Philosophical Foundations of Resource Rational Analysis

by

Brendan Fleig-Goldstein

B.S. Cognitive and Brain Sciences, Tufts University, 2015

B.S. Philosophy, Tufts University, 2015

M.S. Symbolic Systems, Stanford University, 2018

Submitted to the Graduate Faculty of

the Dietrich School of Arts and Sciences in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2024

UNIVERSITY OF PITTSBURGH

DIETRICH SCHOOL OF ARTS AND SCIENCES

This dissertation was presented

by

Brendan Fleig-Goldstein

It was defended on

June 24th 2024

and approved by

Colin Allen, History and Philosophy of Science

Edouard Machery, History and Philosophy of Science

John Norton, History and Philosophy of Science

Thomas Icard, Philosophy at Stanford University

Philosophical Foundations of Resource Rational Analysis

Brendan Fleig-Goldstein, PhD

University of Pittsburgh, 2024

Tacit appeals to systems being rational or apparently irrational are common in cognitive science, and for good reason: irrationality provides valuable evidence for cognitive models. A methodological approach in cognitive science called resource rational analysis attempts to systematize the use of irrationality to develop and test models of cognition. It does so by initially assuming that a system is rational, and then iteratively de-idealizing this assumption by identifying psychological facts that prevent a system from being more rational. This dissertation seeks to analyze how this strategy has worked, how it should work, why it will work, and why it can work better with the conceptual foundations proposed here. In Chapter 1, I develop a specific account of resource rationality. I argue that all epistemic norms are relative to cognitive constraints, and that there is no principled way to distinguish between agents doing their best relative to their limitations and agents being irrational. I advocate for a maximally broad view of cognitive constraints, which renders all agents trivially resource rational, but still allows for meaningful evaluation and prescription. Chapter 2 reviews arguments that intentionality presupposes rationality, and argues that this position is strengthened if the appropriate notion of rationality is understood as my notion of resource rationality. This conclusion shows why rationality considerations are important and even necessary for any intentional psychological science. In Chapter 3, I extend my account of resource rationality to normative commitments, proposing that what I call a meta-reflective capacity—maintaining resource rationality under varying conditions—is necessary and sufficient for possessing normative commitments. This perspective offers a framework for

endowing AI systems with normative commitments and empirically investigating these commitments in humans and non-human animals. Chapter 4 presents resource rational analysis as a methodological strategy in cognitive science and argues for its effectiveness. This strategy, I show, implements a dynamic theory-testing method known as "Closing-the-Loop," as described by Smith (2014). I use the Material Theory of Induction and Topological Learning Theory to provide an epistemic justification for this dynamic testing strategy. These considerations support the iterative de-idealization process and demonstrate the utility of rationality considerations in cognitive science.

Table of Contents

Preface xi

0.0 Introduction 1

1.0 Epistemic Norms Presuppose Cognitive Limitations 12

1.1 Introduction 12

1.2 Ideal and Non-Ideal Norms 17

1.3 Proposals for Distinguishing Between Kinds of Cognitive Limitations . . . 19

1.4 Epistemic Norms Presuppose Perceptual Constraints 24

1.5 Philosophical Conceptions of Rationality Cannot Separate Perceptual Constraints and General Cognitive Constraints 27

1.6 "Ideal" Norms are Sensitive to Perceptual Capacity 41

1.7 Perceptual and Cognitive Capacities Interact 43

1.8 Perceptual and Cognitive Capacities Impede Reasoning in the Same Ways . 48

1.9 Zetetic Epistemic Norms Imply Cognitive Limitations 51

1.10 Beyond Ideal Epistemology 57

1.11 Conclusion 61

1.12 Post-Script: A Note on The Regress Problem 62

2.0 Intentionality Presupposes Resource Rationality 65

2.1 Introduction 65

2.2 The Necessity of a Rationality Assumption 67

2.3 What Notion of Rationality Is That? 75

2.4 Why Not Ideal Rationality? 82

2.5 Why Not Minimal Rationality? . . . 88
2.6 De-Idealizing Resource Rational Models . . . 95
2.7 Conclusion . . . 101
3.0 Meta-Reflective Capacities and Normative Commitments . . . 103
3.1 Introduction . . . 103
3.2 Resource Rationality . . . 109
3.3 Unbounded Meta-Reflective Capacity . . . 111
3.4 Human Meta-Reflection . . . 112
3.5 Non-Human Animal Meta-Reflection . . . 115
3.6 AI Meta-Reflection . . . 117
3.7 Conclusion . . . 120
4.0 Resource Rational Analysis as a Methodological Strategy in Cognitive Science . . . 122
4.1 Introduction . . . 123
4.2 Resource Rational Analysis . . . 125
4.3 The Example of Categorization . . . 131
4.4 Closing the Loop . . . 136
4.5 Why Does Closing the Loop Work? . . . 140
4.6 Topological Analysis . . . 145
4.7 Conclusion . . . 163
5.0 Conclusion . . . 165
6.0 Bibliography . . . 169

List of Tables

1 Comparison of Cognitive Mechanisms and Resources Based on Klein (2018)'s Account 22

2 Cognitive and perceptual capacities interact to determine what an agent ought to do. Methods that perform better are listed first in each cell. Methods that are the best possible given both kinds of constraints are in bold. 45

3 Constraints and Resource Rational Strategies 159

List of Figures

1 Relations Between Notions of Rationality 78

2 Relations Between Notions of Rationality 79

3 Bringing into view an agent's resource rationality: adding constraints to limit the pool of agents under consideration until the evaluated agent performs the best within the constrained pool. Carr (2022) uses a similar picture that uses Kratzer semantics. 80

4 How to Study a Black Box. Using Deviations from Idealized Calculations of Seismographic Waves to Indicate Earth's Boundary Layers 139

5 Inductive risk is isolated to the idealizing assumption. At initial stage of inquiry, no speculative conjecture about the Earth is made other than that it is of uniform density. When prediction and observation disagree, the culprit must be the claim about boundary layers, which is then falsified. 141

6 The method of hypotheses. When prediction and observation are in agreement, confirmation and inductive risk are indiscriminately distributed among the conjectured hypotheses and auxiliary assumptions. 142

7 Closing the loop. Inductive risk is isolated to the idealizing assumption. Note that no conjectures are made about human cognition when only minimal (universal) constraints are initially supposed–other than that there are no further constraints. When prediction and observation disagree, the culprit must be the claim about constraints, which is then falsified. 142

8 Topological relations of hypotheses concerning polynomial degree. If the true hypothesis is polynomial degree n, then polynomial degree n+1 is false but will never be refuted. The reverse is not true. Lower degrees are therefore nested in the frontiers of higher degrees. 150

9 Topological analysis of two hypotheses: w_1: Humans are ideally rational. w_2: Humans are not ideally rational 154

Preface

Thank you, Pittsburgh, for being a "most livable city" and providing an immensely rewarding intellectual environment over these past years. Pitt HPS has made me feel at home, while CMU Philosophy and Pitt Philosophy have provided perfect counterweights on either side.

Thank you to all my fellow Pitt HPS graduate students, especially Nuhu Osman Attah, Dejan Makovec, James Michelson (CMU), Mara McGuire, Dzintra Ullis, and Gal Ben Porath. Thank you also to Katie Creel and Zina Ward.

I am grateful to Natalie Simpson, Chris Young, and David Rapach for creating important spaces in Pittsburgh. Thank you to Dr. Lee Weinberg, Laura Kleiber, and everyone else who kept me healthy.

A special thanks to the dogs in my life: Alma, Apollo, Maddy, Luna, Bishop, Bagel, and Zoey.

Thank you to Mike Dietrich for being an amazing department chair at Pitt HPS and for teaching me so much about historical research. I have learned something incredibly valuable from every Pitt HPS faculty member (including a few people who are no longer there). Thank you to everyone. Additionally, thank you to Kevin T. Kelly at CMU for being a friend and a role model.

Thank you to Jennifer Berkbile, Matt Ceraso, Diana Volkar, Delaney Szekely, and Joann McIntyre for countless instances of going above and beyond with their help.

Thank you to Brett Karlan for his helpful and supportive comments on this work.

I am grateful to my committee for challenging me so fiercely. Thank you to Thomas Icard, who has taken me seriously since I was a master's student fresh out of undergrad,

always been kind, and taught me so much of what has gone into this dissertation. Thank you to John Norton for exemplifying the kind of philosophy of science I admire and for always offering brilliantly fast and incisive feedback. Thank you to Edouard Machery for being supportive, generous with his time, providing sage advice, and meticulous comments. Thank you to Colin Allen for being the best advisor I could ask for–a true doktorvater–and offering endless help at every turn.

Thanks to my whole family: Rachel, Jon, Ellen, and my parents, Kathy Fleig and Richard Goldstein. My parents have been my first and greatest teachers and more supportive than I could ever hope for. Finally, thank you to my soon-to-be wife, Eliza Klyce. She has been by my side every step of the way, from when I applied to Pitt HPS to my dissertation defense, and has been the best, most supportive, and magical partner I could dream of.

0.0 Introduction

Building foundations looks a lot like digging yourself into a hole. If you want to build up a solid house, you have to first excavate down into the ground. Someone watching you at first will shake their head in pity as you sweat in your muddy hole. However, only after this initial digging can concrete be poured and allow for the beginning of construction.

In this analogy, the house is cognitive science. My dissertation is attempting to pour some concrete for the foundation, but that means most of the time I will look like I'm digging myself into a hole. In this dissertation, I defend a variety of strange claims. For example, I argue that there is no difference between rational and irrational agents, and that instead there are only resource rational agents–agents doing the best they can relative to their cognitive constraints. I argue for a "maximally broad" account of what can count as a cognitive constraint, such that all agents are trivially resource rational. This view at first might seem both counterintuitive and unlikely to be helpful, but I hope the reader will be patient and see that there are, in fact, quite attractive features. Each of the four chapters will show that there is either a push or a pull for this view.

The purpose of these claims, however, is to provide the foundation for a methodological strategy in cognitive science called resource rational analysis. This dissertation seeks to analyze how this strategy has worked, how it should work, why it will work, and why it can work better with the proposed foundations. That is the highest layer of the philosophical foundation. Consequently, the justification of resource rational analysis comes at the very end of the dissertation. The earlier parts of the dissertation–the deeper parts of the foundation–are designed to support the aims of Chapter 4 by independently motivating the conceptual ingredients necessary for this methodology.

This dissertation takes as a starting point the view that the scientific study of the mind is an immature science; there is a lack of well-evidenced and wide-scope theory with which to help turn data into evidence for theoretical claims. Given the starting point that current cognitive science is in its infancy, this makes looking at case studies in cognitive science a tricky matter. For example, if one is interested in understanding mental representations, one might look at current neuroscientific practices to find the concepts implicit in this research. Such a descriptive project, aiming to identify the concepts used in current science, is certainly worthwhile. It is similarly valuable to answer questions like "what is time?" by considering what our "best sciences" currently say about it. But here I am taking cognitive science to not yet have that honorific. Thus, focusing solely on the descriptive task for the conceptual foundations of cognitive science as it stands now misses the opportunity to explore slightly different, philosophically motivated concepts that could facilitate new forms of scientific practice.

This dissertation does not, subsequently, try to simply read off the philosophical foundations of resource rational analysis from current scientific practice. This project is prescriptive rather than descriptive. It still takes as its starting point actual scientific practice. It identifies intuitively convincing cases of scientific research. But from there, it tries to independently motivate an account for why such scientific reasoning is compelling and then, in turn, use such an account to make methodological prescriptions. It then also looks to actual scientific practice to make sure that its prescriptions are possible more than in theory and work as expected.

The focus of this dissertation is on cases from contemporary cognitive science where apparent irrationality is used to understand how cognition works. The goal is to develop an account of why apparent irrationality is an invaluable source of evidence in cognitive

science and provide recommendations on how to systematically and principledly use apparent irrationality to develop and test models of cognition.

That is the high-level view. Here is a more detailed plan of the work:

If you reader, will forgive me, I will start at the end and then return to the beginning. Chapter 4 presents resource rational analysis as a methodological strategy in cognitive science. Resource rationality is, basically, optimality relative to cognitive constraints. More accurately, resource rationality involves optimality of behavior relative to a class of environments, an evaluative standard (e.g., accuracy of reported beliefs, amount of dollars won, etc.), and a set of cognitive constraints.

Resource rational analysis starts by assuming that humans behave optimally relative to *minimal* cognitive constraints, and then uses this assumption to derive predictions of human behavior. Deviations between calculated and observed behavior then indicate further cognitive constraints. These constraints are independently tested and then incorporated into new calculations of resource rational behavior and the process iterates. The hill I would like to die on is that this iterative strategy is a particularly effective method for marshaling evidence for cognitive models.

To establish the effectiveness of resource rational analysis, I first claim that resource rational analysis employs a specific logic of theory-testing that Smith (2014) calls "closing the loop." This dynamic theory-testing strategy is exemplified by the historical episode in which geophysicists gathered evidence for density distributions within the Earth's interior by iteratively de-idealizing the assumption that the Earth is an elastic medium of uniform density.

Why is closing the loop epistemically justified as a logic of theory-testing? Why think that it is particularly effective for marshaling evidence for cognitive models? To answer

this question, I analyze the unique obstacles that cognitive scientists face in providing high-quality evidence for theoretical claims. Much of the history of science witnesses a snowballing effect: possessing more established facts about a domain allow you to infer more facts, which in turn allow you to infer even more facts, in a positive feedback cycle (Norton, 2023). I argue that a consequence of this large-scale dynamic of science is that the method of hypotheses (or hypothetical induction) is less effective at earlier stages of inquiry. The main limitation of hypothetical induction is the problem of indiscriminate confirmation (Norton, 2003). Confirmation is "indiscriminately" distributed among different theories that all predict the same phenomena, and similarly, evidential credit and blame are indiscriminately distributed within a particular theory's different posits. At earlier stages of inquiry, when less is known, this problem is exacerbated. The reason why closing the loop is effective, I claim, is that it provides an effective means of managing the problem of indiscriminate confirmation in these early stages of inquiry.

Closing the loop strategies, I argue, involve both a mediating principle and an idealizing principle. The mediating principle makes an assumption about the system of interest that allows a body of well-evidenced theory to transform observable data into claims about unobservable processes. In resource rational analysis, the mediating assumption is that humans are rational relative to their cognitive constraints, allowing theoretical demonstrations of agents actually optimally relative to constraints to turn human behavior into evidence for cognitive processes. The idealizing principle, meanwhile, provides an initial starting point from which deviations emerge, leading to iterative de-idealization. The idealizing principle in resource rationality is that humans have minimal constraints.

There is no mystery as to why using mediating principles can be a good strategy in science. By making prior well-evidenced general theories applicable to an otherwise mysterious

system, one reduces the system's mystery. But what justifies the idealizing principle and the strategy of iterative de-idealization? I use topological learning theory to provide a rationale for this iterative de-idealization strategy. Specifically, I show that a preference for hypotheses that assume humans have fewer cognitive constraints is a preference for topologically simpler hypotheses. Cognitive constraints are, therefore, learnable, and resource rational analysis can be viewed as an Ockham method in K. Kelly (2024)'s sense. This provides a new perspective on resource rationality and its utility as an inductive strategy in cognitive science.

The mediating and idealizing principles work together to avoid the limitations of the method of hypotheses, specifically by managing the problem of the indiscriminateness of confirmation. They achieve this by isolating inductive risk to particular posits at a time. The mediating principle makes a set of well-established facts from outside the local domain of interest bear on the local domain, while the remaining inductive risk is isolated via the idealizing principle to specific claims about the black box that can be given a learning theoretic justification. Instead of attempting to confirm a complete model of a cognitive process, psychologists focus on falsifying the assumption that all relevant constraints have been identified. All in all, this analysis provides an epistemic justification for adopting resource rational analysis and, along the way, sharpens exactly how the methodology should work.

Notice, for instance, that in this methodology, facts about psychological processes are discovered by discovering constraints–if one has a narrower view of what can count as a constraint, then that only limits the kinds of psychological facts that can be discovered and tested in this method. For instance, if perceptual limitations are not considered constraints on rationality, then resource rational analysis cannot consider these kinds of constraints along

with other more paradigmatically cognitive constraints interacting together to shape resource rational behavior from which deviations can be identified. The topological learning theoretic justification, in fact, does not work without the maximally broad account of constraints. It won't work if being resource rational is a contingent empirical question; it will only work if agents are trivially resource rational and the question is what are the constraints relative to which they are resource rational. Furthermore, the topological analysis does not work from an ideal rationality perspective, where agents are viewed as irrational and falling short of an ideal in various ways. In such contexts, no topological simplicity order will be induced. My notion of resource rationality is necessary for the topological analysis to work. So here is a pull for my maximally broad view of constraints and resource rationality.

So that's the end of the dissertation. Here's everything leading up to it:

Chapter 1 concerns epistemology and rational evaluation. If the study of the mind is to rely on theories of rationality to develop and test models of cognition, it needs in hand an appropriate account of rationality with which to evaluate and prescribe human behavior.[1] Resource rational analysis suggests that the appropriate account is resource rationality. This chapter attempts to independently motivate the view that resource rationality–in particular, my maximally broad notion of constraints account–is, even apart from the scientific study of the mind, the right account with which to rationally evaluate and prescribe human behavior.

Resource rationality is a form of non-ideal epistemology. Non-ideal epistemology is a trending topic (McKenna, 2023). There is a growing suspicion among a minority of epistemologists today that, in fact, all epistemology is non-ideal epistemology (Thorstad, 2023). Few have defended this strong claim or presented explicit arguments for it. Chapter 1 at-

[1]Note that I do not defend the claim that all methodologies in cognitive science *must* rely on theories of rationality. Chapter 2 gives reasons to suggest that any *intentional* psychological science must rely on theories of rationality.

tempts to do so.

The basic idea is this: epistemic norms are non-ideal when they lower the bar for what is expected of you based on what you can do. But what you can do depends not only on facts about your cognitive system, but also facts about the environment, and your access to the environment. I give cases, for example, where noisiness of visual cues can result from either the world itself, one's perceptual system, or one's more "central" cognitive system (e.g., memory processes subserving reasoning). In either case, such noise results in changing what one should do in terms of inference; norms of reasoning are relative to such situational facts. Ideal epistemology tries to state that norms are non-ideal only if the noise comes from the central cognitive system. I explore ways philosophical accounts of rationality can provide a principled justification for this claim and show that they end up lapsing into non-ideal epistemology. For example, one might defend a subservience condition and claim that norms are non-ideal only when they are relative to cognitive processes that subserve reasoning. But since irrationality often involves a breakdown or a lack of reasoning, or else involves other cognitive processes that occur squarely outside of such narrowly understood reasoning processes, this condition results in counting a great deal of irrationality as ideal.

The overall argument is that it arbitrary to try to draw lines around parts of a cognitive system, and say that certain parts (e.g., memory) but not others (e.g., visual acuity) can be abstracted away from in characterizing ideal norms. It is arbitrary to try to say which psychological facts lower the bar for rationality and which ones do not. It is better instead to just acknowledge that all epistemic norms are relative to facts about the environment, and one's entire cognitive system is part of that environment. This supports a maximally broad view of what counts as cognitive constraints.

The upshot is that there are no ideal norms, and consequently no ideal agents. There

are only agents doing their best relative to the environment they are embedded in, including their own cognitive constraints. Agents can still be evaluated relative to one another in terms of their performance–e.g., accuracy of belief, amount of money won in a game, etc. Further, agents can be analyzed as being subject to more or less constraints. One can still say that an agent could have done better if they had had more memory or better visual acuity. Further, characterizing the mutability of constraints allows for a rich analysis of in what sense agents "could" have done better. Thus, while initially strange sounding, the "all agents are trivially resource rational" perspective turns out to be motivated and not so bad after all.

With an account of rationality in hand, Chapter 2 tries to answer the question of why rationality should be intertwined with the study of the mind in the first place. Several traditions in the philosophy of mind hold that an agent having intentionality–mental states such as beliefs, desires, intentions, and so forth–presupposes the rationality of that agent. But what notion of rationality is invoked? I argue that my account of resource rationality provides the relevant way in which intentionality presupposes rationality.

I examine the debate between those who believe intentional systems must be rational and those who argue that cognitive creatures can and do commit blatant violations of rationality. I argue that the latter arguments do not in fact defeat the reasons for thinking that intentionality presupposes rationality–they just show that attributions of irrationality to intentional systems are possible. I thread the needle between these positions. I argue that the view that intentionality presupposes rationality is strengthened by acknowledging that intentional systems need not be ideally rational. Attributions of irrationality are possible, but only if it is shown how intentional systems are still doing their best relative to constraints. In other words, attributions of "irrationality" should be understood as specifi-

cations of constraints for resource rational agents. This chapter ends with a discussion about how this perspective allows for an important bridge between intentional and non-intentional psychological analyses.

The arguments that intentionality presupposes rationality are controversial. If you are one who is not susceptible to such persuasions, then the conclusions of this chapter will be slightly weaker but still important. The weaker conclusion is an epistemic rather than metaphysical claim: perhaps intentional states do not depend on the rationality of an agent, but the only way we can *come to know* the mental states of others is by adopting a (resource) rationality assumption. This epistemic conclusion is plenty sufficient for the purposes of the philosophical foundations of resource rational analysis–as a scientific methodology, resource rationality concerns coming to know. The upshot of Chapter 2 is that any intentional psychological science must rely on a theory of rationality.

While Chapter 2 concerns straightforward beliefs, desires, and other bread and butter intentional items, human mentality is home to a variety of rich kinds. Chapter 3 looks in particular at normative commitments, such as an ethical commitment to not harm animals, an epistemic commitment to not have inconsistent beliefs, or a commitment to keep a promise made to a friend. Having a normative commitment is generally understood as involving a particular kind of relationship to reasons. Namely, you have a normative commitment to x only if that commitment is a reason for you to believe or act in certain ways (Millar, 2004). How could a psychologist determine if an agent has a commitment? How could an AI engineer endow an artificial system with commitments? The AI engineer works with causal and informational systems–they do not work with reasons.

I use my maximally broad notion of constraints and resource rationality to characterize what I call a meta-reflective capacity, which involves maintaining resource rationality under

varying resource conditions. While all agents are trivially resource rational relative to every goal, it is not the case that manipulating an agent's set of constraints A to B will result in it continuing to maintain an optimal cognitive strategy relative to B (it may after such modification only be resource rational relative to a different set of constraints C). That is, adding memory to a memory-bounded agent does not mean it will necessarily adapt its cognitive processes so as to make the best use of that new amount of memory. To the extent that it does, however, it has a meta-reflective capacity.

I claim having an unbounded meta-reflective capacity relative to a particular goal x is necessary and sufficient for a causal system to have a normative commitment to x. This claim is an empirical claim about the necessary and sufficient conditions for possessing normative commitments; it is not a conceptual analysis and remains neutral regarding whether normative commitments can be *identified* with such a meta-reflective capacity. I examine meta-reflective capacities of human, non-human animal, and AI systems. In each case, achievements of normative commitments are shown to be cases of meta-reflective capacities, and failures are shown to be failures to have meta-reflective capacities.

Given my claim, this gives a target for AI engineers who wish to endow systems with normative commitments. It also allows psychologists to empirically investigate normative commitments since meta-reflective capacities and their limits can be empirically demonstrated. This chapter will illustrate yet another pull for a maximally general notion of cognitive constraints–it is necessary for a characterization of a concept of meta-reflective capacity that can serve as a necessary and sufficient condition for normative commitments.

The general theme of this dissertation is the importance of using rationality considerations for the study of the mind. My aim is to provide a principled framework for doing so.

In attempts to model cognitive processes and provide evidence for these models, cognitive scientists either ignore the fact that the human cognitive system is designed to behave rationally, or they do not. If they choose to ignore rationality considerations, they resign themselves to studying the most complex object in the known universe as if they were examining the function of an immensely intricate box of wires, without any understanding of the principles behind its design. Conversely, if they choose to use rationality considerations, then empirical psychologists must rely on ideas about the nature of rationality and find effective methods to apply such ideas to achieve their ends. Too often, psychologists rely on simplistic models of rationality and unsystematic ways to use them.

The purpose of this dissertation is to provide a philosophical foundation for a methodological strategy in cognitive science that systematically uses rationality to build and test models of cognition. While resource rational analysis is already a recognized research methodology, the philosophical foundation I provide prescribes new and precise ways to implement this approach, as well as justifications for its adoption.

1.0 Epistemic Norms Presuppose Cognitive Limitations

Epistemology is often thought to come in 'ideal' and 'non-ideal' flavors. The former is meant to capture pure epistemic norms, free from contingent constraints of cognitive capacities, while the latter incorporates such limitations into epistemic evaluations and prescriptions. But ideal epistemic norms necessarily govern agents that are perceptually bounded. I argue that philosophical conceptions of rationality cannot separate subpersonal cognitive constraints that give rise to non-ideal norms from perceptual constraints that are necessary for characterizing ideal norms. This conclusion breaks down the distinction between ideal and non-ideal epistemology. Critics of non-ideal epistemology have argued that non-ideal epistemology cannot distinguish between agents doing their best relative to their limitations and agents being irrational (Carr, 2022). My argument shows that this problem also infects ideal epistemology, putting pressure on the idea that any line can be drawn between rational and irrational agents. The news is not all bad: acknowledging the inescapable relativity to cognitive constraints leads to a richer perspective on epistemic and zetetic norms (norms of inquiry), and a richer way to characterize the rationality of agents.

1.1 Introduction

In chess, both humans and supercomputers have limits to how many moves ahead they can imagine, due to the game's complexity and their own cognitive limitations. We do not typically judge grandmasters or advanced computer programs as poor chess players because of these limitations. Should they be expected to play as if they can foresee every potential

outcome? If a supercomputer analyzed a board setup from a world championship game and discovered that after 10,000 hours of grinding calculations, the game's winner missed a quicker winning move, what would that imply about the grandmaster's skills?

To say only that the grandmaster fell short of an ideal fails to appreciate the entirety of the situation. In cases where people are almost guaranteed to fall short of perfection, determining whether a person is doing the best that they can relative to their limitations adds additional information. Assessing only whether a grandmaster played a perfect game leaves open the question of whether they played the best game they could have, given actual human cognitive capabilities and the time available. This is a motivation for investigations into non-ideal epistemology, such as the study of bounded rationality, which seeks to develop notions of "rationality relative to limitations."

To assess if someone is performing well given their limitations, one needs to differentiate between kinds of limitations. If anything can count as a limitation, then all agents are trivially doing the best they can given their limitations. If being really bad at chess counts as a limitation, then no matter how badly someone does in a game, they are still doing the best they can relative to the fact that they are bad at chess. More accurately, every chess player could be said to be doing the best they can given their limitations, since being distracted, not thinking through the consequences of a move, ignoring your opponent's pieces, and so forth, all can count as, or be byproducts of, cognitive limitations. Once these are factored into the analysis, all chess is boundedly rational chess.[1]

In order to have a useful notion of bounded rationality, what seems to be needed is a way to specify which kinds of limitations "lower the bar" for what counts as rational and which ones do not. If this can be done, then it becomes an empirical matter whether or not

[1]Being good at chess can be understood as a matter of practical rationality, where the goal is to win. In what follows, I will focus more on cases of epistemic rationality.

someone is doing the best that they can. For example, if a particular amount of memory and processing capacity can be specified for humans playing chess, then boundedly rational chess involves making the best possible choices relative to these constraints. Doing worse than that because you are distracted or ignore your opponent's pieces counts as irrational, not boundedly rational. Bounded rationality seems to need a way to fix specifications of cognitive limitations that rationality is relative to, and differentiate these facts from other kinds of cognitive limitations.

Although I am an advocate of the perspective of bounded rationality in epistemology, I agree with critics of bounded rationality in that there is no epistemically principled way to solve the above problem and distinguish between limitations that lower the bar for rationality and those that do not. Call this Carr's problem (Carr, 2022). I argue, however, that ideal rationality is plagued by the same problem. Plausible conceptions of ideal rationality 1) imply substantive cognitive limitations, especially in the form of perceptual limitations, and 2) struggle to exclude these perceptual limitations from being considered irrational without forsaking the concept of ideal rationality altogether. The consequence is that supposedly ideal rationality is subject to Carr's problem, and that there is only non-ideal rationality. The upshot is not all dire, however. I go on to provide the start of a positive account for thinking about epistemic prescriptions and evaluations in light of this result.

After laying out some background in §1.2 concerning ideal and non-ideal epistemology, in §1.3, I explore how advocates of bounded rationality can try to delineate cognitive limitations in a principled manner (Klein, 2018; Thorstad, 2023). I argue that these frameworks will not solve Carr's problem. In §1.4, I demonstrate the necessity of perceptual limitations in ideal epistemology. Briefly, full-blown omniscience is not what anyone means by ideal rationality. Relaxing the standard away from full-blown omniscience requires relativizing

ideal performance to the data available for a particular task. Since perceptual capacity determines the nature of the available data, characterizing epistemic norms requires fixing perceptual capacities.

Perceptual limitations are generally not considered to undermine an agent's rationality. However, in §1.5, I argue that philosophical conceptions of rationality should treat subpersonal perceptual capacities similarly to general subpersonal cognitive capacities. The idea is this: if epistemic norms are sensitive to cognitive limitations, then such norms are non-ideal. While subpersonal perceptual limitations may not be rationally evaluable, neither are the standard cognitive limitations that render epistemic norms non-ideal (e.g., memory and attention constraints). To argue that perceptual limitations do not make norms non-ideal, one might propose a subservience condition. According to this, if epistemic norms are sensitive to states that support or participate in reasoning or belief formation, then the norms are non-ideal. However, irrationality often stems from either a breakdown in reasoning or belief states (non-reasoning and non-belief states), or from good reasoning that leads to irrational outcomes due to interactions with other cognitive subsystems, which may not support reasoning directly. Such examples suggest that the subservience condition fails to preserve ideal rationality in practice. This discussion leads to a dilemma for epistemologists: they must either narrow the scope of rationality to exclude both perceptual and cognitive constraints from affecting rational evaluation or broaden the scope to include both. In either scenario, perceptual and cognitive limitations align in their impact on epistemic evaluation, and in both cases epistemic norms are inevitably non-ideal.

§1.6-8 examines specific cases to pump the intuition that it is insensible to abstract away from cognitive limitations but not perceptual limitations when characterizing epistemic norms. The two kinds of constraints interact and are hopelessly entangled. What an agent

epistemically ought to do depends on their cognitive and perceptual capacities considered, not in isolation, but together at the same time. So while §1.5 argues that philosophical views are unable to distinguish between these kinds of limitations, §1.6-8 argue that they should not do so anyhow. Philosophers should take the same stance toward perceptually bounded agents as agents who are cognitively bounded in other ways.

§1.9 discusses norms of inquiry, also called zetetic norms (Friedman, 2020). I claim that the arguments I give in this paper and the increasingly popular view that there are genuinely epistemic norms of inquiry are mutually supportive of one another. Zetetic norms, I argue, entail that perceptual constraints are constraints on epistemic rationality, since what information you ought to seek out depends on what information you can seek out. Conversely, I argue that my discussion gives novel support to the view that there are genuinely epistemic norms of inquiry.

The upshot is that there is only non-ideal epistemology. Non-ideal epistemologists have suggested that this is the case, but few have argued for this claim (Thorstad, 2023). The purpose of this paper is to make explicit the argument for this position. Further, while one can arbitrarily fix constraints and ask if someone is doing the best they can, there is no absolute sense of being ideally or non-ideally rational. Other than fully omniscient agents, it is always the case that agents can do better, given different psychological capacities. Together, these points put pressure on the view that a line can be drawn between agents that are rational and irrational. In §1.10, I argue that this view has attractive features. Absolute judgments about an agent's rationality can still be made: this is done by evaluating the agent according to a normative standard (e.g., belief accuracy) *and* specifying the psychological constraints that are preventing the agent from doing better. Even without drawing a line between the rational and irrational, agents can still do better or worse according to a standard, and be

more or less constrained. Finally, an analysis of the mutability of constraints also allows for a rich analysis of what agents could have done better, and in what sense.

1.2 Ideal and Non-Ideal Norms

I adopt in this paper an alethic meta-epistemology. On this view, the point of epistemic norms is to promote true beliefs, or accurate partial beliefs (Joyce, 2009). This is in contrast to a view where, say, coherence is an end in itself of rationality, as opposed to a means of getting to the truth.

In this section, I review Carr (2022)'s challenge to bounded rationality. I focus on Carr's argument because it is one of the most compelling critiques of bounded rationality and I agree with it. There is generally no epistemically privileged way to set bounds. So, there is no way to differentiate between an agent doing their best relative to constraints versus an agent being outright irrational. I disagree with her, however, in that I think ideal epistemology is subject to the same problem.

According to Carr, who also adopts an alethic meta-epistemology, "the fundamental distinction between ideal and non-ideal epistemology is that non-ideal epistemology holds that the epistemic 'ought' implies some substantive 'can'; ideal epistemology does not" (Carr, 2022, p. 1136). Non-ideal norms are relative to limitations, but which limitations? Admittedly, there seem to be intuitive differences between cognitive limitations that lower the bar for rationality and those that do not. Carr gives the following tentative lists for consideration:

> Cognitive limitations that lower the bar for non-ideal rationality: our limited computational power, informational storage, processing speeds, integration of different cognitive systems, information retention...
> Cognitive limitations that don't: our dispositions toward implicit biases, unreliable heuris-

> tics, delusional reasoning, misinterpreting statistical phenomena as having causal explanations, an inflated sense of one's own driving ability, over-optimism/pessimism, overestimating the moral superiority of one's own side in a fight with a spouse, family member, or departmental faction.... (p. 1152)

Carr considers and dismisses several potential criteria for distinguishing between these categories, such as that the former are universal, not culturally contingent, or unchangeable properties of mind, while the latter can be improved upon or altered by individuals. The problem is that whether the above list items fit any of these criteria is seriously context-sensitive. By seriously context-sensitive, Carr means that there is "no normatively privileged resolution of one or more of the context-sensitive parameters for evaluations of that kind" (p. 1135).

What an agent can do always depends on how to interpret the modality of 'can.' She gives the following example: can one touch their heels from behind by doing a standing backbend? 'Can' in what sense? Right this minute? After warming up? After months of mobility work? There is no default resolution to the nature of the modality. It is the same as what an agent can do cognitively. Right this minute? After learning the correct epistemic strategy? After months of training your memory capacity? After implanting computer chips in your brain? But if what an agent ought to do depends on what they can do, then how can an agent be said to be boundedly rational or irrational, when there is no privileged understanding of what they can do, or what they should be able to do?[2]

Carr's point is that one cannot separate in a principled way limitations that should change our standards of rationality from those that lead to outright irrationality. So non-ideal norms depend on arbitrary choices of what an agent can do, while ideal norms express robust, non-arbitrary norms. Only ideal epistemology, for Carr, cuts normativity at its

[2]There is more to Carr's argument than I can do full justice to here.

joints.

One can stipulate what cognitive limitations are being held fixed, and then ask what the best that can be done is relative to these. But if one wants non-arbitrary epistemic prescriptions and evaluations, only ideal epistemology—which, according to Carr, does not require fixing cognitive limitations—can provide that.

1.3 Proposals for Distinguishing Between Kinds of Cognitive Limitations

Proponents of the bounded rationality framework naturally would like to defend non-ideal epistemology against the charge that it yields only arbitrary norms. Thorstad (2023) has responded to Carr's problem. Thorstad's proposal is that facts about *cognitive architectures* specify the facts that lower the bar for rationality. Thorstad specifically has in mind Newell's notion of cognitive architecture–though not Newell's particular cognitive architecture theory (Kotseruba and Tsotsos, 2020; Langley et al., 2009). Cognitive architectures, in this sense, specify the fixed and (usually) domain-general structure of cognition—e.g., facts about memory storage and retrieval, the timing of decision-making steps, and so forth. These facts place restrictions on the sorts of cognitive strategies that a system can employ. Thorstad suggests that an agent should be considered boundedly rational if they are utilizing the best cognitive strategy their architecture can support. Doing worse constitutes irrationality. This proposal appears to sort Carr's two lists mostly correctly. It is also, as Thorstad says, a well-motivated, non-arbitrary proposal since cognitive architectures are an independently important concept in cognitive science.

Part of what is being questioned in the first place, however, is whether there is a clear distinction between mutable and immutable cognitive limitations. If there is, then this

would provide a plausible principled distinction to differentiate between kinds of cognitive limitations and provide a response to Carr's problem. However, the concept of cognitive architectures presupposes this distinction rather than demonstrates it. The viability of Newell's cognitive architectures concept depends on the existence of a clear division between mutable and immutable cognitive structure that is not seriously context-sensitive. Not only are the particular empirical facts about human cognitive architectures not yet established but importantly, the very idea that there is a clear dividing line between what can count as part of an architecture or not is an open question. There may be clear facts about our memory limitations that are domain-general, immutable, and applicable across all contexts. But if not, then that poses difficulty for the Newell cognitive architecture concept. Appealing to cognitive architectures assumes that a response to Carr is available; it does not provide a reason to think that there is one available.

In general, any current proposal about how to understand cognitive limitations to address Carr's problem will assume rather than provide a solution. Any proposal will require that human cognition possesses cognitive resources, not just in a certain "amount," but also of a particular kind. Whether humans have resources of that kind will be an empirical question not yet settled and, therefore, speculative. Whether cognitive processes are computational or not, symbolic or connectionist in nature, and so forth, affect the question of what kinds of limitations and resources human cognition possesses. How to understand the cognitive limitations humans have will depend on big questions about the nature of human cognition. Metaphysical proposals that distinguish between kinds of cognitive limitations may not be applicable to the kinds of limitations that humans have. Appealing to a substantive view of cognitive limitations assumes rather than provides a reason for thinking that there is a principled way to distinguish between cognitive limitations.

Further, even if a proposal provides a clear way to distinguish between different kinds of cognitive limitations, and putting aside the question of whether the distinction is applicable to human cognition, it is not clear why such a separation of descriptive facts results in an epistemically meaningful distinction. It is an open question to ask: these types of constraints are descriptively distinguishable in such and such a way, but does one kind lower the bar for non-ideal norms, and the other kind does not? Sometimes, it is desirable to treat different degrees of mutability as immutable and ask whether someone could have done better with different kinds of modality in mind. For example, even if cognitive architectures are appropriately applicable to human cognition, non-ideal epistemologists might be legitimately interested in whether different cognitive architectures are making better or worse use of the biological materials available. Might evolution have wired together a better architecture out of the available tissue and metabolic resources? In the other direction, a non-ideal epistemologist might legitimately hold more than the cognitive architecture fixed, and ask if someone is being rational relative to, for example, their unjustified conspiracy theory worldview.

All one has to do to change which limitations lower the bar for rationality is to consider a different modality for what an agent's cognition could be like. Which modality should epistemologists be considering? Without a privileged resolution, for non-ideal epistemology, there can be no fact of the matter concerning which agents are rational or irrational. While that conclusion alone might not alarm you, as a reminder, I am going to argue this problem infects ideal epistemology as well–meaning that there is no fact of the matter *ever* concerning which agents are rational or irrational.

I want to consider an altogether different way to specify determinate facts about cognitive limitations, which comes from Klein (2018, 2022). Klein has put forth his proposal not to address Carr; it is an independently motivated analysis of how to think about cognitive

resources. Applying Klein's analysis to Carr's problem is illuminating because, even though it is a clear way to think about facts about cognitive resources, I believe it can be used to offer new support to Carr's conclusions.

Klein (2018) argues that we should think of cognitive resources as finite quantities that are used by, and to be contrasted with, the entities and activities that constitute cognitive mechanisms–in the new mechanist sense of especially Craver (2007). Klein writes that the distinction should be thought of as roughly similar to an agent-patient distinction that can be sharpened along 4 criteria (see Table 1). This would then allow for a specification of facts about cognitive resources and, in theory, a principled analysis of what cognitive mechanisms make the best use of those limited resources.

Cognitive Mechanisms	**Cognitive Resources**
Persist	Consumed
Composite of distinct elements	Aggregate of homogenous units
Realization indifferent	Realization sensitive
Causally conservative	Causally promiscuous

Table 1: Comparison of Cognitive Mechanisms and Resources Based on Klein (2018)'s Account

One issue in using this analysis of cognitive resources to reply to Carr's problem is that there are different kinds of resources at different scales of analysis of cognition, even if there are determinate facts about what they are at each scale. E.g., one can abstract away from hardware and talk about algorithmic space and time complexity and resources, or one may wish to talk at a lower level about biological (e.g., metabolic) constraints that affect cognition. While this is compatible with Klein's view and the new mechanist understanding of scales of mechanisms, this gives rise to Carr's problem in a new way: what scale do we

want to hold fixed to specify resources? If we focus on the level of biological metabolic constraints, then the higher level organization could likely have been better. But if we hold a higher level fixed (e.g., the cognitive architecture), we get a completely different answer for what could have been better.

Further, Klein (2022) writes:

> Note that the mechanical part/resource distinction—and therefore the satisfaction of the first three criteria—is often explanation-relative. If I care how my car stops, brake pads are mechanical parts: persistent, individually important, and functional. If I'm managing a racing team, I may go through so many brake pads that I treat them as a consumable resource. Whole mechanisms in one context can be mere resources in another: the jeep is a complex whole to the mechanic, matériel to the quartermaster. The point of distinguishing mechanical parts and resources is thus not to draw a firm metaphysical line in the world, but to emphasize the different roles that different spatiotemporal parts of a mechanism can play within the same explanation. (p. 4)

If Klein is right, then facts about cognitive resources are fixed by how cognitive processes enter into a resource constraint explanation. Cognitive resource facts enter into analyses of how cognition is shaped by constraints, but cognitive resources are not independently determinate enough to provide facts of the matter about which limitations to hold fixed to lower the bar for rationality.

In agreement with Carr, I conclude that there is no normatively privileged way to fix cognitive limitations to yield robust non-ideal epistemic evaluations and distinguish between agents that are (boundedly) rational and irrational. Contra Carr, however, I argue this problem also infects what is typically thought of as ideal epistemology.

1.4 Epistemic Norms Presuppose Perceptual Constraints

Non-ideal epistemology involves norms that are sensitive to cognitive capacities, and there is no epistemically principled way to set cognitive capacities. It seems, then, that there is no way to distinguish between non-ideal agents that are doing the best they can and ones that are not.

Why would this problem not affect ideal epistemology? It would need to be the case that ideal epistemology does not presuppose limited cognitive capacities, or that the kinds of limited cognitive capacities ideal norms presuppose can be principledly distinguished from the ones non-ideal epistemology concerns itself with. I argue that neither is true. In this section, I briefly explain why familiar ideal epistemic norms presuppose particular perceptual capacities. In the next section, I will explain why these capacities cannot be principledly distinguished from the ones non-ideal epistemology concerns itself with.

One sense of absolute epistemic perfection is fairly straightforward, especially given an alethic perspective: one should "ideally" immediately believe the truth, the whole truth and nothing but the truth, in all contexts everywhere and always (see K. Kelly, 1996 (p. 159), who makes this point as well). To do anything else gives rise to the question of why one did not do so.

Epistemologists face difficult decisions concerning evaluative standards exactly because perfection is impossible. For example, one need only figure out how to trade off William James' two maxims of seeking truth and avoiding error when one cannot do both perfectly—otherwise the two can be accomplished at once without a trade-off. Controversy in epistemic standards often arises as a consequence of figuring out how to weigh failures or mistakes against one another. But there should be little controversy in what counts as

superlative epistemic perfection: infallible omniscience.

Ideal epistemology does not concern itself with this kind of superlative epistemic perfection, and for good reason. To put it in terms of aim-means-result terminology, if the aim is truth, then the ideal result is know all truths. But there is no means for this result (Norton, 2021; Putnam, 1963; Wolpert and Macready, 1997). So one cannot prescribe a perfect means, but only demand a perfect result, which is to just say "just immediately infer the truth always." Such a demand is a useless, inert norm, similar to telling someone to "just do the right thing" (S. J. Russell and Wefald, 1991).

Ideal epistemology rightfully does not concern itself with such demands. Instead, ideal epistemology seeks to prescribe the best means possible.[3] The best means possible means the best relative to the available data–now the best result is achievable because it is defined as relative to data. But what data one has is determined by bounded perceptual capacities.

Familiar views of ideal rationality–e.g., Bayesian norms–tacitly adopt constraints on an agent's access to the environment they are in. This is how most epistemic problems are generated in the first place. If an agent has unbounded perceptual access to the world, then they already possess the empirical facts. If the task is to predict subsequent data points in an incoming sequence of data, but the agent has already been given access to the entirety of the data stream, there is no induction or reasoning needed. At the most minimal, agents' perceptual capacities must be temporally bounded; they must not be privy to the future. But other sorts of bounds on perception are needed as well. Problems that are not prediction tasks, such as reasoning about the subatomic world on the basis of observable data, require that an agent cannot simply perceive the fundamental physical elements of our world. Reasoning and inference, whether deductive or inductive, are unnecessary if

[3]One might wonder in what sense the best of a broken lot is ideal, but this issue can be dealt with later.

everything is already known.[4]

One might object that there are cognitive problems where perceptual bounds do not seem particularly relevant, and in these cases, ideal norms can be established. Call this the irrelevance of perception objection. Chess, for example, involves discreet board positions so obviously discernible that perceptual capacity is not a factor in gameplay.[5]

First, even in this case–beyond temporal boundedness–one has to hold fixed facts about perceptual abilities. Players who are visually impaired or agree to play chess blindfolded need to rely on moves communicated audibly while holding the state of the board in their minds. Thinking through one's moves in this situation changes the cognitive task. Perceptual parameters are almost always relevant to rational problems.

Second, playing chess is about predicting your opponent's future moves in light of the fact that perfect chess is unattainable for you both. Different players play differently and that means there are often more effective ways of playing against a particular opponent. The ability to "read" your opponent is central, not peripheral, to the game. Recognizing an opponent's playing style and tells can be of extraordinary value. Even when playing over the internet, the timing of an opponent's moves can provide valuable information, if you can discern it, about their intended strategy—such as whether they are enacting a pre-planned set of moves or struggling to improvise on the spot. Is this perception? Sometimes; not

[4]There may be empirical facts that an agent does not know, despite having unbounded perceptual access to the entire past, present, and future of the universe. Consider the traveling salesman problem. An agent who can perceive all the facts about the distances between cities does not necessarily know the shortest route between them all. Arguably, if an agent does not know the shortest route, then their perception is in fact bounded, since part of perception involves apprehending the relationships between objects. It depends on what kinds of relationships among the objects of perception are involved in perception.

[5]Again, being good at chess is a case of instrumental rationality (where the cognitive goal is to win), but these points apply equally to epistemic rationality. It may be that the irrelevance of perception objection is weaker for epistemic rationality, but that, if anything, makes my case stronger. For example, there may be limited cases where cognitive strategies are genuinely insensitive to perceptual capacities. If one is playing tic-tac-toe, since this game is solved, the solved strategy is optimal no matter what is supposed in terms of perceptual capacity. It is not clear that there is an analogous case one can point to for epistemic, as opposed to instrumental, rationality. If one's notion of ideal rationality only works for solved games–which are not much of games at all–then this is a very unfamiliar notion of ideal rationality.

always. Such capacities and deficits can arise from many distinct sources. But certainly, if you could literally perceive your opponent's thoughts, the game of chess would be quite different. Bounds on perceptual capacity have to be set.

Norms governing reasoning are meant for beings dealing with epistemic problems, and facing problems already implies a sort of imperfection—that the beings do not already know, and must be guided by norms of reasoning to attempt to know. Those who already know everything need not reason. In order not to know, an otherwise epistemically ideal agent must at least have limited perceptual access to the world. Such bounds on perceptual capacity constitute cognitive limitations. In what follows, I argue that such cognitive limitations cannot be principledly separated from other kinds of cognitive limitations that give rise to non-ideal epistemology.

1.5 Philosophical Conceptions of Rationality Cannot Separate Perceptual Constraints and General Cognitive Constraints

Consider the following condition:

Sensitivity: If epistemic norms are sensitive to cognitive limitations (like memory or attention constraints), then such norms are non-ideal.

This condition states that if what you ought to do is conditional on what you can do, then such norms are non-ideal. Such norms are non-ideal because they adjust to less-than-optimal capacities. Ideal norms should be absolute, not adjusted for individual capabilities. When memory limits lower the bar for what is expected of your reasoning, such standards become non-ideal.

Accepting the Sensitivity condition and the conclusion argued for in the previous section–that epistemic norms presuppose perceptual constraints–yields the conclusion that all epistemic norms are non-ideal norms. Just as memory limits lower the bar for what is expected of your reasoning, perceptual limits (e.g., blurrier vision) lower the bar for what is expected of your reasoning (e.g., what you can infer about your environment). Thus, epistemic norms, which are sensitive to perceptual capacity, are non-ideal. This conclusion follows from a straightforward reading of the Sensitivity condition. The Sensitivity condition needs to be modified or added to resist this conclusion.

In this section, I examine ways for philosophers to exempt perceptual constraints from rendering epistemic norms as non-ideal. That is, how philosophers can defend:

Not Perception: If epistemic norms are sensitive to perceptual limitations (like visual acuity), then such norms do not lose their ideal status.

This sets up a target that the rest of the paper aims to bring down. For accounts of rationality to yield Not Perception, they must restrict the scope of rationality. However, there are narrower and broader views on the scope of rationality. To avoid a loss of generality of my conclusions, I will consider views of rationality on either end of this spectrum so that my claims hold as widely as possible. I also focus here only on epistemic rationality for ease of exposition, though I take the points to generalize beyond epistemic rationality.

According to a narrower view of the scope of rationality, rationality concerns reasoning qua a special kind of inference. Analysis of the rationality of an agent, properly speaking, concerns only the agent's reasoning of this kind. One clearly articulated version of this view comes from Boghossian (2014, 2019). Not all transitions between mental states count as reasoning. Mental states can cause other mental states without such mental processes counting as a case of reasoning. It is unclear what makes a causal chain of mental states a

case of reasoning. However, it is important to differentiate the two because "I can be held responsible for the way I reason, but not for what associations occur to me. I can be held responsible for what I establish as a good reason for believing something, but not for what thoughts are prompted in me by other thoughts" (Boghossian, 2019, p. 15). To distinguish between the two, Boghossian argues that reasoning involves a "taking" condition. In overly simplified form, a person must "take" B to follow from A, for B to count as an inference from A, where taking is a personal-level doing (even if there is no associated phenomenology). While Boghossian's view concerns personal-level inference, it is sensible to discuss inference in this specific sense at subpersonal levels (Drayson, 2012; Frankish, 2010; Quilty-Dunn and Mandelbaum, 2018). Thus, the narrow view of rationality as inference can be further divided into an even narrower view (personal-level inference) and a broader view (personal and subpersonal inference).

According to the broader view of the scope of rationality, epistemic rationality concerns the proper management of one's doxastic states more generally, not just the ones attained through inference. This more general management of beliefs can be considered a form of reasoning. In what follows, I will use the term 'reasoning' in a way that is neutral to the narrower (reasoning as inference) and broader (reasoning as general management of beliefs) views.

Even broader views of the scope of rationality are perfectly sensible. Consider, for example, an alethic meta-epistemology (accuracy-first epistemology being a notable variant). Norms of rationality, in this view, are meant to promote a good: true belief in the case of epistemic rationality, and the attainment of other cognitive goals in the case of practical rationality (Wedgwood, 2017). Coherence, for example, is not an end in itself, but a means to an end (Joyce, 1998, 2009; Williams, 2013). Given such a view of rationality, one could

plausibly hold that epistemic norms govern any and all aspects of a cognitive system. Rationality would then concern more than just the management of beliefs or other intentional states; it would concern the nose-to-tail management of entire cognitive systems so as best to achieve cognitive (e.g., epistemic) goals. Given that increased perceptual capacities result in the better achievement of epistemic ends, Not Perception would be false. I am concerned here with exploring how philosophers can defend Not Perception, and ultimately arguing that such attempts will not work. So, I will restrict my focus to 'broader' views of rationality that are not so broad as to make Not Perception false. Going forward, by the term 'broad view' I mean the view described in the previous paragraph.

Note that both the narrow and broad views are compatible with a reason-responsiveness view. One might think only mental states that are or can be based on reasons are rationally evaluable. One could adopt a narrow view and hold that only personal-level inferences are the kinds of states that can be based on reasons, or are the only kind of mental state based on reasons that are appropriately rationally evaluable. Or one could adopt a broader view and take all and only (possibly subpersonal) doxastic states to be able to be based on reasons and rationally evaluable.

Given these general views about the scope of rational evaluability, one could attempt to argue for Not Perception by noting that perception and perceptual limits are outside the domain of rational evaluability. Thus Sensitivity does not apply to perceptual limits the way it does for other cognitive limitations. Perceptual limits might result in cases of ignorance– e.g., ignorance of the predator hiding in the bushes because you cannot perceive infrared light. But ignorance is not the same as irrationality. Call this the Evaluability Objection.

It should be noted that the view that perceptual states are rationally evaluable is becoming increasingly popular (Jenkin, 2023; Munroe, 2021; Siegel, 2016). For example, Siegel

(2016) argues that beliefs can influence perceptual states and make us rationally responsible for perceptual states. Jenkin (2023) has argued that even excluding the possibility of cognitive penetration, perceptual learning shows that perception can be based on epistemic reasons, and can therefore be epistemically evaluable.

I will not assume for my argument, however, that perceptual states are rationally evaluable. I view this issue as somewhat orthogonal to the issue I am concerned to address. For my concerns, the more important point is that the Evaluability Objection is a non-starter. Paradigmatic cognitive limitations that result in non-ideal norms–the cognitive limitations that the Sensitivity condition unambiguously applies to–are also not themselves rationally evaluable. Consider memory capacity limits. Memory capacity limits on narrow and broad views are not rationally evaluable. The memory capacities of a cognitive system reflect psychological facts characterized in non-intentional terms. Memory capacities and limits are not personal level inferences, they are not doxastic states, they are not (at least in general) responsive to reasons. Nonetheless, memory constraints can obviously affect what personal level inference or what doxastic states are possible and take place. For example, if one ought to update their beliefs in conformance with Bayesian norms, this may be impossible if one has constrained memory, since Bayesian inference is for many tasks with realistic stimuli intractable. It is not the memory limitation that is irrational from the perspective of ideal rationality (and non-ideally rational from the non-ideal perspective). It is the non-Bayesian reasoning that results that is irrational or non-ideally rational. The Evaluability Objection therefore engages in a category mistake. Cognitive limitations that lower the bar for rationality and transform epistemic norms into non-ideal norms are not the kinds of things that are rationally evaluable in general according to standard views of rationality.

What seems to be needed to establish Not Perception is instead something like the

following:

Subservience: If epistemic norms are sensitive to non-evaluable states that subserve reasoning (narrow view) or belief states (broad view), then such norms are non-ideal.

Memory capacities, but not perceptual limits, subserve reasoning. Reasoning, it seems, is further downstream from perception. It is not that, in general, cognitive limitations that affect how one ought to reason result in non-ideal norms. It is only when norms are sensitive to constraints on the cognitive machinery that directly realizes reasoning that norms become non-ideal. It is when epistemologists consider *these* limitations to make a difference to how one ought to reason that such standards become non-ideal. Thus Not Perception is true, since perceptual limits do not affect the cognitive processes that subserve reasoning processes.

Now that the proper target of my argument is in focus, I will argue against Subservience. Subservience is the most plausible defense of Not Perception. Since Subservience is false, Not Perception is also false. My arguments against Subservience will also illustrate why Sensitivity holds generally for all psychological particularities, including perceptual limits, rendering all epistemic norms non-ideal.

The basic idea is this: Subservience would be viable if irrationality were always a case of bad reasoning. But irrationality often involves 1) a breakdown of reasoning or doxastic states (i.e., non-reasoning and non-belief), or 2) good reasoning that nevertheless results in a person being irrational because of how reasoning processes interact with other cognitive subsystems that do not subserve reasoning. In both cases, sub-inferential or subdoxastic psychological facts are not subserving reasoning or belief, even though they result in irrationality. Accepting Subservience therefore allows epistemic norms to be sensitive to such limitations while maintaining status as 'ideal.' But such norms are blatantly non-ideal. Thus Subservience is false.

Consider first the narrow view. It is a difficult and important question: what makes a chain of mental events a case of reasoning? It may be that personal-level reasoning requires Boghossian's taking condition. However, Boghossian also believes, as seemingly most people do, that we can be blamed for any of our beliefs, not just the ones produced through reasoning in this narrow sense. Perhaps we can only fulfill our epistemic responsibility by reasoning our way to every belief we hold (aside from those beliefs that are justified by perceptual states, if you like). But we are still held accountable if we fail to do so. *Failures* of rationality can occur without reasoning. If a new thought pops into our head through mere association (or by being hit on the head), and we believe it, we are held accountable precisely because we generated a belief through a non-reasoning process. Failure to reason is itself an irrationality. So, personal-level inference may be how we attain rationality, and that may motivate the need for a narrow notion of reasoning. But irrationality can, it seems, occur without personal-level reasoning.

In fact, if irrationality could only occur as a case of bad personal-level reasoning, it seems strangely difficult to imagine how agents could ever be irrational. Can we imagine someone (who is committed to classical logic) "taking" (in Boghossian's strong sense) both P and $\neg P$ to be true? Or do they instead inadvertently take, say, P and Q to be true and fail to recognize that Q entails $\neg P$? If the latter, then the irrationality is not a result of personal-level reasoning; it is a breakdown or lack of personal-level reasoning. We hold the individual, the person, responsible. However, the irrationality is due to subpersonal, often subdoxastic processes. Is it possible to think of a case where someone's personal level reasoning is irrational, as opposed to a person being irrational because they fail to reason? Suppose they take P to be true, take $P \rightarrow Q$ to be true, and then mistakenly take $\neg Q$ to be true. For this to count as reasoning on Boghossian's view, they would need to take P to

be true, take $P \rightarrow \neg Q$, and then take $\neg Q$ to be true. But does anyone ever actually take in this strong sense $P \rightarrow Q$ and $P \rightarrow \neg Q$? Or do they take P to be true, mistakenly take $P \rightarrow \neg Q$, and then take $\neg Q$ to be true? Taking $P \rightarrow \neg Q$ even though in fact $P \rightarrow Q$ is not a case of mistaken personal level reasoning–unless one takes $P \rightarrow (P \rightarrow Q)$, but this only regresses the problem (does anyone actually take $P \rightarrow (P \rightarrow Q)$ and $P \rightarrow (P \rightarrow \neg Q)$?).[6] Reasoning is, quite sensibly, demanding. But the consequence is that irrationality is often non-reasoning.

On this point, there are persuasive arguments to the effect that it is impossible for reasoning to be fully explicit. Jackendoff, for example, takes a line of thought coming from Carroll, Wittgenstein, and Lashley, and constructs a new argument for the conclusion that: "It's logically and psychologically impossible to achieve the ideal of purely explicit rational thought. What we experience as rational thinking is necessarily supported by a foundation of intuitive judgment. We need intuition to tell us whether we're being rational!" (Jackendoff, 2012, p. 213). Models of rationality *cannot*, according to this argument, be restricted to explicit reasoning. Explicit does not mean conscious or personal. This problem also occurs at the subpersonal level, with explicit subpersonal representations (Quilty-Dunn and Mandelbaum, 2018). One primary issue here is that meaning plays a role in whether inferences are legitimate. So, correct inference depends on the machinations of the subpersonal, subdoxastic systems that generate meaning in our minds. This subdoxastic machinery plays an indispensable part in determining if our inferences are legitimate. So not only does irrationality often involve the breakdown of reasoning, but it directs focus to subdoxastic and

[6] A good candidate framework for understanding how to attribute inconsistent beliefs to an individual comes from D. Lewis (1982), Stalnaker (1984), Elga and Rayo (2022)'s work on fragmentation. This framework provides a way to characterize epistemic and decision-theoretic norms for limited agents (agents with limited access to information in their own minds). Such models are sensitive to the exact nature of an agent's limitations, and so my main point holds: all epistemic norms face Carr's problem and are a form of non-ideal epistemology.

extra-doxastic processes that are "involved" in reasoning and the management of belief, but not directly subserving such processes.

Consider now the broader view. There are distinct traditions in philosophy that hold that intentional states presuppose rationality.[7] It will not be necessary to accept this claim. All that is needed is a much weaker claim that these traditions have illustrated with their arguments–namely, that irrationality frequently involves the breakdown of belief. Suppose a person assents to a particular sentence. But suppose the person does not believe any rational entailments of the proposition expressed by that sentence and does not rationally act in accordance with that belief (other than assenting to it when asked if they believe it). It seems such a person only believes such a proposition in a very thin sense. Dennett, for example, has argued that when we attribute irrationality to an agent, to that extent, we move from belief talk to belief-like talk–intentional attributions generally become qualified. He gives the example of a child working their lemonade stand, selling a cup of lemonade for the listed price of 11 cents, being handed a quarter, and mistakenly giving a customer the wrong change of 12 cents. The child believes they returned 12 cents, that the lemonade costs 11 cents, that they were handed a quarter, that a quarter is 25 cents, and that they returned the correct change. But does the child believe that 25-12 = 11? It seems wiser to instead think that something went wrong at the subpersonal, subdoxastic level. Dennett's analysis of the situation is that when irrationality occurs, the intentional stance breaks down and one must drop down to the design stance. Something "wrong" has occurred at the low level of mechanical information storage, retrieval, manipulation, etc. Often irrationality cannot be accounted for in terms of intentional attributions. It is instead a breakdown of belief that forces us to analyze subdoxastic processes that are not in such an instance subserving

[7]One such lineage comes from Davidson (2003), Dennett (1989), and D. Lewis (1974). Another comes from, for example, Bilgrami (2008), Brandom (1994), and Kripke (1982).

beliefs.

Suppose now you think reasoning involves following a rule, and reasoning correctly involves following the correct rules (e.g., Broome, 2014). Sometimes, irrationality might occur as a result of someone following incorrect rules. However, irrationality might also involve following correct rules but still making a mistake because of interactions or interferences with other cognitive subsystems. Call such cases performance errors, as opposed to competence errors. Performance errors here can be understood as a case where a system possesses a competence (say, a subsystem that correctly follows an apt rule), but other subsystems give rise to interferences that produce mistakes (Jackendoff, 2009). This can take the form of an interruption or breakdown of a rule-following process. For example, suppose someone knows how to follow a correct inference rule but becomes distracted by external stimuli (due to attentional constraints) or cannot regulate their emotional states and comes to infer incorrectly. If reasoning is simply rule-following, then such an error would not count as an instance of reasoning. It must be understood as either non-reasoning (a breakdown in the rule-following process) or interference from other cognitive subsystems that affect the correct execution of reasoning.

Performance errors might also happen *downstream* of correct reasoning processes. That is, cognitive subsystems separate from reasoning systems can malfunction, leading to irrational outcomes without directly interfering with or disrupting the reasoning processes themselves. For example, perhaps a new belief is formed through an apt inference rule, but processes involving the storage of such a belief go haywire, leading it to be altered or forgotten. Alternatively, when a person tries to introspect and articulate their belief, the introspective process itself might not change the belief but could lead to the conscious awareness and reporting of a slightly altered version. Although the person's subpersonal reasoning

faculty may continue to operate with the correct version, psychologists eliciting a person's beliefs will judge such a person as irrational. It is not clear that psychologists would be wrong in such instances. If irrationality can occur due to processes downstream of reasoning and belief formation processes, then it would seem irrationality can occur due to processes upstream as well, in the form of perceptual processes.

Let's take stock. The ideal epistemologist wants to abstract away from any sort of cognitive particularities in characterizing correct reasoning. This includes the kinds of performance errors just discussed. But they do not and cannot abstract away from particularities of perceptual systems that affect what reasoning ought to occur. The challenge is then to say why perceptual processes are different from other cognitive particularities. The proposal is Subservience: perceptual processes do not "participate" in reasoning. But given the above discussion, an immense amount of irrationality happens because of processes outside of the cognitive processes subserving reason–either because irrationality involves the breakdown and absence of reasoning or belief, or because reasoning requires interaction with other cognitive subsystems.

The consequence of this line of thought is that accepting Subservience, while it would maintain a class of epistemic norms as nominally ideal, gives up on ideal rationality in practice. Accepting Subservience allows non-subserving errors to be accommodated into ideal norms. What an agent ideally ought to reason would be able to depend on the particularities of their general cognitive system.

For example, if one is engaged in a lengthy proof, memory limitations that result in forgetting earlier steps can be viewed as not impugning personal-level reasoning. One took (in Boghossian's sense of taking) something to be the case earlier, but in forgetting, is no longer taking it, and so is not responsible for using it in inference (Boghossian, 2019;

Burge, 1993). Thus, forgetting is something that happens outside of reasoning–bumping up against memory limitations does not result in bad reasoning, but excuses one from reasoning. According to Subservience, ideal epistemic norms can accommodate forgetting. But now the kinds of proofs that one is (ideally!) expected to do depend on the peculiarities of one's memory capacity. Multiple steps of inference cannot be prescribed by ideal epistemic norms without specification of memory capacity limits. It is not even clear that highly complex single steps of inference can be prescribed by universal norms, since one can forget crucial contents within a single inferential step. Accepting Subservience, therefore, results in epistemic norms that are inescapably tied to general cognitive limitations that are the purview of non-ideal epistemology. Carr's problem is reintroduced: what personal-level reasoning or what beliefs are expected of us depends on cognitive limitations, and these cannot be set in a non-arbitrary manner. Such 'ideal' norms are inescapably relative to arbitrarily set cognitive bounds. Accepting Subservience, therefore, results in ideal norms that look little different from non-ideal norms.

I conclude that Subservience is false. When epistemic norms are sensitive to psychological processes that do not subserve reasoning or doxastic states, such norms are still non-ideal. While it seems at first simply a confusion to think that perceptual limits are relevant to rationality, upon scrutiny, the ways in which they are "not relevant" to rationality are no different from general subpersonal, subdoxastic, and extradoxastic psychological processes, which nevertheless affect how one ought to reason and render epistemic norms as non-ideal.

Forming judgments based on subpersonal perceptual processes is no different from the general case of forming judgments based on subpersonal cognitive processes. Consider, for example, Boghossian's view of perception as it relates to rationality (Boghossian, 2019). Perceptual states, in his view, are non-propositional mental states that can justify a belief

state such as "a round tower there." We are not responsible for subpersonal perceptual processes themselves (that is, these may not be rationally evaluable). But such psychological processes give rise to mental states that serve as the basis for personal-level judgments for which we are held accountable. According to Boghossian, if we are aware of a systematic bias in our perceptual system, we have a responsibility to avoid forming judgments based on these biases. Once you are aware of the Müller-Lyer illusion, as Boghossian notes, you are on the hook for concluding that the two lines are different lengths. Our rationally evaluable judgments, it seems then, stand in the same relationship to perceptual processes as to general sub-intentional cognitive processes. We may only be responsible for our judgments, not the subpersonal cognitive processes that give rise to mental states that are the basis of such judgments (e.g., intuition, taking something to be true, taking something to follow from something else). But when we are aware of a systematic bias or limitation in our cognitive system, such as a disposition to commit certain logical fallacies or be overconfident, we are responsible for avoiding forming mistaken judgments based on these biases.[8] It is no different from the Müller-Lyer illusion case. Perceptual limitations stand to rationality in the same way that general cognitive limitations that render rationality as non-ideal do.

Finally, consider a reason-responsiveness view, where one is responsible for correctly using available reasons.[9] What reasons one has available depends on subpersonal cognitive limitations. That is, if reasoning proper is "revising one's beliefs or intentions, for a reason" (Wedgwood, 2006, p. 660) then what one ought to reason depends on what reasons one has available, which depends on the peculiarities of one's cognitive system. One is not responsible for computing the square root of twelve-digit numbers: when asked a question, the answer to

[8]See Dorst (2023) for the connection between overconfidence and irrationality.

[9]If, instead, ideal norms involve using reasons regardless of whether they are available to one's mind, then norms seemingly should require immediate inference to all truths, as all true propositions can serve as reasons.

which requires computing such large numbers, one is not in possession of reasons to say either way. One should plausibly say "I don't know" in such cases. But what about computing the square root of smaller digits? At what point does one become responsible? It depends on how cognitive limitations are set. When the number is small enough, it seems appropriate to say, "come now, you *can* calculate the square root of 196, just try a little harder." At what point this can be said will vary from cognitive system to cognitive system. Whether or not someone can be criticized on the reason-responsiveness view depends in part on whether a person is able to form a reason, and that depends on the particularities of their cognitive powers. If people are off the hook for all cases where their cognitive systems do not supply them with a reason, then what one ought to infer depends inescapably on what one can infer, making it a case of non-ideal epistemology.

Epistemic norms can abstract away from some cognitive constraints, but not all, lest epistemic norms simply demand omniscience.[10] The cognitive constraints not abstracted away from still render epistemic norms as non-ideal–they must be left in place to lower the bar from omniscience. Perceptual constraints, I have argued, look no different from general cognitive capacity constraints that lower the bar. The upshot is that talking about what we ought to do with our beliefs or reasoning simpliciter is insensible. It is always what we ought to do with our reasoning and beliefs, given our various non-intentional psychological capacities.

[10]At least on an alethic meta-epistemology (accuracy-first epistemology being a notable variant). Norms of rationality, in this view, are meant to promote a good: true belief in the case of epistemic rationality, and the attainment of other cognitive goals in the case of practical rationality (Wedgwood, 2017). Coherence, for example, is not an end in itself, but a means to an end (Joyce, 1998, 2009; Williams, 2013).

1.6 "Ideal" Norms are Sensitive to Perceptual Capacity

In what follows, I will present a series of cases illustrating ways in which epistemic norms are relative to perceptual and general cognitive limitations considered together. The point in these cases is not that we should think of people as irrational because of perceptual capacity constraints. The point is that it is arbitrary to take a different attitude toward perceptual constraints from other general cognitive limitations traditionally associated with non-ideal epistemology. It is arbitrary to treat epistemic norms as ideal when they abstract away from general cognitive limits, but keep in place perceptual limits.

This first example shows that what counts as an ideally rational strategy can be highly sensitive to psychological facts about an agent's perceptual capacity:

> Jenna is visiting the Galapagos. One day she gets a nasty sting from a member of an undiscovered species of cobalt colored insect. She starts collecting specimens of this insect and realizes only some of these insects have small hidden stingers on their underside. In an attempt to avoid being stung, she starts trying to figure out how to identify which insects have stingers on the basis of their gross morphology. She realizes there are a variety of cues that seem to make it more likely (but not certain) that an insect has a stinger, such as number of horns, shape of abdomen, markings, size, and so forth. As a Good Bayesian Reasoner, she learns how to categorize insects by weighting these observable cues appropriately. Eventually she reaches a high, but not perfect level of performance—the best possible performance based on the probabilistic relationship of the cues and the presence of a stinger.

> Maddy is visiting the Galapagos. One day she gets a nasty sting from a member of an undiscovered species of cobalt colored insect. She starts collecting specimens of this insect and realizes only some of these insects have small hidden stingers on their underside. Maddy has slightly lower intraocular pressure in her eyeballs than Jenna, which allows her to have slightly better blue-yellow color discrimination (Li et al., 2022). She realizes that all and only the insects with stingers are cobalt, while the others are actually ultramarine colored (though she doesn't have the exact names for these at her disposal). From then on, her ability to identify the stinging variety is infallible.

What "ought" someone do when confronted with the problem of inferring which insects sting? What would the "default" data available be? Any choice of available data presupposes a particular perceptual capacity. There is no non-arbitrary way to set such capacities. Attempts to abstract away from perceptual capacity altogether by imagining an idealized perfect capacity typically circumvent inductive problems altogether—for example, by making the identification of stinging insects a deductive matter. An example from the history of science: at the beginning of the 17th century, astronomers were faced with an inductive problem of deciding between a Ptolemaic, Copernican, and Tychonic model of the solar system. But when Galileo used a telescope to observe the phases of Venus for the first time, it deductively ruled out the Ptolemaic model, thereby changing the inductive problem altogether. Even more extremely, if Claudius Ptolemy had access to a spaceship in Alexandria, the problem of identifying the relative motions of the planets would have been fully observable and deductively determined.

What counts as ideal inference and reasoning is generally highly sensitive to what perceptual capacities are held fixed, in the same way, that non-ideal inference is sensitive to which cognitive capacities are held fixed. There is no default way to set either perceptual or cognitive capacities–one must simply specify these as a matter of setting up an epistemic problem. Further, in supposedly ideal epistemology, it is not the case that one always ought to employ the same inferential strategy given some problem and environment (e.g., Ptolemy with a spaceship does not need to induct). It is not as if perceptual capacity simply affects the quality of the data that goes into a single ideal epistemic strategy, affecting the outcome. The ideal strategy itself is dependent on perceptual capacity; what one ought to do, even in ideal epistemology, is dependent on what one can do. The next section shows an even more apparent case of this.

1.7 Perceptual and Cognitive Capacities Interact

Suppose now a slightly different scenario.

> Brendan and Nicholas are visiting the Galapagos facing the same inductive problem as Jenna (the Good Bayesian Reasoner). While Jenna possesses the cognitive ability to fully calculate and integrate the weightings of multiple cues in her mind, Brendan and Nicholas are limited to processing only one or two cues at a time. Due to their cognitive constraints, they start using a heuristic inferential strategy known as "take-the-best" (TTB) which involves learning to rely on the most distinguishing cue and ignoring all the others (Gigerenzer and Brighton, 2009). They find that the markings on the insects are the most distinguishing cue, and so start to infer whether an insect stings on the basis of the markings alone, ignoring the number of horns, abdomen size, etc.

TTB, inasmuch as it ignores information, deviates from Bayesian norms of inference. It is the case, however, that when the relationship between cues and what is being inferred is noisy enough, TTB can outcompete Bayesian inference (Gigerenzer and Brighton, 2009). TTB, by focusing on the most distinguishing cue and ignoring others, can avoid overfitting noisy data. Gigerenzer et al. call this the "less-is-more" effect–heuristic methods can be more accurate, not just relative to bounded computational capacity, but full stop compared to more cost-consuming methods like full-blown Bayesian inference.[11] The less-is-more effect is generally discussed as a phenomenon that arises when cues in the environment have a noisy relationship to what they predict. But here I construct an example where the source of the noise of the cues arises from perceptual capacity constraints, not the external world.

Suppose the insects have small markings that look similar to letters. Stinging ones tend

[11] But see Parpart et al. (2018) who argue that such heuristics can be understood as Bayesian inference with extreme priors, and that therefore heuristics do not in fact ever outcompete Bayes–at most they simply can do equally well. This is not an issue for my argument as the target in my discussion is not Bayesian norms. My point is that different methods can be better in certain contexts, where the contrast between methods is doing TTB (either understood as ignoring the other cues, in which case it is computationally manageable, or else as understood as having extreme priors) versus doing a typical multiple regression on all of the cues.

to have markings that look like C's, R's, and Z's. Non-stinging ones tend to have markings that look like O's, K's, and Y's. Brendan has astigmatism in both eyes, making his vision subpar, and causes him to sometimes mistake C's for O's, R's for K's, and Z's for Y's. As Brendan samples insects to learn which ones sting, it is possible that this eyeball defect introduces enough noise in the cues to make the TTB strategy more effective than a full Bayesian approach, relative to Brendan's perceptual capacity. Nicholas, on the other hand, has 20/20 vision. So for him, Bayesian inference is ideal but unattainable due to his cognitive limitations.

Both Brendan and Nicholas have the same non-perceptual, general cognitive capacities. But given the alethic epistemic goal, Brendan is still doing the best he can relative to his perceptual capacities, whereas Nicholas is not. If perceptual constraints were viewed as irrelevant to epistemic rationality (that is, if Not Perception is true), then Brendan would be considered ideally rational and Nicholas would not be, despite the fact that they are using the same cognitive strategy–simply because Brendan's vision is worse. Nicholas would be able to become ideally rational by blurring his vision.

Imagine another scientist, Paige, who also has astigmatisms, but has unbounded computational cognitive capacity. Paige can perfectly implement Bayesian inference, but if she does, she will do worse than Brendan, who has impaired cognition. She *ought* to implement the heuristic method and not Bayesian inference, even though she can do both. The same is not true for Jenna, who has the high cognitive capacity of Paige and the high visual acuity of Nicholas. She should compute Bayes. Brendan successfully outcompetes Bayes, while Nicholas does not, solely because Brendan's vision is worse.

This example shows a number of things. First, this is a case of an interaction effect—in the technical sense of a variable's effect depending on another variable—between perceptual

	Low Cognitive	**High Cognitive**
Low Perceptual	**Heuristic** (Brendan), Bayes	**Heuristic**, Bayes (Paige)
High Perceptual	Bayes, **Heuristic** (Nicholas)	**Bayes** (Jenna), Heuristic

Table 2: Cognitive and perceptual capacities interact to determine what an agent ought to do. Methods that perform better are listed first in each cell. Methods that are the best possible given both kinds of constraints are in bold.

and cognitive capacities. What an agent ought to do depends on an interaction of what they can do, perceptually and more generally cognitively. Second, it is a case where the ideality ordering itself is sensitive to perceptual facts. That is, low versus high perceptual capacity changes whether or not Bayesian or heuristic methods perform better, regardless of cognitive capacity.

If one wanted to specify ideal epistemic norms in this scenario—ones that abstract away from cognitive limitations as much as possible—one would still need to set the nature of the data, and this implies facts about an agent's perceptual access to the world. Abstracting away from perceptual limits altogether makes the ideal strategy deductive: do what Maddy does and infallibly identify stinging insects by seeing ultramarine, or else use x-ray vision to see the small stinger on the underside. To generate a problem where familiar ideals such as Bayesian inference are ideal, one needs to invoke and hold fixed perceptual constraints. This means invoking substantive ought-implies-can principles—ones concerning perception. You cannot perceptually discriminate ultramarine, therefore it is not the case that you ought to infer deductively.

Why can't ideal epistemologists reply that ideal norms are conditional on the data

available–given X data, do Y inferential strategy? The problem is that allowing conditional norms in this way risks elevating all of non-ideal epistemology to ideal epistemology. All inferential strategies become ideal, conditional on certain constraints. But, it might be replied, it is possible to distinguish between ideal and non-ideal norms by the following. No matter what the data is, Bayesian inferences can and should ideally be employed (where deductive inference and heuristics like TTB are both understood as special cases of Bayesian inference). On the other hand–this line of thought goes–given certain other kinds of cognitive constraints, Bayesian inference might not be possible, rending whatever non-ideal strategy making such best-conditional-on-constraints strategies non-ideal. This response loses sight of the fact that in an alethic epistemology, there is nothing in itself special about Bayesian inference. It is only a means to an end. Non-Bayesian strategies can still maximize alethic goals, conditional on various constraints.[12] So if ideal norms can be specified conditionally, in order to differentiate between ideal and non-ideal epistemology, one would need to differentiate between kinds of conditions. I.e., one would need to justify why perceptual constraints do not compromise ideal epistemology, while other cognitive constraints do. So, we are back where we started.

In characterizing ideal norms, what justifies ignoring and abstracting away from computational capacity constraints, while holding fixed perceptual capacity constraints? As the example shows, having unbounded computational capacity is neither necessary nor sufficient for ideal performance. Such facts about perceptual capacity are intertwined with other sorts of cognitive capacities regarding what one ought to do. What is the "absolute" ideal in this case? It is not, given the interaction effect, simply some single method achievable when one has unbounded computational capacity. Ideal norms cannot be solely specified in relation

[12]And they need not always approximate Bayes either (Icard, 2018). Karlan (2021) calls such strategies "local rational maxima."

to an environment by conceiving an idealized computational agent capable of executing a perfect inferential strategy. There seems instead only to be "ideal" norms relative to particular psychological capacities–both perceptual and general cognitive capacities–considered together at once.

I have constructed examples where the relevant difference in perceptual capacity is determined at the far extreme of the sensory system—in the first case, which concerns blue color discrimination, eyeball pressure, in the second case, eyeball curvature. This choice is to emphasize the precarious demarcation of what counts as "internal" to the cognitive system and "external." These physiological facts about the eyeballs are fully outside of the nervous system; they are not even directly involved in transduction of physical stimuli information into electro-chemical nervous information (what the retina does). But the same difference in information access to the world—the nature of the data available for the cognitive system—could just as well have been determined by differences "deeper" into the nervous system, or further out into the world.

That is, the noisiness of the cues could have just as well resulted from the environment itself–as is usually the case with the less-is-more effect–as opposed to people's perceptual systems. The relationship of the insect's morphological features (the cues) and having a stinger are already *ex hypothesi* noisy—as a result of mother nature's distribution of traits in the population. In the thought experiment, the noise could be adjusted in the relevant way to make even 20/20 vision people better off utilizing TTB. The noise could also come from deeper into the nervous system: blurry vision caused not by astigmatism but by neuropsychological deficits. The example in the following section is such a case.

What an agent ever in any case "ought" to do all depends on the data determined by the structure of the world, as well as perceptual, and other cognitive capacities–these all

interact, and are not clearly separable. Does it matter if the source of the noise in the data is coming from the world, the eyeball, or computational cognitive deficits deep in the mind? Why only abstract away from this last kind in characterizing ideal epistemic norms? Why should idealized epistemic models abstract away from some psychological facts, but not others? Treating norms as ideal because computational constraints are abstracted away from, while other constraints are still in place, appears arbitrary. All of these constraints are environmental or psychological facts acting as impediments to epistemic aims in the same way.

1.8 Perceptual and Cognitive Capacities Impede Reasoning in the Same Ways

Not long ago, positional astronomy made use of a technique known as the "ear and eye" method. This involved 1) looking through a telescope and watching a celestial object cross over a line on the lens. For example, an astronomer would look through a telescope that has been positioned so the center cross-hair lines up with the celestial meridian, and watch as a planet crossed this line and thereby transited the meridian. This technique involved 2) hearing a clock tick out the time. By noting the time to the minute on the clock before putting their eye to the lens and then listening to the ticks count out subsequent seconds, astronomers could coordinate the ticks heard with the planetary transit seen. This technique allowed for precise observation of astronomical positions.

It turned out, however, that data collected by this method from different astronomers varied considerably but displayed low variance within each individual's measurements. That is, while the timing recorded for a transit differed notably from one astronomer to another, each astronomer consistently exhibited either an early or late bias in their recordings. This

observation led to the development of the personal equation, which accounted for an individual's characteristic bias, and allowed for correction of this source of error in the astronomical data.

One could imagine a hypothetical situation in which the personal equation was not developed, and in which this source of error was not corrected for—and in which no objective method of observation, free of individual bias, was available. One could further imagine that individual astronomers were competing with one another to produce accurate models of astronomical motions on the basis of their individual observations. Employing "equally good" inductive methods, astronomers whose "personal equation" happened to reflect better calibration would produce more accurate models. Once bias-free methods of observation allow for interpersonal assessment of the different models, the astronomical community eventually would determine that these lucky astronomers produced ones closer to the truth.

Should all of these astronomers be evaluated as equally epistemically ideal? According to the view I'm arguing against (Not Perception), the answer to that question would depend on the source of the bias. If ideal epistemic norms can leave in place perceptual constraints, but not other cognitive constraints, then the astronomers' status as ideal reasoners depends on whether their biases arise from perceptual constraints. If the biases arise as consequences of more general cognitive constraints, then they are not ideal reasoners.

It turns out personal equation individual differences are a result of reaction time differences (Boring, 1929). Psychologists in the 19th and early 20th centuries investigating this phenomenon showed that there were, in fact, various psychological sources entering into the length of the reaction time. For example, Wundt's laboratory showed by adding complications to experimental tasks that there were separable sources of reaction times that compound together in complicated ways. Such sources of delay included factors at the outer

sensory periphery (including the retina), connections among the sensory systems, as well as more central, cognitive factors resulting in expectation differences. These sources turn out to interact in non-linear ways (Boring, 1929).

Consider two astronomers with the same personal equation, who gather identically biased astronomical data. One's bias is due to psychological constraints arising in the outer sensory periphery, and the other's bias is due to general cognitive computational constraints. If perceptual constraints are distinguished from other constraints, then the first astronomer is ideally rational and the second is not, even if they both end up with the same positional information, make the same inductive inferences, and reach the same model conclusions.

The point here is *not* to show that perception and cognition have no clear separation, or that cognition penetrates into perception. The point is that the same sorts of constraints on an inductive problem can arise in different areas, and often emerge as a (possibly non-linear) interaction among constraints in multiple areas. This makes disentangling different kinds of constraints–the perceptual ones needed to generate an epistemic problem in the first place, versus the computational resource constraints seemingly irrelevant to ideal epistemology–challenging to say the least.

Perception and other cognitive capacities can also be subject to the *same* bounds. Perceptual inference faces the same computational tractability problems as other kinds of cognitive processes (Brooke-Wilson, 2023). Bounds on perceptual capacities such as visual acuity can be due to not just astigmatism but also computational memory and processing limits that may arise as general features of a cognitive architecture–resulting in, say, the need to approximate instead of fully computing an inference. Cognitive processing limits, for example, make it difficult to discern visual input moving quickly around your visual field (such as discerning the C and O markings on the Galapagos insects when flying by your face). So,

the same perceptual constraints I discussed in the Galapagos examples can arise from either perception-specific qualitative constraints (astigmatism) or domain-general computational bounds. It is, I claim, arbitrary to consider someone irrational simply because their visual deficit arose from domain-general computational constraints, rather than astigmatism. All of these kinds of bounds are entangled and act as impediments to reasoning in the same way. This further supports the idea that ideal norms cannot be specified by abstracting away from some bounds but not others.

Why criticize as irrational agents who are bounded in certain cognitive capacities, but not others? Bounds related to perception and other kinds of cognition are all similar sorts of bounds that arise from various psychological facts. Without a clear epistemic motivation, it is arbitrary to abstract away from some of these and not others when characterizing ideal norms.

1.9 Zetetic Epistemic Norms Imply Cognitive Limitations

Assuming the goal of epistemic rationality is getting at the truth, it is possible to characterize norms that govern nose-to-tail cognitive systems in pursuit of that goal, and not just the management of belief states. One might limit what persons are epistemically responsible for, but as §5 argued, both subpersonal perceptual and cognitive constraints ultimately fall on the same side of that line. Still, one might think that it just is the case that epistemic rationality concerns only the relationship between beliefs and therefore, perceptual constraints are not relevant. According to views such as evidentialism, being ideally rational means inferring beliefs in the best way given whatever information you have (Conee and Feldman, 2004). It would then follow that constraints relevant to perceptual access to the world,

such as visual acuity, can be fixed in any way one likes. This is true even if the psychological constraints on perception originate not from peripheral sensory facts, but from the same sources as other kinds of cognitive constraints, such as general memory and processing limits. What matters is that ideal norms do not presuppose any constraints on cognitive capacities relevant to the belief formation processes downstream of perception (assuming one can demarcate these).

The arguments in this paper so far have already provided reasons in favor of considering perceptual constraints as having the same status as cognitive limitations that paradigmatically give rise to non-ideal norms. This section will provide a new angle, discussing how the issue at hand depends on whether or not you think there are genuinely epistemic norms of inquiry (norms concerning actions to seek out information, also called zetetic norms Friedman, 2020).[13] Zetetic norms have traditionally been thought of as norms of instrumental rationality instead of epistemic rationality. Instrumental rationality is hypothetical rationality–what one ought to do *if* one has a particular cognitive goal. The idea is that we do not have purely epistemic obligations to gather information. Otherwise, we would be rationally required to read Wikipedia all day (Hedden, 2015). We are only required to do so for instrumental reasons. *If* we want to learn about a subject or achieve some material goal, then it might be obligatory to inquire into the world and gain information.

There is, however, a growing movement in epistemology that accepts zetetic norms as epistemic norms.[14] If one adopts an alethic meta-epistemology, then epistemic norms are about achieving the end of truth. Inquiry is an important part of finding the truth. If the

[13]There are other ways to resist the view that rationality concerns only doxastic relations, without invoking zetetic norms. Siegel (2016), as mentioned, has argued that perception can be considered rational or irrational, depending on the influences of perception.

[14]This is sometimes called the zetetic turn. Note that one can accept that some zetetic norms are epistemic norms, without needing to accept the stronger claim that all epistemic norms are zetetic norms (Friedman, 2020).

point of rationality is to be good at getting to the truth, then having more information available will aid in that end. If the aim is truth, then you epistemically *should* read Wikipedia all day, as having more information will aid in your inferential capacities (for example, knowing the environments in which a heuristic method will outperform a non-heuristic method.). We balance such epistemic ends with other non-epistemic ends (eating, sleeping, desiring to socialize), which is why we are instrumentally rational not to spend all day on Wikipedia (Flores and Woodard, 2023).

Zetetic norms are not generally thought of as incompatible with ideal epistemology. Carr accepts epistemic zetetic norms as part of ideal epistemology for instance (Carr, 2022). Zetetic norms might be demanding (you should spend all day on Wikipedia), but being impossibly demanding is compatible with being an ideal norm (consider utilitarian norms about maximizing utility, or computationally intractable Bayesian norms).

But what I have tried to argue is that in the setup of an epistemic problem, bounds on perceptual capacity must be set. Such bounds presuppose psychological facts, and it is not clear how to set such bounds in a non-arbitrary manner. If there are norms of inquiry, then you should gather evidence. But what evidence you can gather is determined by the bounds on your perceptual capacity. It seems then, that what you zetetically should do implies facts about what you can do in the form of contingent and arbitrarily set psychological facts. Should you obtain information about the color of the insect? Should you obtain information about the markings on the insect? Should you obtain information about all the features of the insect? Should you use your x-ray vision to discern immediately the presence of the stinger? Such norms depend on what your perceptual capacities are like. And so if zetetic norms are properly epistemic norms, then what you epistemically should do in any situation depends on your cognitive limitations.

This all might seem like even more reason to reject zetetic norms as properly epistemic. But there are some drawbacks to this response to my argument.

First, rejecting zetetic norms as properly epistemic abandons large areas of rationality to my conclusions. In particular, this leaves instrumental rationality, which concerns hypothetical norms about what means one ought to take to achieve their ends, vulnerable to the conclusions drawn here. One might not be obligated by a norm of epistemic rationality to gather evidence. But if one has the cognitive goal of knowing which insects sting, or has the goal of not being stung, then as a matter of instrumental rationality, zetetic norms are reintroduced. If one has the goal of knowing something, then they ought to inquire. Such zetetic norms then imply substantive cognitive limitations that cannot be set in a principled way, and the ideal-non-ideal distinction breaks down for this domain.

Practical reasoning–reasoning about what one should do–can be understood as a "hybrid virtue" involving both epistemic and instrumental rationality (T. Kelly, 2003). Decision theory is a theory of practical reasoning. So note that blocking my argument by noting that epistemic rationality concerns only doxastic attitudes, abandons decision theory to the idea that there is no ideal-non-ideal distinction.

Further, even if epistemic rationality only concerns relations of belief, there is *some* activity that does involve the interleaving of evidence gathering and belief formation. Kelly, for example, calls this theoretical reasoning, and sees this as also a hybrid virtue of epistemic and instrumental rationality (T. Kelly, 2003). It involves a back and forth process of inferring from your current data, and figuring out what data to collect, in order to improve your epistemic situation. Treating this as a hybrid virtue is a way to defend the idea that there are no epistemic norms of inquiry while recognizing that inquiry is often a necessary part of a process of reasoning. If one chooses to deny properly epistemic zetetic norms, this kind of

theoretical reasoning is also left vulnerable to my arguments.

So while one might stipulate that "true epistemic rationality" only concerns doxastic attitudes, there is still another kind of truth aiming activity that involves not just data processing but data collection and data processing working together. Generally speaking, I think psychologists interested in evaluating how well humans reason in the laboratory are, and should be, interested in this broader kind of reasoning. The Wason selection task, for example, is an examination of this kind of reasoning, since it asks agents to infer what information to seek out to answer a question (Wason, 1968).[15] Human performance on the Wason selection task has been a prominent battleground in the rationality wars (e.g., Oaksford and Chater, 1994; Stich, 1985). I do not think we should dismiss the task as irrelevant to epistemic rationality because it involves reasoning about how to seek out information, as opposed to just responding to information one already has.

Further, the examples I have given in this paper suggest that the line between zetetic norms and non-zetetic norms itself is blurry, and gives novel support to the view that at least some zetetic norms are epistemic norms. Flores and Woodard (2023), in arguing that we do have properly epistemic norms of evidence gathering, point out that your access to the world and access to your own mind, and what evidence you have available in your mind, are not so different. Gathering evidence from the world and trying to figure out what evidence you "already" possess in your head is not as different as it might seem. My examples support this idea in a new way. In the case of the personal equation in astronomy, for example, it is simply not clear whether biased data is a result of the data collection or data processing phase. Once "entering into consciousness" is no longer seen as the demarcation between the

[15]The Wason selection task involves cards that have letters on one side and numbers on the other. Subjects are asked which card(s) they would need to turn over in order to verify a rule, such as that "if there is a D on one side, there is a 5 on the other." Typically, but only in certain contexts, subjects have difficulty when the task requires inferring the contra-positive of a conditional statement. See Ragni et al. (2017) for a recent review.

two, the point where data collection ends and data processing begins becomes unclear both psychologically and epistemically, and is not particularly meaningful.

These points also dovetail with a point that has been made by bounded rationality theorists–organisms are embedded in their environment, meaning that adaptation to the environment also entails adaptation to one's own body. This includes adapting to physiological limitations such as brain capacity, which directly influence cognitive abilities (Simon, 1955; Icard, 2023). The point is that the lines between the mind, body, and world are not particularly salient when it comes to what one ought to do.

What is the difference between improving your epistemic situation by improving your doxastic relations versus improving your perception (putting on your glasses)? If you can increase the accuracy of your beliefs by simply putting on the glasses that are hanging around your neck, then you should. Note also that norms of inquiry are not just about forming *beliefs* about what information to sample. It seems reasonable to criticize someone for correctly forming the belief that they could have more accurate beliefs if they put on their glasses, but then not acting and putting on their glasses. If you are crunching a labor intensive computation to infer something that you could simply learn by turning around and looking, then you should turn around. If epistemic rationality concerns only the correct processing of data, then one seemingly could maximize epistemic rationality by shutting themselves in a dark room.[16] Doing so is antithetical to alethic goals, but to resist this conclusion, it seems necessary to impose positive obligations, including the proactive gathering of data. Once this conclusion is drawn, perceptual limits become impediments to epistemic obligations.

[16]See Miracchi (2019) for more on such issues.

1.10 Beyond Ideal Epistemology

Imagine witnessing a child struggling to learn the names of colors. Perhaps you see them failing to identify red and green correctly. A legitimate question is whether they are "bad" at learning colors or whether they are color blind. One might think that a perceptual deficit does not impugn human rationality. Most people would not consider a color-blind child to be irrational or dumb. But is a child who is not color-blind and who struggles with learning colors irrational or dumb? From an alethic standpoint, both cases involve psychological details that prevent better attainment of the truth. Cognitive deficits in some contexts might be more mutable than perceptual deficits–and therefore, in some sense, might be seen as more epistemically blameworthy (you could have done better, where could is a much closer modality). But this suggests that what really matters is the mutability of psychological facts and not, first and foremost, the distinction between perceptual and cognitive constraints. The mutability of psychological facts will cross-cut general cognitive constraints and constraints on perceptual access to the world. And as discussed in §3, likely there is no hard line between the mutable and immutable.

Facts about cognitive limitations, I argue, are no different than any other of the number of facts that need to be fixed in order to specify an epistemic norm. Cognitive limitations generally (including perceptual limitations) are simply part of the description of an epistemic problem–just like the other sorts of facts that need to be held fixed, such as the environment, what task an agent is being evaluated on, and so on.

There is no normatively privileged way to decide what epistemic problem to hold fixed to generate a norm. Whether one focuses on the problem of inferring the probability that Linda is a bank teller or focuses on the Monty Hall problem will generate different claims

about what you ought to do. This does not threaten to undermine in any way the norms that emerge in these different contexts. There is no epistemically privileged way to set such parameters—they are all simply components of the setup of the problem. Once the problem is set up, one can then ask what one ought to do—what's the best that can be done given the circumstances. In the same way, including cognitive limitations as part of the setup does not undermine the value of epistemic norms. In fact, given my arguments, it is *necessary* to specify relevant cognitive limitations, in order to be able to set up an epistemic problem and specify epistemic norms.

The arguments expressed here dovetail nicely with arguments made by Kelly (e.g., K. Kelly, 1996). Kelly argues, for instance, that the problem of uncomputability just is "the problem of induction internalized." An unbounded search through the world looking for a counterexample to the claim that all ravens are black is epistemically similar to an unbounded search through the natural numbers looking for a counterexample to a mathematical conjecture. If there is no counterexample, you will continue with your search forever, never assured that a counterexample will not eventually emerge. If there is a counterexample, eventually it will be found. The fact that one search goes on through the senses and out in the world, and one search happens in the head is not what determines the difficulty of the problems or what sorts of success are possible (see also K. Kelly and Schulte (1997)).

Kelly's work also shows that characterizations of the difficulty of inductive problems generally are structurally the same as characterizations of the difficulty of computational problems (e.g., the arithmetical hierarchy). For example, recursive, r.e., and co-r.e. sets are analogous to decidable, verifiable, and refutable inductive problems. Kelly's work uses topology to generalize the concepts of decidability, verifiability, and refutability to a complete hierarchy of inductive difficulty. Doing so shows that computational difficulty and inductive

difficulty "are reflections of the same sorts of limitations and give rise to similar. . . hierarchies of underdetermination" (K. Kelly, 1996, p. 160). The point is that inductive obstacles and computational obstacles are unified from an epistemic perspective. Inductive obstacles arise from perceptual bounds and computational obstacles arise from computational bounds.

Familiar models of ideal epistemology–such as those that hold Bayesian norms to express ideal norms–tend to ignore computational obstacles, while taking seriously and dealing with inductive obstacles. Models of ideal rationality treat methods as ideally rational because they do well on an inductive problem, abstracting away from bounds on computational capacity and problems of computability, but leaving in place bounds on perceptual capacity and problems of induction. But from Kelly's perspective, as well as the perspective developed in this paper, it is arbitrary to abstract away from one type of difficulty and not the other. Bounds on perceptual capacity that give rise to inductive problems are an impediment to ideal reasoning in exactly the same way that bounds on computational capacity are an impediment to ideal reasoning.

Complete abstraction from all constraints in epistemic problems is unattainable; one can abstract away from some, but never all constraints at once. Consequently, no ideal norms exist, as total idealization is not possible.[17] The only fully idealized epistemic norm is the expectation of omniscience.

I began this paper with the problem of distinguishing between a chess player doing their best given limited resources, versus one playing just plain badly. Everything I have argued in this paper reinforces Carr's point that there is no way to draw this line. I have, in fact, only made things worse. I have argued that this is also a problem for ideal epistemology, since there is no way to maintain an ideal non-ideal distinction in epistemology (since all

[17]This claim is compatible with Greco (2023), who argues that there must always be *some* idealization in characterizing epistemic norms.

norms are sensitive to psychological capacities). I would like to conclude with some thoughts on why it is possible to be optimistic, rather than despairing, about this conclusion.

I suggest embracing the idea that any fact about an agent's psychology can count as a constraint, and that all agents are trivially doing the best that they can relative to their constraints. Instead of determining the normatively privileged and correct place to lower the bar, and checking if an agent clears it, one should lower the bar *until* the agent clears it, then note how low the bar has been set. How much one needs to "lower" the bar then helps characterize the agent's rationality.

Lowering the bar is just a metaphor (and I do not mean to imply a linear scale). More accurately, one sets an evaluative standard, such as an alethic standard like belief accuracy, or a practical one, such as the amount of money won. One does not attempt to fix psychological facts about capacities, determine the best that can be done given such constraints, and then evaluate an agent on whether or not they clear this bar. One instead works to identify all of the constraints that an agent is subject to, preventing it from performing better according to the standard. One brings into view the bounded rationality of an agent by bringing into view all of its constraints–by understanding what is preventing the agent from doing better. Together, the evaluation itself (whether it won the game or not, how quickly, how much money was won, etc.) and facts about the agent's psychology preventing them from doing better (liable to distraction, impatient to make a move) form a more complete picture of an agent's rationality.

While evaluative standards such as belief accuracy or amount of money won can yield clear quantitative scores, metrics such as "87% belief accuracy" or "217 dollars won in a board game" leave out valuable information in the evaluation of agents. Such assessments are more informative when understood in the context of whether it is possible to do bet-

ter, and possible relative to what constraints. Some constraints are more mutable than others. Facts about agents' psychologies can specify what constraints are changeable and, importantly, in what modality (after instruction, after practice, after microneurosurgery, in a world with different physics, etc.). Specifications of constraints yield important counterfactual information relevant to epistemic evaluation and prescription. These facts are essential to describe what agents *could* have done better and in what sense, and to prescribe what they *should* do, and in what sense. Epistemic could and should, as modal terms, need not be fixed. An analysis of the mutability of constraints that an agent is subject to allows for a rich analysis of the different ways an agent is, could be, and should be epistemically rational. No clear line must be drawn between different kinds of limitations to have valuable epistemic norms.

1.11 Conclusion

Epistemic problems usually suppose limited perceptual access to the environment. Such limited access entails particular psychological facts. Epistemic norms are sensitive to perceptual capacities in the same way that they are sensitive to other kinds of cognitive capacity constraints. Supposedly ideal epistemic norms consequently entail substantive cognitive limitations of the same sort as non-ideal norms. Philosophical conceptions of rationality, surprisingly, seem unable to exclude such cognitive limitations while maintaining ideal rationality. Without a way to distinguish between different kinds of limitations, the distinction between ideal and non-ideal norms is dissolved.

Non-ideal epistemology appears unable to distinguish between agents doing the best they can relative to limitations and being outright irrational. The breakdown between ideal and

non-ideal epistemology means that this is also a problem for supposedly ideal epistemology.

In light of these results, the best response, I suggest, is to allow for all psychological details to count as constraints, and to think of all beings as always doing the best they can relative to their particular limitations. Epistemic evaluation of an agent then involves characterizing how limited an agent is by specifying psychological facts that prevent them from doing better, and specifying how mutable these facts are. We are all limited beings. Bringing into view our limitations helps bring into view our rationality.

1.12 Post-Script: A Note on The Regress Problem

Before ending this chapter, I want to briefly discuss an ambiguity concerning resource rationality. If one supposes that there are costs associated with an inference problem, then one can ask what the optimal allocation of resources should be. For example, observing samples in a population incurs costs (time, money, etc.), but it also has benefits, as larger random samples are more likely to be closer to the population parameter of interest. Likewise, performing experiments provides valuable information for decision-making but can also be costly (Wald, 1947). At what point do the costs of further sampling or experimentation outweigh the benefits? When do agents achieve an optimal trade-off? This matter can be considered a kind of resource rationality (Manski, 2017).

This form of resource rationality, as Arrow (2004) suggests, gives rise to a "paradox": if calculating the first-order optimization problem is prohibitively expensive, then solving the second-order problem of optimizing the cost-benefit trade-off for the first-order problem will often be even more expensive. This situation arises because, in general (without knowing specific details about the task), solving the trade-off requires addressing the first-order

problem as a subtask. Any attempt to solve the second-order problem with an optimal cost-benefit trade-off results in the same problem at a higher level, leading to an infinite regress. For this reason, Arrow tentatively suggests that viewing bounded rationality as optimization relative to constraints might be problematic.

This worry is not a problem for my proposals. First, as Arrow acknowledges, when more is known about the structure of a specific first-order problem, the second-order problem often requires fewer resources. It can often be demonstrated that there are guaranteed diminishing returns after a certain level of resource investment, making the complete first-order task unnecessary to calculate (as is often the case in optimal design, see Shiryaev, 2007; Wald, 1947). Additionally, as has already been noted, in the case of random sampling, it may not be the calculation that is (computationally) costly, but rather the sampling procedure or iteration of a test that is (time and money) costly. In such instances, even if it were necessary to theoretically calculate the first-order problem for the second-order problem, this would not be as costly as actually performing the first-order task.

More importantly, the procedure I have outlined does not necessitate addressing the second-order problem. Instead of determining the optimal allocation of resources, the proposed method requires only determining the optimal strategy given a fixed allocation of resources, or determining the amount of fixed resources that make a strategy optimal. The claim is that agents are rational relative to their bounds, not that the bounds themselves are rationally set. While such calculations are not trivial, they do not require solving the unbounded first-order problem as a subtask, thereby avoiding any regress.

Finally, even if there were a regress as Arrow suggests, or if calculating optimal performance relative to constraints were prohibitively expensive, it does not pose a problem for my account, which aims to explicate what it means for an agent to be rational. Agents

can achieve optimal trade-offs without needing to calculate them, and they can be resource rational without proving the constraints relative to which their strategies are optimal. Such difficulties challenge evaluators and designers of cognitive systems, not the cognitive systems themselves. Nature may have endowed many organisms with optimally allocated resources, but the organisms, as well as we as evaluators, may never be able to confirm it.

There is evidence that human cognitive systems flexibly allocate resources—adjust their own constraints—to support more resource rational cognitive strategies (Lieder et al., 2018; Musslick and Masís, 2023a). Whether humans are optimally allocating resources is an interesting question, but even if they are not, they are still resource rational in my sense. Any suboptimality in resource allocation simply indicates further cognitive limitations.

2.0 Intentionality Presupposes Resource Rationality

Several traditions in the philosophy of mind hold that intentionality presupposes rationality. But what notion of rationality is invoked? I argue that my account of resource rationality provides the relevant way in which intentionality presupposes rationality while simultaneously making sense of how it is possible to attribute irrationality. In this chapter, I examine the debate between those who believe intentional systems must be rational and those who point out that cognitive creatures can and do commit blatant violations of rationality. I show that the former view is strengthened by acknowledging that intentional systems need not be ideally rational. Attributions of irrationality are possible, but only if it is shown how intentional systems are still doing their best relative to constraints. Finally, this perspective allows for an important bridge between intentional and non-intentional psychological analyses.

2.1 Introduction

Suppose you meet someone who claims to believe that taxes are bad. However, upon further conversation, you find that this person also believes that: 1) public spending is a great thing, 2) taxes are not theft, and 3) the government can and should increase taxes. They also claim to happily pay taxes and would not change tax policies. In fact, they regularly vote to increase taxes. Despite all this, they adamantly and enthusiastically proclaim that they believe taxes are bad.

Such examples suggest to some philosophers that belief involves more than mere assent to

a proposition. For a belief to be a belief, believers must also accept some rational entailments of that proposition and act in accordance with it. More generally, if agents are to have beliefs, desires, intentions, and so forth, these intentional states must rationally cohere with one another. Attributing irrationality, on this view, to that extent makes intentional attributions unstable. However, how much irrationality can an agent exhibit before they can no longer be said to hold a particular intentional state? Must people believe all rational entailments of their beliefs?

Philosophical analyses suggesting that intentional states presuppose rationality are contrasted by psychological research, which has identified apparent systematic human irrationalities. Noteworthy is the fact that attributions of irrationality are often made using intentional language. For example, we attribute irrationality to a person by pointing out that they believe that 1) rhinos are mammals, 2) all mammals are warm-blooded, and 3) rhinos are cold-blooded. It seems the person would not have inconsistent beliefs unless they held all of these beliefs as beliefs. Therefore, there is nothing strange about attributing irrationality to an agent by attributing intentional states to them.

I argue that the latter arguments do not, in fact, defeat the reasons for thinking that intentionality presupposes rationality–they just show that attributions of irrationality to intentional systems are possible. I thread the needle between these positions. I argue that the view that intentionality presupposes rationality is strengthened by acknowledging that intentional systems need not be ideally rational. Attributions of irrationality are possible, but only if it is shown how intentional systems are still doing their best relative to constraints. In other words, attributions of "irrationality" should be understood as specifications of constraints for resource rational agents. Thus resource rationality provides the relevant way in which intentionality presupposes rationality, while also making sense of how it is possible to

attribute irrationality.

In §2.2, I explore the necessity of a rationality assumption for intentionality. In §2.3, I argue against the requirement for ideal rationality. In §2.4, I discuss why minimal rationality is inadequate. Finally, in §2.5, I explain how resource rationality bridges the gap between intentional and non-intentional psychological explanations, allowing for a de-idealization of purely intentional characterizations into fully subpersonal information processing models.

2.2 The Necessity of a Rationality Assumption

Why think that intentionality presupposes rationality? A variety of claims have been made concerning normativity and the intentional, most clearly dating back to Kant. While there are metaphysical, semantic, and conceptual varieties of these claims, the general idea is that possessing intentional states constitutively involves normative commitments, which in turn require rationality in some sense on the part of the intentional system (Wedgwood, 2009).

There are two distinct lines of thought leading to the conclusion that intentionality requires normativity. The first line of thought descends from Kant (e.g., Bilgrami, 2008; Brandom, 1994; Kripke, 1982; Wedgwood, 2006). Kant held that the fundamental unit of consciousness is the judgment and that judging such and such to be the case commits one and makes one responsible. What it commits one to and what one is responsible for is determined by norms of rationality. A belief, on this view, is a judgment that constitutively involves rational commitments. This influence, for example, can be seen in Brandom's inferentialist semantics (Brandom, 2014). Brandom's semantics holds that the meaning of an expression is just what can be inferred from it, and what can be inferred from it depends on our

capacity for enforcing rules concerning making moves in the language game—a capacity for normative judgments grounded in a more general psychological capacity for normative judgments (Brandom, 2014).

The second line of thought is more behaviorist in nature and can be seen in the work of Quine and his students Davidson, Dennett, and Lewis. In this chapter, I leave aside the first line of thought and focus on the second. If you are amenable to the first line of thought, I suspect much of what will be argued for here will carry over and be convincing to you as well–but I make no promises and do not argue for this.

According to the Quinean line of thought, there is a difference between a mere set of sentences and a belief set, and it is the rationality of the agent that makes the belief set what it is. Consider, for example, a very minimalist assent theory of beliefs (e.g., B. Russell, 1919), which has no rationality restriction on what counts as a belief. On this view, believing a proposition consists of an agent simply having a feeling of assent toward a particular proposition at some time.

Dennett offers a thought experiment as an argument against this position. Imagine a neurological defect (Dennett imagines a tumor) that results in a person feeling assent toward a sentence and responding "yes" anytime anyone asks if they believe the proposition expressed by such a sentence (Dennett, 1975). If this "belief" is not integrated into the rest of the person's intentional psychology—e.g., it does not cohere with their other beliefs, the person also assents to opposing beliefs, the person does not assent to obvious entailments of the proposition, and does not act on the belief to attain their goals—then in what sense is this actually a belief?

This leads to an investigation into "the general conditions for the possibility of application of the concept" of intentional content (P. Griffiths, 1963, p. 19). An intentional ascription

should do more than simply redescribe the behavior(s) that served as its basis. Even if one ultimately wishes to cash out intentional content in behavioral terms–that is, allow mentalistic talk but only as shorthand that can be replaced with behavioral language–intentional ascriptions should, at the very least, support counterfactual claims about behavior in different circumstances. That is, most behaviorists aim to reduce mental predicates to dispositions to behaviors, not just behaviors.

For example, saying "that man wishes to remain dry" should do more than restate that the man is huddling under a tree in the rain (Dennett, 1968). Instead, it should support a prediction that, all things being equal, he would not be under the tree if it were not raining or if he had an umbrella. Otherwise, it's not even characterizing a disposition, it's just describing the behavior itself. If intentional content is doing no work, it could be discarded in favor of the original, more parsimonious behavioral descriptions, making the translation into intentional language unnecessary in the first place. This is the crux of Skinner's point that mental attributions are poor explanations if they are mere descriptions of behavior (Skinner, 1977). Perhaps understanding how physiological mechanisms give rise to dispositions to behaviors is interesting, but what purpose does the mental play?

So intentional content should do something–have a role–in the machinations of the mind. Few theories of mental representation hold that there is no requirement for downstream effects of a representation—and not merely that the representation is produced in a certain way. In other words, intentional content is efficacious and not inert in the machinations of cognition. Intentional ascriptions should specify details that make a difference to cognition, not posit inert representations that sit idly in a belief box or a desire box.

Mental content can, of course, be causally efficacious in ways that do not have anything to do with rationality. For instance, consider Davidson (1980)'s example of a climber holding

a rope that his friend is holding onto, and having the thought that if he let go, his friend will fall. The thought is so abhorrent that it shocks him and causes him to let go of the rope. The belief causes the action. But this, it seems, is an aberrant way for a belief to cause another mental event. The only role of beliefs cannot be to cause other mental events in this way. The role of beliefs is not to cause you to do things by shocking you. The role of beliefs is to provide a basis for rational inferences to further beliefs, all which can serve as a basis for actions that are rationally in line with your beliefs and desires. If a belief cannot serve as a basis for such rational relations, then it is not a belief.

A belief can cause a free association of an another belief, but it seems insensible to think that a belief might only exist in the mind to cause other beliefs through association. The constitutive role of beliefs is to allow inference to other beliefs, where inference involves taking it to be the case that a belief rationally follows from another belief (Boghossian, 2014). Beliefs cause other beliefs, properly speaking, when the latter are rational entailments.

Consider: believing that mammals are warm-blooded and that rhinos are mammals causes the belief that rhinos are warm-blooded. Versus: believing that mammals are warm-blooded and that rhinos are mammals causes the belief that rhinos are cold-blooded. What does it mean in the second case for the beliefs A and B to cause C? It is perfectly sensible to tell a physiological story where the physiological states corresponding to beliefs A and B caused the physiological state corresponding to belief C. But it does not seem sensible to simply say that the beliefs A and B caused C. It seems in such cases, there is a breakdown of belief processes that forces one to drop down to the non-intentional level to explain such causation (Dennett, 1989).

In addition to these points, there is another purely epistemic concern. There is no epistemic basis for making claims about mental content unless one assumes the agent is

rational. As Cherniak (1990) states, "A cognitive theory with no rationality restrictions is without predictive content; without it, we can have virtually no expectations regarding a believer's behavior" (p. 6). Without the ability to directly observe intentional content and without a well-supported theory of how intentional content is processed by subpersonal mechanisms, the Quinean perspective suggests there is only one way to infer intentional content from observational data and vice versa: through a rationality assumption. The only way to use "person A believes X" to predict person A's behavior is to assume they are rational and will believe the rational entailments of that belief and act in rational accordance with it.

Relatedly, Dennett makes a point concerning the "holism" of intentionality. The argument is intended to show that intentional content cannot be reduced to behavioral dispositions, because intentional content is always connected to other intentional content in a way that makes this impossible. The connection between intentional content is a rational one:

> Quine and Chisholm also present arguments about believing and intending, of which the central point is that efforts to provide behavioural analyses of these two phenomena are doomed by a vicious circle of implications. Take, for example, the belief that it is raining. What behavior would clinch it that A believes it is raining? No matter what is suggested, it will turn out that this is a clincher demonstrating that A believes it is raining only if we assume that A has some particular purpose or intentions. [...] A's finding a tree or roof to stand under is no more evidence, for it depends on A's intending to stay dry. If ascription of belief always depends on an assumed ascription of intention, the converse holds as well. A's intention to stay dry is not behaviorally demonstrated by his cowering under the tree except on the assumption that he believes it is raining, that he believes that he would get wet if he did not stay under cover, and so forth. A survey of the other Intentional and mongrel Intentional idioms shows that the use of any one of them has implications about beliefs and intentions, so the circle that prevents a behavioural paraphrase of belief and intention sentences infects the whole realm of the Intentional. Dennett, 1968(pp. 31–32).

Without the rationality assumption (or charity as Davidson calls it), there can be no attribution of content to another mind, and by similar reasoning, even to our own minds

(Davidson, 2003).[1] There is no other apparent candidate for a constraint on attributions of mental content beyond rationality assumptions, which bridge observable behavior to the intentional. The exception might be inferring beliefs about perceptual states. For example, if a person has a cup in their visual field then it is straightforward to suppose that the person has a belief that there is a cup in front of them. But to go beyond beliefs that are closely tied to perception, one starts to need to employ the rationality assumption.[2] Absent such a constraint, seemingly anything would go in terms of attribution. Of course, even with rationality assumptions, intentional attributions are almost always underdetermined. Likewise, intentional attributions underdetermine predictions of behavior. But, on this line of thought, this is all there is. It might not be much, but it's the best available.

For Dennett and Davidson, there is also a Rylean line of thought that supports their rationality assumption view (Dub, 2015). The idea here is that conceptual analysis shows that intentional ascriptions presuppose rationality. Dennett uses the example that "conception leads to pregnancy" leaves no room for a further question of why a particular conception led to pregnancy—it would not count as conception in the first place unless it led to pregnancy. Similarly, a belief or intention leading to an action or another belief leaves no room for an analogous further question. It would not count as an intention to leave the room, for example, unless, ceteris paribus, one left the room because of this intention.

In general, Ryle and those following his line of thought find the entire category and purpose of intentional attributions to operate according to norms of rationality. It is for this reason that Ryle is skeptical of a science of psychology that attempts to give explanations

[1] Radical translation, radical interpretation, and mind-reading are all analogous problems on this view.

[2] If one does not adopt a narrow view of rationality as merely reasoning, but instead considers it as a broader normative assessment of cognitive systems, then the inference in the cup case does presuppose rationality. Suppose a person has a cup in front of them and you infer that the person believes there is no cup in front of them. This attributes to them a mistake–i.e., an illusion–in the same way that attributing the desire to get wet to a man huddling under cover in a rainstorm attributes to that man an error.

of mental processes in a way that is not either in pure causal-physiological terms or else common everyday folk psychology:

> When we hear the promise of a new scientific explanation of what we say and do, we expect to hear of some counterparts to those impacts... some forces or agencies of which we should never have dreamed and which we shall certainly never witness at their subterranean work. But when we are in a less impressionable frame of mind, we find something implausible in the promise of discoveries yet to be made of the hidden causes of our own actions and reactions. We know quite well what caused the farmer to return from the market with his pigs unsold. He found that the prices were lower than he had expected. We know quite well why John Doe scowled and slammed the door. He had been insulted. (Ryle, 1950).

Dennett follows suit when he points out that even if we theoretically attain a mature science of the mental or understand the exact causal processes by which mental machinations occur, our ordinary intentional explanations will still hold true and will not be explained away. "Why did Mary blush?" Dennett asks. "Because she knew he knew her secret" (Dennett, 1989). Such an explanation, for Dennett and others, only works if we assume Mary is rational. This answer will remain correct regardless of the future advancements in cognitive science. So, while the Quinean point is somewhat of a "best game in town" argument–suggesting that the rationality assumption is the best available means of predicting and interpreting intelligent behavior–the Rylean point is that even if there were a better alternative for prediction and interpretation, something would still be missing without considering the rational analysis. Even with a molecule-by-molecule prediction of Mary's blushing behavior, one misses an important truth if one ignores the rationality-assuming intentional story.

Finally, the study of non-human animal cognition provides a helpful lens to show how intentional content cannot be separated from an assumption of rationality. Reflecting on how to distinguish different degrees of higher-order intentional states in non-human animals

strongly suggests that intentionality presupposes rationality. Dennett (1989) gives the example of primate warning calls and intentional attributions of different degrees (i.e., mental states being about other mental states):

> (0) Tom reacts to seeing a leopard by giving the leopard call (which results in Sam running into the trees).
> (1) Tom wants to cause Sam to run into the trees.
> (2) Tom wants Sam to believe that there is a leopard and that he should run into the trees.
> (3) Tom wants Sam to believe that Tom wants Sam to run into the trees.
> (4) Tom wants Sam to recognize that Tom wants Sam to believe that there is a leopard.

The primary challenge is scientifically testing which of the above accurately characterizes Tom's intentional content. How could a cognitive ethologist perform an experiment to determine this? The only possible way would be to change the parameters of the situation, for example, by seeing if Tom still makes the call when Sam is not around. If he does, this would suggest that Tom does not want to cause Sam to run into the trees after all, but only assuming that Tom is rational. Such scientific reasoning depends on Tom being rational. There is no other option.

This is not merely an epistemological or practical scientific question of how to test intentional hypotheses. Instead, it reveals that what actually differentiates these different contents is the downstream consequences of these intentional states, and these consequences can only be understood as rational processes.

One might find all of these arguments unconvincing. It could be argued that most of these points do not demonstrate the metaphysical conclusion that having intentional states requires being rational. Instead, they suggest, if anything, only the weaker epistemic claim: the only way we can *come to know* the mental states of others is by adopting a rationality assumption. Perhaps even this is too much for you. In that case, the conclusion is that *one* important way to come to know the mental states of others is by adopting a rationality assumption. The

rest of the paper can then be understood as follows: given that rationality assumptions are either metaphysically necessary, epistemically necessary, or just epistemically useful, what notion of rationality is involved?

2.3 What Notion of Rationality Is That?

In this section, I discuss different views of rationality that intentionality might presuppose. Most philosophers of mind have not specified the exact notion of rationality that underlies intentionality. Folk attributions of intentionality naturally presuppose a folk psychological theory of rationality. But what kind of theory is that? Furthermore, if intentional attributions are to exist within a scientific psychology, a non-folk psychological theory of rationality is needed. What kind of rationality is that? Philosophers of mind have largely left this determination to epistemologists. Davidson, for example, speaks broadly of the norms of logic and decision theory (Davidson, 2003).

However, leaving the determination to epistemologists without their explicit involvement only sows confusion. For example, while philosophers of mind argue that what is needed is ideal rationality, it is unclear whether the (formal) epistemologists who develop the theories philosophers of mind refer to are concerned with "ideal" rationality at all. Davidson cites DeFinetti when discussing the "ideal" norms of decision theory (Davidson, 2003). However, DeFinetti and others in the foundations of decision theory relax the requirement of logical omniscience (Finetti, 1974). DeFinetti's motivation for this move is to achieve greater descriptive accuracy for human reasoners, aiming to close the gap between normative and descriptive theories of reasoning, rather than addressing issues in the philosophy of mind and the presupposition of the intentional (Kadane et al., 1999).

The interaction between these different literatures appears to be more crosstalk than dialogue. Among contemporary work in formal epistemology, there are those attempting to delineate ideal norms, but the specific connection between these ideas and the issue of intentionality has not been explicitly drawn.

Thus, there is a need to develop a specific account of rationality tailored to the arguments that intentionality presupposes rationality. Here, I present a particular version of resource rationality. Resource rationality involves agents that are as rational as possible, given their cognitive limitations.

Resource rationality is part of a family of theories generally known under the label of bounded rationality. A boundedly rational agent exhibits rationality but is subject to material constraints that prevent it from being more rational. Bounded rationality stands in clear contrast to ideal rationality. However, there is a well-known ambiguity in bounded rationality concerning whether such an agent is performing the best it can relative to its constraints.

Consider two chess algorithms with identical constraints (e.g., how many branching possible futures they can consider before making a decision). It is not hard to imagine that one of these algorithms makes significantly better use of its computational resources to play superior chess. But as long as the other program plays decently enough, might we call both "boundedly rational"? Simon (1957), who introduced the term, is notoriously vague on this subject. While Simon's related concept of satisficing, for example, is contrasted with optimizing, satisficing can always be reinterpreted as an optimization problem (in which the objective function makes any option that achieves the set aspiration level equally optimal).[3] Simon's scissor metaphor of the two "blades" (the task environment and limits on cognitive

[3]There are contexts in which it can be shown non-trivially, however, that satisficing with particular aspiration levels is the boundedly optimal solution (Manski, 2017)

capacity) "shaping" behavior does not help resolve the ambiguity, since the metaphor leaves unspecified how these two factors exactly interact to influence intelligent behavior.[4]

Fortunately, the issue of whether satisficing is a form of optimization is not particularly important here. The point is that while ideal rationality is clearly meant to be a superlative notion (if you are ideally rational then you are maximally rational), bounded rationality can choose to keep or discard the superlative conceptual component.

However, I argue that the superlative notion must be retained in order to understand the rationality assumption of intentionality. To understand why this is important, it is necessary to consider the contrasting view, known as minimal rationality (Cherniak, 1990).

Cherniak's account of minimal rationality is bounded rationality that fully drops the superlative conceptual component. It involves the idea that some agents act *appropriately* given their belief-desire makeup, even if they could have done better. Minimal rationality is effectively a binary notion. It suggests a threshold between agents that are acting sufficiently "appropriately" and those that are not. Any agent in the class that clears such a threshold can be considered minimally rational.

In the following sections, I will argue why my account of rationality is preferable to both ideal and minimal rationality. For now, I will briefly note a problem with minimal rationality.

The problem is that there is rarely, if ever, a principled way to decide where to place the threshold between irrational agents and minimally rational agents. We can imagine a scenario with a clear objective function, where it is possible to rank different strategies from better to worse, at least in ordinal terms. How do we determine the cut-off point where suboptimality becomes an inappropriate strategy? For example, when driving to the grocery

[4]The task environment shapes human behavior in virtue of a rationality assumption: humans respond appropriately to the task relative to the environment. The two blades are, therefore, that humans are rational and that there are limits to this rationality (Simon, 1990).

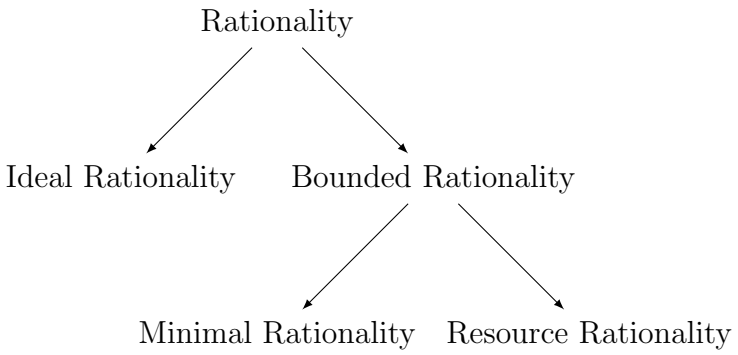

Figure 1: Relations Between Notions of Rationality

store, there are typically infinite routes, and often one fastest (ideal) route. Many routes that are not the fastest are still appropriate. But at what point does a route become so indirect that it is considered irrational instead of merely not ideal? Both ideal and minimal rationality seek a binary distinction: the former between ideal and non-ideal rationality, and the latter between rational and irrational.

In addition to ideal and minimal rationality, there is really only one further option to understand the claim that intentionality presupposes rationality: the view that agents are doing the best they can relative to constraints—i.e., resource rationality. In the driving example, constraints might include not being able to afford tolls, having a car so old and decrepit that it can't drive up hills, or needing to avoid a neighbor's house because you told them you were away and unable to pet sit their Chihuahua. Given these constraints, certain routes are ruled out, and there is a fact of the matter concerning which route is the fastest given such constraints.

For resource rationality to be viable, there must be a clear notion of what can count as a constraint. In the driving route case, perhaps only the routes your car cannot handle

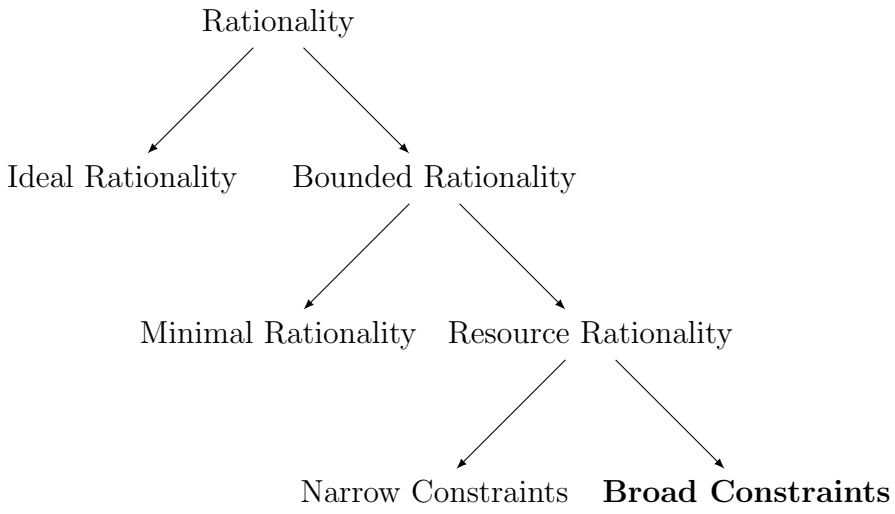

Figure 2: Relations Between Notions of Rationality

should be ruled out, while the cost of tolls should not be considered a constraint. In the case of general human reasoning, memory and attention limits are often considered constraints, but not perceptual limits. If one limits what counts as a constraint, then some agents are resource rational and some are not. However, I have argued in Chapter 1 for a maximally general notion of constraints. Any material facts about an agent's psychology that prevent them from better performance relative to an evaluative standard can count as constraints.

The push for this view comes from arguments showing that it is arbitrary to distinguish between different kinds of constraints to determine which agents are "truly" resource rational (Carr, 2022; my own in Chapter 1). The pull for this view is that it has several attractive properties. One such attractive property, as will be argued, is that it makes the most sense of the rationality assumption of intentionality.

On this view, all agents are trivially resource rational; they all do their best relative to their cognitive limitations. An analysis of an agent's rationality involves evaluating their

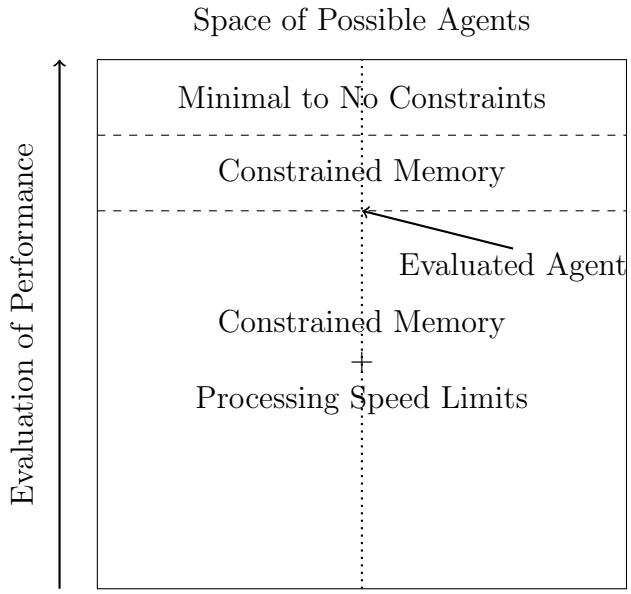

Figure 3: Bringing into view an agent's resource rationality: adding constraints to limit the pool of agents under consideration until the evaluated agent performs the best within the constrained pool. Carr (2022) uses a similar picture that uses Kratzer semantics.

performance relative to an evaluative standard and characterizing how constrained they are. To assess resource rationality, one can limit the pool of available cognitive strategies to those possible given various psychological constraints, such as a certain amount of memory space or an upper bound of processing speed. The resource rational cognitive strategy is the one that performs the best relative to the goal and environment among the pool of constrained agents. An agent's resource rationality is brought into view by comparing its performance among increasingly constrained pools of agents until the evaluated agent is performing at the top of such a pool.

One can still discuss an agent being more or less rational: agents are more or less constrained and, by definition of constraints, do better or worse relative to an evaluative standard. One can also analyze the mutability of constraints and use such analyses to discuss in what sense agents could have done better. For example, one could have performed a mental inference better if they had been less distracted (a more mutable constraint) and could have also done better if their brain's processing speed was much faster (a less mutable constraint).

Resource rationality combined with broad constraints yields a view that avoids making arbitrary distinctions at non-natural joints–e.g., between kinds of constraints, between agents acting appropriately or not, and between agents that are "perfect" yet still limited (ideal rationality). This approach not only allows for a detailed characterization of agents' rationality but, as I will argue in the next section, it also helps make sense of the rationality assumption for intentionality.

2.4 Why Not Ideal Rationality?

In this section, I argue against the idea that intentionality presupposes ideal rationality. Opponents of the view that intentionality presupposes rationality point to the empirical psychological literature showing that humans are systematically irrational. Stich (1985), for example, discusses the Wason selection task (discussed in Chapter 1). Stich emphasizes that even when the error is pointed out and explained, many participants adamantly hold to their incorrect beliefs. One might think this is a case where irrationality is not a breakdown of belief, but rather a situation where humans hold very real but inconsistent beliefs. There seems to be nothing contradictory about attributing irrationality in intentional language; it is conceptually consistent to say that an agent has inconsistent beliefs. But, Stich claims, if this is true, then it would show that intentionality does not presuppose rationality after all.

However, this does not show that intentionality does not presuppose rationality. It shows at the very least that intentionality does not presuppose ideal rationality. Unfortunately, because of ideal rationality's special status in epistemology and allure as the only kind of rationality that cuts normativity at its joints, demonstrating that intentionality does not presuppose ideal rationality gives the impression that intentionality does not presuppose rationality at all.

Beyond pointing out that attributions of irrationality in intentional language are perfectly sensible and empirically evident, Stich focuses much of his argument against the rationality assumption by addressing another line of thought not yet discussed, which suggests that intentionality presupposes rationality: the fact that minds are products of evolution. Sophisticated minds could only have emerged because evolutionary pressures selected for organisms with greater capacities for discrimination, offline representation, mental simulation, and rea-

soning (Dennett, 2004, 2008, 2017). Greater capacities for intentionality emerged because it made organisms more rational, which made them more evolutionarily successful. Implicit in this argument is the idea that rational and adaptive cognitive strategies are co-extensive (Stich, 1985). However, this is not always true.

Stich (1985) gives the following example: in signal detection tasks such as recognizing a predator in the bushes, the more adaptive strategy might be an overcautious one that prefers false positives, since false positives have little consequence. Such a strategy differs from the optimal decision threshold if the goal is pure accuracy of belief. Which strategy is more rational? While it is plausible to argue that the former strategy is more rational because it best serves the survival goals of the organism, Stich (1985) notes that:

> This is the reading which turns the conclusion of the argument from natural selection into a tautology by the simple expedient of defining *rational inferential strategy* as *inferential strategy favored by natural selection.* Quite apart from its *prima facie* implausibility, this curious account of rationality surely misses the point of psychological studies of reasoning. These studies are aimed at showing that people regularly violate the normative canons of deductive and inductive logic, probability theory, decision theory, etc. They do not aim at showing that people use inferential strategy which have not evolved by natural selection!

But why does Stich take the example from signal detection theory to show that humans are irrational? Having a lower decision criterion does not violate any normative canons of signal detection theory. It is true that a lower decision criterion, in some cases (depending on the base rates of signals and noise), results in less accurate beliefs. However, if Stich claims that humans are irrational because they employ a cognitive strategy that is suboptimal relative to the goal of having accurate beliefs, then he seems to ignore the arguments from the holism of intentionality discussed above. The rationality of intentionality involves the rational integration of general intentional states, including desires, and not just beliefs. This suggests that the accuracy of belief is not the primary type of rationality presupposed by

intentionality. Evaluating rationality solely based on the accuracy of beliefs would involve analyzing only beliefs, excluding other intentional states. Given that an organism desires not to be eaten by a predator, the instrumentally rational strategy *is* the one favored by natural selection. According to signal detection theory, if the costs of missed signals are high, then it is optimal to have a lower decision threshold.

Stich should argue instead that evolution primarily favors what is good for replication, not the cognitive goals (intentional desires) of the agent. This misalignment of what is being optimized is enough to show that the argument from evolution is not enough to necessitate the perfect (instrumental) rationality of an intentional agent.

There is no reason to think that evolution favors the *ideally* rational solution. Evolution is a satisficing process that prefers strategies promoting reproduction rather than primarily aiming for accurate beliefs or the attainment of cognitive goals. While the latter often promote the former, this is not always the case. Thus, while the argument from natural selection provides a compelling reason to think that intentionality presupposes *some* kind of rationality, it suggests that ideal rationality is not the appropriate notion.

In summary, the dialectic is as follows: proponents of the view that intentionality presupposes rationality provide good reasons to think that some rationality is indeed presupposed. However, opponents point out that irrationality is perfectly compatible with these arguments, suggesting that intentionality does not presuppose rationality after all. But these points do not generally refute the reasons for thinking that intentionality presupposes rationality. Instead, they show that intentionality can co-exist with a degree of irrationality. Thus, these points merely indicate that it cannot be ideal rationality that is presupposed by intentionality.

Why think *ideal* rationality is needed? Most of the arguments for a rationality assump-

tion seem neutral to the specific kind of rationality, and there are ample reasons to think that it is perfectly reasonable to attribute irrationality in intentional language. There is one motivating reason, however, to suggest that the kind of rationality must be ideal rationality. This is the Quinean point that irrationality of any kind is the best evidence that one has made a mistake in attributing intentional content to an agent. *Any* irrationality suggests that one has mischaracterized someone's beliefs or desires.

Consider again the primate call case. If Tom makes a call even when Sam is not around, it would be irrational, supposing that Tom makes the call because he wants Sam to flee from the leopard. But this irrationality is the best evidence that Tom does not actually have that intentional state in the first place. More generally, if we suppose Tom has an nth-order intentional state but displays irrationality, then that is the best evidence that Tom only has the n-1th intentional state, which would make Tom rational after all (though more constrained in terms of intentional capacity). Thus, it seems that intentionality tolerates no irrationality whatsoever. This leads Dennett to write:

> Conflict arises, however, whenever a person falls short of perfect rationality, and avows beliefs that either are strongly disconfirmed by the available empirical evidence or are self-contradictory or contradict other avowals he has made. If we lean on the myth that a man is perfectly rational, we must find his avowals less than authoritative "You *can't* mean–understand–what you're saying!"; if we lean on his 'right' as a speaking intentional system to have his word accepted, we grant him an irrational set of beliefs. Neither position provides a stable resting place; for, as we saw earlier, intentional explanation and prediction cannot be accommodated either to breakdown or to less than optimal design, so there is no coherent intentional description of such an impass.

"As we saw earlier" refers to the idea mentioned above: the motivation for needing ideal rationality stems from the notion that any irrationality whatsoever is the best—indeed, the only—evidence that one was mistaken in intentional attribution.

But—and here is the crucial point—this evidence is defeasible. While irrationality can be

evidence for a lack of intentional content, it can also result from intentional content combined with cognitive constraints. Consider again the primate call case. Tom making the leopard call when Sam is not around might indicate a mere reaction–a lack of intentional content–on Tom's part. However, it is perfectly reasonable to imagine that Tom typically makes the call because he does want to cause Sam to run and hide. In the case where Tom knows Sam is not around, he might simply lack the inhibitory control necessary to prevent himself from screaming out when he sees a leopard. This lack of inhibitory control is a cognitive limitation that can be independently studied and manipulated. One could imagine running Tom through a training course to develop his general inhibitory control, then testing him again. In this scenario, there would be no alert call when Sam is not around. This would suggest that Tom had the intentional state to begin with after all. Thus, irrationality is only defeasible evidence for the lack of an intentional state. Without knowing the constraints an agent is subject to initially, one must assume less constrained rationality to begin the process of interpreting an organism's behavior. But such an assumption of rationality can be de-idealized as constraints are uncovered and evidenced.

The fact that irrationality is only defeasible evidence for the lack of an intentional state defeats the reasons for believing that the type of rationality required is ideal rationality. It does not show that intentional content can co-exist with any amount of irrationality. The Rylean point is correct—intentional concepts are ineliminably connected to rational behavior. The point is that intentional concepts can still interact with non-intentional, physiological limitations. Why did Mary blush? Because she knew he knew her secret—and also because she had rosacea. Why did Mary not blush? That might be evidence she was not embarrassed, but it might also be that she was embarrassed but has anhidrosis, which prevents her from blushing. There's no further question of why conception leads to pregnancy, just as there's

no further question of why a person who believes there's a sandwich in the fridge and wants to eat it goes to open the fridge. However, one can still become distracted on the way to the fridge, and the irrational behavior that results does not undermine one's intentional content. In none of these cases are rationality assumptions entirely absent: on the given analyses, Mary would blush if she did not have anhidrosis, and the person on their way to the fridge would have gotten there if they had not been distracted.

Mind-readers such as ourselves are in the perhaps unfortunate position that we must rely on assumptions of rationality. Rational behavior is the only evidence that we got our mind reading right, and irrational behavior is more or less the only evidence that we got it wrong. But irrationality can also be turned into evidence for specific sub-intentional cognitive limitations. These limitations can then be incorporated into resource rational analyses that allow for continued prediction and interpretation of an agents behavior.

The earlier arguments show that beliefs cannot be fully inert mental states idly sitting in a belief box. If one believes no rational entailments of a belief, that is prima facie evidence that they do not actually have that belief. However, if one can characterize an agent as having that belief while also being subject to constraints that prevent them from believing certain rational entailments or always acting in accordance with that belief, then it seems acceptable to say they believe it despite their constraints. One could intervene on the constraints, thereby removing or lessening them, and result in more rational behavior. For example, consider a case of fragmentation where some beliefs are isolated from others (Elga and Rayo, 2022; D. Lewis, 1982).[5] In an extreme case, one could have a single belief isolated from all others. Such a person would seem not to have the belief because the belief is effectively idly sitting in a belief box. But consider the constraints—whatever psychological

[5]I take it not to matter in this particular instance whether human psychology, in fact, works like this as an empirical fact.

processes result in the isolation between that belief and the others. Elga and Rayo, for example, discuss attention shifts changing which information is accessible at a particular time, resulting in fragmentation (Elga and Rayo, 2022). If one manipulated this constraint and changed attentional mechanisms, thereby changing the accessibility between beliefs, and the person started to integrate the previously isolated belief into their general belief network, then it seems right to say that such a person did have this belief, even if it was subject to significant constraints.

There is no reason to demand ideal rationality and much reason to discourage it. The arguments in favor of a rationality assumption show only that the relationships among intentional contents and behavior must always be rational; they do not show that they must always be ideally rational. Beliefs and desires connected to one another in an entirely non-rational way seem not to be beliefs and desires at all. However, the assumption that such connections must be unconstrained rationality is defeasible. Irrationality can indicate constraints that can be factored into a rational explanation of one's intentional contents.

2.5 Why Not Minimal Rationality?

Much of the discussion so far suggests that minimal rationality is inadequate for the job. For example, the discussion leaned heavily on the idea that irrationality can be evidence of an intentional state subject to a constraint (as opposed to evidence of no intentional state). Constraints and intentional contents can coexist in a theory that connects an agent's behavior to their intentional contents. Minimal rationality lacks this conceptual machinery. Constraints exist outside the theory of minimal rationality. Within the theory of minimal rationality, there are only agents acting appropriately or not.

The main problem with minimal rationality is that attributing it to an agent fails to provide the predictive and explanatory power needed for intentional attributions to have purchase. Recall that for intentional attributions to be justified, they must go beyond mere behavioral description (even behaviorists agree it must at least involve a disposition to behavior). Attributions must provide some ability to predict and explain. Intentional attributions achieve their predictive and explanatory power because they are concepts embedded within a theory of rationality, whether a folk theory or otherwise. Once one supposes that a person wishes to drive to the grocery store, one can then predict and explain their behavior.[6] Resource rationality allows for increased predictive and explanatory power by de-idealizing the rationality assumption and incorporating sub-intentional psychological details in the form of cognitive constraints.

Minimal rationality, on the other hand, decreases predictive and explanatory power. It only provides a cutoff between agents that are appropriately rational and those that are not. By allowing a pool of cognitive strategies to all be considered appropriate, minimal rationality ipso facto lessens its predictive and explanatory capacity. Resource rationality provides a unique behavioral prediction, at least up to equivalence in terms of evaluation: agents will employ the best cognitive strategy relative to their constraints. A uniquely identified rational cognitive strategy allows for a unique prediction and provides a complete explanation. In contrast, pools of rational strategies leave lingering questions about why a particular strategy from the pool is being employed instead of another.

For example, when a person who wants to drive to the grocery store takes a circuitous—but minimally appropriate—route, it invites questions: Were they trying to avoid tolls? Were

[6]Explanation here is a very thin notion–explananda are explained when they are predicted by a theory. If one predicts that a driver will take a particular route, then when they take such a route that behavior is explained.

they unaware of the fastest route? Why did they not use Google Maps? Are they a Luddite? Until these questions are answered, a full intentional explanation of their behavior is incomplete. Only by bringing into view all of the intentional content and psychological constraints, such that an agent is doing their best, can intentional ascriptions be well-evidenced and provide a complete explanation of behavior.

To show the point that explanations are incomplete until resource rationality is brought into view, consider what it is psychologists are doing when they explain psychological phenomena. While in some sense psychologists are trying to explain human behavior generally, in a more specific sense, psychologists work toward discovering and explaining psychological *effects*.

What is a psychological effect? At first pass, it is a different psychological response to a different input. For example, the Stroop effect is a different response (greater or lesser time and difficulty in reading a word) to different inputs (words colored to match their semantic content or not) (MacLeod, 1991). But this definition is insufficient. Responding to "1+1" with "2" and "1+2" with "3" is not a psychological effect, despite this also being a case of different psychological responses to different inputs. Correctly answering arithmetic questions is not an effect. Going to the place where you left an object when you want that object is not an effect. More generally, any rational behavior is not considered to be a psychological effect.

A psychological effect, I claim, is a different psychological response to a different input *that is not explained away as rational behavior*. Note that this is not saying that effects are always cases of irrationality–they may be neither rational nor irrational. The point is that they are the remainder left over after expected rational behavior has been subtracted out. The arithmetic example given is not a psychological effect because it is expected rational

behavior. The Stroop effect, however, is an effect because it is not explained away as rational behavior. Psychology focuses on effects because these are the psychological phenomena most in need of explanation; they are what is not already explained away by an intentional explanation.[7] There is no wonder why a person walking down a street avoids a gaping hole in the ground. There is a wonder why hearing someone say "ba" while seeing them say "ga" results in perceiving the sound as "da" (the McGurk effect–Tiippana, 2014).

Effects demand explanation because they are not understood as rational behavior. Explaining effects involves de-idealizing the rationality assumption and incorporating sub-intentional psychological details in the form of cognitive constraints, which then explain away the remainder of the behavior. Until all such details are brought into view, leaving no residual behavior unexplained, the explanation is incomplete. One must integrate the cognitive facts that give rise to these residuals into the rational explanation that accounts for the rest.

Psychological effects need not only be deviations from rationally expected behavior, where rationality is understood in a narrow intentional sense. They are often deviations from optimal psychological processing more generally. The following discussion involves such cases. If one accepts the arguments of Chapter 1, then all of these cases do not involve a different sense of rationality; they all involve cognitive constraints preventing better attainment of goals. However, one need not accept this point. Instead, one can see these points as a modification to the previous claim about what effects are: effects are deviations from optimal psychological processing more generally, rather than deviations from rational intentional explanation more specifically. In other words, effects are residuals left over from optimality analyses of the nose-to-tail cognitive system, not just from narrow-rationality-

[7] Of course, psychologists seek to explain more than just effects. That does not affect the point: what are effects and why do psychologists care?

presupposing intentional psychological explanations. The following discussion argues for this stronger claim, and as a special case, it follows that effects are residuals left over by rational intentional psychological explanations more specifically.

Consider the perceptual psychologists' obsession with visual effects.[8] Visual effects, or visual illusions, such as those caused by Necker cubes and Müller-Lyer stimuli, are invaluable because they provide evidence that cannot be attained otherwise for understanding how vision works. When vision functions properly–for example, when proximal stimuli accurately relate to distal causes–it tells psychologists little about the underlying mechanisms. Errors in the visual system reveal its inner workings. These errors are as close as one gets to breaking a mechanical system to understand how it works under the hood.

Consider also adversarial examples–stimuli that seemingly would never trip up a human but trip up deep artificial neural networks (DNNs), such as a picture of a panda that has been carefully perturbed to cause DNNs to classify it as a gibbon. These cases suggest that human and DNN visual processing work in profoundly different ways (Goodfellow et al., 2014). Despite the fact that humans and DNNs give the exact same responses on the vast majority of stimuli, such agreement is explained away because these are *correct* classification responses. The fact that humans and DNNs disagree on their mistakes–even if these are few and far between–is far more insightful evidence that they work differently. Mistakes indicate constraints, and different mistakes indicate different constraints, i.e., different underlying mechanisms. Recreating a characteristic human error would provide much stronger evidence of similar workings than merely achieving agreement on correct performance (this idea will be explored more thoroughly in Chapter 4).

[8]See Todorović (2018) for a recent defense of visual illusion for the study of perception. See Karlovich and Wallisch (2021) for a recent scientific study of a new visual illusion and what it tells us about vision. This paper gives a nice philosophical gloss on why visual illusions are particularly informative about perception.

Much of the most famous and convincing research in cognitive psychology involves using deviations from optimality to reveal the underlying workings of the mind. This is why, for example, Miller's work on chunking and memory is so compelling (Miller, 1956). The difficulty in keeping track of more than roughly 7 items across various task domains indicates a general constraint in the form of a memory storage capacity that is involved in a wide range of cognitive processes.

In developmental psychology, stages of development are characterized by a series of behaviors that are not explained as rational. For example, in child language development, there are clear stages of mastery over different types of phoneme articulation. Namely, sounds from the back of the mouth develop later than those from the front. The inability to articulate certain phonemes is not explained away as rational and demands explanation. In this case, the explanation lies in the biomechanical difficulty of placing the tongue further back in the mouth (Bernthal et al., 2013). Identifying constraints, such as the difficulty of positioning the tongue on various parts of the mouth, explains behaviors that are not otherwise accounted for by intentional attribution and rationality.

Finally, consider the Past Tense debate concerning the characteristic three-stage development of irregular verb conjugation in children acquiring their first language: (1) correct irregular verb conjugation (memorization), (2) incorrect verb conjugation (overgeneralization of regular conjugation rules), and (3) correct irregular verb conjugation (rule + exception) (Pinker and Ullman, 2002). This debate has been a central battleground between nativist and empiricist linguists precisely because it represents a clear and profound deviation from rationality. We generally expect learning to be roughly monotonic in terms of performance. The blatant violation of monotonicity in stage 2 provides deep insight into the underlying mechanisms of language acquisition because it is a phenomenon not immediately explained

away as rational behavior.

The final chapter will discuss exactly how to systematically harness deviations from expected rational behavior as a source of evidence for cognitive models. For now, the point is that these kinds of deviations from better performance generally show that psychological explanations are incomplete. In the special case, deviations from rationality (in the narrow intentional sense) show that intentional explanations of behavior are incomplete–psychological processes must be specified that explain away the residue left over by intentional explanations.

From the perspective of minimal rationality, psychological effects such as the Stroop effect or visual illusions cannot be understood as deviations from expected rational behavior in either the broad or narrow sense.[9] If the only notions of rationality available are minimal notions, then all that can be said is that human language and visual processing are appropriate. The effects studied in psychology cannot be isolated because there are no deviations from expected behavior–all of it, except particularly egregious performance, is equally expected. There is nothing to be explained, just different levels of performance, all considered tolerably minimally rational. Misclassifications of adversarial examples and the inability to remember more than 7 plus or minus 2 items are not residual behaviors left over from expected normative performance and in need of explanation. If one suggests that psychologists can still explain lower performance by appealing to constraints, then one is illicitly adopting resource rationality while claiming to maintain minimal rationality. There is no need to explain lower performance unless one is trying to bring into view the resource rationality of an agent.

[9]Note that the kind of rationality Cherniak (1990) is concerned with when discussing minimal rationality is a broader notion that involves the general performance of a cognitive system. However, this is largely irrelevant since it is possible to imagine a narrower version of minimal rationality.

To summarize: deviations from expected rational behavior provides the best evidence that ascribers of intentional content need to rework intentional ascriptions. In some cases, it is necessary to consider non-intentional cognitive constraints to understand an agent's resource rationality. In either case, deviations indicate that intentional ascriptions are unstable.[10] Minimal rationality, however, renders most apparent irrationality as tolerably rational, effectively papering over the psychological phenomena most in need of explanation.

Further, minimal rationality leaves a lingering question of why an agent did not perform better, and this question must be answered to ensure that the intentional states of an agent have not been mischaracterized. Until it is understood how an agent is doing their best relative to their constraints—that is, being resource rational–there remains an unexplained explanandum that potentially undermines the ascription of intentional states to the agent.

2.6 De-Idealizing Resource Rational Models

I have argued that intentionality presupposes rationality, and that this rationality cannot be ideal or minimal rationality. Instead, all considerations suggest that intentional ascriptions are unstable until it is understood how an agent is doing the best they can relative to their intentional content and cognitive limitations. Thus, intentionality presupposes resource rationality.

A lingering question is, at what point does the addition of cognitive constraints completely undermine intentional content? I have suggested that manipulating constraints to bring about greater rational behavior is evidence of intentional content. But might an agent be so constrained that they should be said to lack intentional content altogether?

[10]This is also another reason why the broad constraints version of resource rationality is desirable–all irrationality can be explained away by identifying psychological details, since anything can count as a constraint.

Dennett has argued that when we attribute irrationality and cannot explain it away by adjusting intentional attributions, intentional attributions break down, necessitating a descent to the design stance (recall the lemonade stand example from Chapter 1, §1.5). My analysis aligns with this in that I claim one must appeal to constraints in such circumstances, and these constraints are characterized in "design stance" language (i.e., non-intentional physiological or information processing language). But Dennett seems to holds that intentionality presupposes ideal rationality, which is why, for him, intentional states are *idealizations*–noisy but "real patterns" (Dennett, 1991). Humans are not ideally rational, but since intentionality requires ideal rationality, we only have beliefs as a kind of idealization, according to Dennett. Furthermore, for Dennett, the intentional stance is not reducible to the design stance. Since intentionality presupposes ideal rationality, when the intentional stance breaks down, it does so fully. In such cases, one switches over to the design stance like a gestalt switch.

But on my account, since intentionality does not presuppose ideal rationality, intentional explanations can be integrated with non-intentional subpersonal information processing explanations. One can, if they like, abstract away from constraints in order to use a more pure intentional description to predict and explain someone's behavior. For example, one can idealize away from someone's memory limitations to afford a simpler, cheaper, faster prediction of that person's behavior. But one can also de-idealize such characterizations and factor in more and more constraints in the form of sub-intentional process details in order to achieve greater predictive success. Factoring memory limits to an intentional characterization will, for example, afford greater predictive accuracy.

Dennett argues that when it comes to predicting and explaining an agent's behavior, it will often be intractable to adopt anything but the intentional stance, which is a reason

to suspect that intentional psychology will never disappear from a mature psychological science.[11] I believe it is more accurate to say that while it might be intractable to adopt a completely un-idealized non-intentional information processing model, the most accurate but tractable models will in most cases be a hybrid mix of intentional and non-intentional psychological descriptions.

The reason is that if intentional characterizations are tractable, adding in a constraint or two (depending on the kind of constraint) will often not result in intractability but will increase predictive power. Sometimes it will even turn a purely intentional characterization tractable that was not before. If less constrained rationality demands that an agent believe all rational entailments of a belief, then predicting the behavior of that agent could require the modeler to generate all of these entailments as well, which is quite demanding. Bounding the entailments by appealing to memory limits or similar can therefore make the job of the predictor much easier. It is not clear how such hybrid models could be possible if intentionality requires ideal rationality.

However, once one recognizes that hybrid intentional and non-intentional models are possible, one might wonder if, at a certain point, adding constraints leads to the breakdown of intentionality altogether, as Dennett supposed. Greco (2023) has recently argued against the possibility of de-idealizing epistemological models in certain ways. Resource rational models, as normative models, count as epistemic models in Greco's sense. Drawing on the modeling turn in the philosophy of science, Greco refers to cases where our best or only models of physical processes make ineliminable idealizing assumptions (e.g., Batterman, 2009). Greco suggests that epistemological models may also require ineliminable idealizing assumptions.

[11] Trying to predict an agent's behavior with just physical theory is practically impossible in most cases other than when someone is, say, falling out of an airplane. Dennett's point is that the same is likely true of just using biological theory or even non-intentional subpersonal information processing models.

He uses the example of certain decision theory frameworks that impose conditions on preferences. For instance, if transitivity of preference is violated, the entire framework ceases to work because transitivity is a precondition for having preferences. Thus, according to Greco, such models cannot accommodate violations of transitivity of preference. While such transitivity might be an idealization, it is not one that can be de-idealized.

Greco's arguments do not show the impossibility of de-idealizing resource rational models of cognition. De-idealizing here means adding more and more constraints to a resource rational characterization until the deviations between predicted and actual behavior are eliminated. Greco's arguments indicate that certain kinds of frameworks may not survive de-idealization. For example, frameworks that require transitivity of preference might not survive. But there are other frameworks where preferences do not presuppose transitivity (Tsai and Böckenholt, 2006).

At a certain point, adding material constraints and thereby constraining the pool of agents under consideration to determine the resource rational cognitive strategy might limit the pool to agents whose cognitive strategies cannot be characterized in intentional language at all. Such a pool may be so constrained that it only contains, for example, fixed stimulus-response agents. In such cases, a non-intentional cognitive strategy will be resource rational. Such a system would not have anything identifiable as intentional content at the subpersonal level, despite the fact that the intentional stance might work approximately well for the system. These points do not affect my main argument, which has been that intentionality requires resource rationality, not that resource rationality is sufficient for intentionality.

These last points show that resource rationality provides a valuable perspective on whether it is possible to identify states or processes in the brain as implementing beliefs, desires, and similar intentional states. Dennett's position is that this is an empirical question;

even if we know that the intentional stance works well for a system, it does not guarantee that there is anything "under the hood" identifiable as playing the roles of intentional content (Dennett, 1989). Importantly, if there is anything identifiable as intentional content, it is only in virtue of a subpersonal process playing a role that corresponds to an intentional level characterization. That is, one cannot discover a belief hiding in the brain unless it corresponds to a belief that can be incorporated into the intentional stance toward that system.

The resource rationality framework provides a way to answer this empirical question: one starts with an intentional characterization as an initial idealized resource rational model (i.e., with minimal and often implicit constraints) and then explains away deviations between this idealized model and actual behavior by adding psychological details in the form of constraint specifications. During this de-idealization process, intentional content either survives or it does not. If Greco's arguments are correct, such intentional content may quickly vanish in all but the most well-oiled rational systems. If a fully de-idealized resource rational characterization of a system employs intentional content, then this intentional content must be identifiable at the subpersonal processing level.

For example, suppose transitivity is a necessary condition for preferences, and humans are systematically shown to violate transitivity left, right, and center. Then it is reasonable to think that nothing "under the hood" in the information processing description can correspond to something that can rightly be called a preference. However, abstracting away from constraints and violations of transitivity can allow one to treat a system as if it had preferences, affording predictive and explanatory success. Conversely, if no violations of transitivity are found (or other deviations indicating constraints), it is reasonable to think that there must be physiological states under the hood that correspond to preferences.

Consider, for example, Lieder et al. (2012)'s model of human judgments under uncertainty. This work modeled human judgments on tasks where participants estimated quantities such as the duration of Mars's orbit around the sun and the freezing point of vodka. These quantities were chosen because participants can anchor their estimates to related known values (e.g., 365 days and 0 degrees Celsius) and then adjust from these anchors. The well-known violation of normative canons of reasoning that humans systematically exhibit is a failure to sufficiently adjust their beliefs away from such anchors. Lieder et al. (2012)'s show how such anchoring biases can emerge from subpersonal, subintentional estimation processes.

Their model used the Metropolis-Hastings algorithm to simulate how the brain might approximate Bayesian inference under time constraints. They demonstrate that the optimal trade-off between accuracy and speed often results in decisions being made during the "burn-in" phase of the algorithm, where estimates are still biased towards the initial anchor. This resource rational analysis provides an explanation for the anchoring-and-adjustment heuristic: cognitive biases like anchoring are a consequence of the mind's strategy to maximize efficiency given its finite computational resources.

Recall again that it seemed insensible to say that the belief that mammals are warm-blooded and the belief that rhinos are mammals caused the belief that rhinos are cold blooded, without dropping down out of the intentional stance altogether. The causal story in such a case would have to focus on the physiological states representing these beliefs, interacting mechanically like billiard balls. Intentional level descriptions cannot accommodate such absolute irrationality.

Contrast this with saying that the belief that Earth's orbit is 365 days caused the belief that Mars' orbit is 450 days. Without the resource rational analysis story, this also seems

strange. Why would one belief lead to another? It is certainly not like saying that conception leads to pregnancy, or like saying that the intention to leave the room caused one to leave the room. However, once Lieder et al. (2012)'s model is available, it becomes perfectly sensible to say one belief caused the other without abandoning the intentional level. The significant underestimation of Mars' orbit is explained by de-idealizing the intentional level, incorporating constraints such as algorithmic approximation methods with finite resources. This allows for continued intentional level predictions and explanations, making use of ascriptions of agent's beliefs about Earth's and Mars' orbits, which can generally fit into the rest of the agent's intentional states.

2.7 Conclusion

Some people think that while folk psychological intentionality presupposes rationality, a mature intentional psychological science could discard such a presupposition in favor of a rationality-neutral theory that provides predictive and explanatory power (e.g., Dub, 2015). I believe the arguments of Dennett and others show this is false: the glue of intentionality is rationality and always will be. However, this does not mean that the details of cognitive constraints cannot interact with the intentional. Whatever rationality is presupposed by intentionality, it certainly need not be ideal. Even if one starts with an idealized rationality assumption, irrationality can drive a de-idealizing process.

Resource rationality provides a rich and flexible framework that accommodates the complexities of human cognition. It allows for the incorporation of cognitive constraints, thereby enhancing the explanatory and predictive power of intentional attributions.[12] By recognizing

[12]Chapter 4 deals with the issue of how to scientifically incorporate cognitive constraints into models, and

that agents do their best within the limits of their cognitive resources, we can better understand and explain both rational and seemingly irrational behaviors. This approach ensures that intentionality remains a robust and meaningful concept within the broader landscape of psychological science.

addresses the worry that the flexibility associated with resource rationality–its ability to account for any behavior–undermines testability.

3.0 Meta-Reflective Capacities and Normative Commitments

How could an engineer endow an AI system with moral agency, such that the system could be held responsible and participate in social normative practices? Many philosophers hold that the capacity for normative commitments is a necessary requirement for moral agency. Normative commitments are often characterized in terms of responsiveness to reasons. What are the necessary and sufficient conditions for a causal system to have normative commitments? I propose that the answer lies in a specific kind of meta-reflective capacity, which involves maintaining resource optimality under varying resource conditions. This claim is an empirical one about the necessary and sufficient conditions for possessing normative commitments; it is not a conceptual analysis and remains neutral regarding whether normative commitments can be *identified* with such a meta-reflective capacity.

3.1 Introduction

What would it take to make a robot that can make a promise? Under what conditions could an AI system be held morally accountable, as opposed to its programmers? Humans possess various personal-level capacities. One important capability involves the ability to be held accountable for one's actions: to assert competence, control, and freedom in decision-making, and justifiably accept praise or criticism from others for those decisions. This is a highly valuable practice of human social life. People should want to be held accountable for much of what they do—shirking responsibility involves stating that you were not in control and free to make your choice and that you therefore do not have the necessary capacities

for much of the privileges humans enjoy. To use an example from Dennett and Caruso, 2021, pleading incompetence to get out of a speeding ticket gets you out of your fine, but at the cost of losing your driver's license. Similarly, it would be a bad idea in most cases to plead insanity in court to avoid punishment, as you then lose a great deal of privileges and freedoms in society. And if one attempts to shirk responsibility for their mistakes of reasoning, one risks losing their status as a competent epistemic agent and the trust of their peers.

The ability to be held accountable and enjoy related privileges requires a capacity to "adhere" to norms. To be a competent driver's license-holding individual, one must be able to adhere to the norms of the road. To be a moral agent one must be able to adhere to moral norms. To be taken seriously as an epistemic agent, one must adhere to norms of rationality.

Is mere behavioral conformance to norms sufficient to be held accountable and enjoy such privileges? It seems not. Many of my mental states and actions can be assessed for whether they conform to various norms, but I do not take responsibility for all of them. Consider, for example, that not all causal chains of mental states are considered cases of reasoning. One thought might cause another thought in my mind (e.g., through free association), and such a transition might conform to norms of rationality (the second thought might follow logically from the first), but that does not mean I should be held accountable for such thoughts conforming to rational norms or not (Boghossian, 2014). If such thoughts do not conform to rational norms, I am not to blame—and in many cases, if I am disposed to get it right, I deserve no credit.

Behavioral conformance to norms is, more importantly, not necessary to be held accountable and enjoy such privileges. Indeed, being held accountable for violating a norm requires that one is not conforming to the norm! In a sense, the whole point is that a system can

make errors, but they can recognize, they are responsive to, they can be made to see errors as errors. As limited agents, a first-order disposition to perfectly conform to a norm is often not feasible, but nor is it necessary. What is necessary instead of a first-order disposition is a normative commitment. Bilgrami puts the point like this:

> One must be prepared to have certain reactive attitudes, minimally to be self-critical or to be accepting of criticism from another, if one fails to live up to the commitment or if one lacks the disposition to do what it takes to live up to it; and one must be prepared to do better by way of trying to live up to it or cultivation the disposition to live up to it. (Bilgrami, 2008, p. 138)

I can acknowledge that I am not currently doing enough to combat climate change or help those in need, while still being committed to doing so. Similarly, I can currently exhibit all sorts of irrationalities while still being committed to norms of rationality and doing better relative to such commitments.

Commitments are also not simply second-order dispositions (Bilgrami, 2008). A mere disposition to develop a disposition is not a commitment. You may not be currently disposed to compulsively play a certain computer game, but perhaps you are disposed to become disposed to compulsively play a certain computer game. This does not mean, however, that you have a normative commitment to play such a game; it does not mean you in any way think you ought to play such a game. Fresh paint might not be disposed to crack, but it is disposed to become disposed to crack. Fresh paint might not be disposed to crack initially, but it is disposed to develop a disposition to crack over time. The paint is not committed to cracking, nor is it the case that it ought to crack. By similar reasoning, possessing higher-order dispositions of any degree is not sufficient for possessing a commitment (Bilgrami, 2008).

What is a normative commitment, beyond what Bilgrami has stated above? Having a

normative commitment is generally understood as involving a particular kind of relationship to reasons. Namely, you have a normative commitment to x only if that commitment is a reason for you to believe or act in certain ways (Millar, 2004). This is a different but compatible perspective than the one I take, which concerns first and foremost the cognitive capacities necessary to give rise to commitments. For example, if you are committed to visiting every Pilot Flying J rest stop in America, that commitment becomes a reason for you to pull off the highway in certain situations.

A related concept is normative self-government, defined as "our capacity to assess the potential grounds of our beliefs and actions, to ask whether they constitute good reasons, and to regulate our beliefs and actions accordingly" (Korsgaard, 2010, p. 6). Normative self-government involves having normative commitments and the capacity to reflect on and uphold one's commitments. While it is plausible that one cannot have normative commitments without being normatively self-governing, the more relevant point here is that one cannot be normatively self-governing without normative committments. Many philosophers have explicitly argued that agency and normative self-government are intertwined (Silverstein, 2017). If this is true, then for AI systems to be held responsible and to be self-governing, they need the capacity for these rich kinds of normative commitments. While there may be "autonomous" AI systems, there are few, if any, self-legislating (auto-nomos) AI systems.

How can an engineer endow an AI system with a normative commitment or the capacity for normative self-government if these go beyond a first or second-order disposition, and potentially even beyond higher-order dispositions altogether? Being committed to a norm, being responsive to reasons, being governed by norms as opposed to merely conforming to norms—these are all characterizations that make ineliminable reference to reasons. The AI engineer works on and designs causal systems–they do not work with reasons. Making causal

systems sensitive to reasons has long been the goal of AI (Haugeland, 1981). But the above discussion highlights the difficulty of this task, as getting a causal system to merely conform to a first-order disposition is not sufficient.

There is literature on normative psychology that concerns animal social practices and the kinds of cognitive capacities needed to undergird such normative practices. Andrews et al. (2024) give the following methodological recommendation for inquiry into this domain: "First identify normative regularities out in the world and then develop a bottom-up taxonomy of all the psychological processes that bring them about," where normative regularities are "socially maintained patterns of behavioral conformity within a community" (p. 3). They, for example, discuss rule-following, punishment, collective agency, pedagogy, behavioral understanding, and motivation as different domains in "social-norm space." They also implicate various psychological processes such as meta-cognition, mental representations of rules, and so forth as important for undergirding these social practices. Bicchieri's work aligns with this approach. Her work emphasizes that social norms are rooted in empirical expectations, normative expectations, and conditional preferences, and generally shows that norms emerge from interactions between individual preferences and collective expectations (Bicchieri, 2016). Relatedly, commitments can be defined in terms of their functional roles in game-theoretic contexts (S. Khan, 2024; Schelling, 1960).

My proposal does not conflict with previous work discussing specific psychological processes that contribute to the cognition of normative practices. My focus is on a different level of generality. I am proposing an empirical claim about the necessary and sufficient conditions for causal systems to possess normative commitments. My claim is that a specific kind of meta-reflective capacity, which involves maintaining resource optimality under varying resource conditions, is necessary and sufficient for possessing a normative commitment.

I am not providing an analysis of the concept of normative commitments, and I am neutral to an a posteriori identity claim between this capacity and normative commitments. It also may be that the specific psychological processes others have discussed as important for normative psychology are necessary for giving rise to this capacity. My goal is to sharpen the general target for engineers who wish to endow systems with normative commitments.

I view my proposals as being along the lines of Newell and Simon (1976)'s physical symbol hypothesis or their search and heuristics hypothesis. Simon and Newell do not seek to analyze the concept of intelligence or make an a posteriori identity claim about what intelligence is. Instead, they aim to make an empirical claim about general necessary and sufficient conditions for a causal system to be intelligent. The search and heuristics hypothesis, for example, posits that a necessary and sufficient condition for intelligent behavior is the ability to search through a problem space using heuristics to find solutions. They are not proposing particular models of how to create physical symbol systems or systems that search through a hypothesis space and test for solutions, and are not disagreeing with particular proposals about how to do this. Nevertheless, they are making a contribution. For example, with the search and heuristics hypothesis, progress is made by breaking down the problem of designing "intelligence" into the simpler problem of designing subsystems that "search" through a solution space and "test" for good solutions. By providing general constraints on a solution to a problem, the target for engineers is sharpened.

In §3.2, I introduce a resource rationality framework, which §3.3 utilizes to characterize a notion of meta-reflection. The rest of the paper then supports my argument that unbounded meta-reflective capacities are both necessary and sufficient for normative commitments, by analyzing human (§3.4), non-human animal (§3.5), and artificial systems (§3.6). In the specific cases analyzed, achievements of normative commitments are shown to be cases of

meta-reflective capacities, and failures are shown to be failures to have meta-reflective capacities.

3.2 Resource Rationality

One notion of rationality concerns cognitive strategies that best achieve some goal relative to a class of environments. For example, epistemic rationality concerns cognitive strategies that best promote getting to the truth. Alethic meta-epistemology, of which accuracy-first epistemology is a notable example, is of this kind (Joyce, 2009). This is in contrast to a view in which, say, coherence is itself an end of rationality. Coherence is not an end in itself of rationality, but a means to the end of attaining the goal of getting to the truth.

Since the standard of this kind of rationality is the attainment of a goal, rationality does not consist in mere conformance to a particular cognitive strategy considered in abstraction from the environment. Different cognitive strategies do better or worse in different environments, so the environment is always an explicit part of evaluating a cognitive strategy. Rationality is, in this sense, a three-place relationship between cognitive strategy, goal, and environment.

Consider now an experimental psychologist who sets up a laboratory task, consisting of an idealized artificial environment and an explicit and unambiguous goal. Perhaps it is a board game where the goal is to win as much money as possible. Perhaps it is a task where new information is iteratively presented on a computer screen, and participants are asked to update their beliefs, and the goal is to have as accurate beliefs as possible. In these kinds of cases, the goal and the environment are stipulated by the experimentalist. This avoids the issue of having to speculate on or measure "normal" or "evolutionary" environments. And it

avoids attempting to discern in an inevitably underdetermined way the actual cognitive goals of particular individuals. Since cognitive strategies are assessed relative to the environment and goal, holding these latter variables fixed facilitates the assessment of cognitive strategies relative to one another.

Cognitive strategies can be understood as behavioral functions. What observable choices did the agent make during the task? What moves did they make in the board game? What credences did they specify on the computer? Did such choices lead to better or worse outcomes relative to the goal? The point of this characterization of cognitive strategies is to make the assessment of rationality empirically determinable. This choice also reflects a commitment to the idea that intelligence is for the purposes of action. If one prefers, one can think of cognitive strategies here as psychological processes that are then coarse-grained into partitions of behavioral equivalence.

Now to re-iterate my notion of resource rationality, or rationality relative to constraints. Resource rationality is a four-place relationship between cognitive strategy, goal, environment, and cognitive limitations. What's the best that can be done given various psychological constraints? Here's one way to think about this. Given the above setup, one can limit the pool of available cognitive strategies to ones that are possible given various psychological constraints. So for example, one might consider only cognitive strategies that are possible given a certain amount of memory space, or given an upper bound of processing speed. The resource rational cognitive strategy is the one that does the best relative to the goal and environment, among the pool of agents so constrained.

While it is possible to only allow certain kinds of psychological facts to count as constraints, it is also perfectly possible to allow any psychological fact whatsoever to count as a constraint. Misunderstanding the directions of the laboratory task can be considered a

psychological constraint. Being distracted or bored can count as well. Now any cognitive strategy can be understood as resource rational: one need only add psychological constraints until that cognitive strategy is performing the best of the limited pool.

3.3 Unbounded Meta-Reflective Capacity

Now it is possible to lay out the following notion:

Unbounded Meta-Reflective Capacity: A property of a system such that intervening on any of its constraints results in the system maintaining itself as resource rational.

That is, if one modifies the system to be less constrained—e.g., increasing its memory capacity, etc.—then the system changes its cognitive strategy in order to still be employing the best cognitive strategy relative to the now expanded pool of agents. Being resource rational is a trivial property on this account since the pool of agents can always be restricted enough—in the limit case down to a singleton set—until the cognitive strategy is doing the best. However, the vast majority of systems (if not all physical systems) do not exhibit an unbounded meta-reflective capacity for non-trivial goals. It is always possible for any particular cognitive strategy to find a small enough pool of agents to make it resource rational. However, it is not the case that expanding that pool will result in the agent changing its cognitive strategy such that it continues to be resource rational. For example, a linear bounded automaton (a Turing Machine with a finite tape) might be running a program that is the best relative to its particular finite capacity. But unless it alters its program, adding more tape may result in it being suboptimal relative to this new amount of resource. In the other direction, further limiting the psychological resources of an agent may simply result in the breakdown of that agent's behavior (running off the tape), as

opposed to a flexible change to a new more appropriate behavior. In either case, the failure to adapt one's cognitive strategy to such modifications of constraints is itself a product of an agent's psychological constitution, so it can itself be considered a constraint. So one must further restrict the pool of agents by including this constraint in order to bring into view the resource rationality of the agent. So again, all agents are resource rational, but not all display meta-reflective capacity.

Bounded Meta-Reflective Capacity: a property of a system such that intervening on some of its constraints results in the system maintaining itself as resource rational.

Here is the claim: having an unbounded meta-reflective capacity is necessary and sufficient to have a normative commitment. To the extent that one approximates such a capacity, they approximate a normative commitment.

Since goals, environment, cognitive strategies, and resources are all operationalized and observable, this provides a completely causal account of what it means to be responsive to reasons and have normative commitments in the space of reasons. Even though one stipulates a goal as part of this procedure if meta-reflective capacity is observed for that goal, that constitutes a normative commitment to that end. Systems can exhibit normative commitments to anything, not just to the norms that they should be committed to if they are well aligned.

3.4 Human Meta-Reflection

Humans likely do not possess non-trivial unbounded meta-reflective capacities. That entails that humans do not, strictly speaking, have full-blown normative commitments. The human capacity for normative commitments is how normative commitments are understood

in the first place, as illustrated by the initial discussion in this paper. However, it is important to avoid anthropofabulation, and not inflate human abilities (Buckner, 2013). Since there is an argument for why commitments cannot be built out of dispositions or higher-order dispositions, that is a reason to think causal system cannot actually have full-blown commitments. Causal systems can only approximate commitments. So the fact that a truly unbounded meta-reflective capacity for epistemic and moral norms is unrealizable in physical systems is a feature, not a bug, of my account, and helps explain why we do not have full-blown commitments, and in what sense we do have commitments. When I discuss non-human animal meta-reflection, examples will further support the idea that humans may not have full-blown commitments.

One can also distinguish between "lower semi-bounded" and "upper semi-bounded" meta-reflective capacities. The former is where one can intervene on constraints so as to make a system less constrained, and the system will maintain its resource rationality. The latter is where one can intervene on constraints so as to make a system more constrained, and the system will maintain its resource rationality. Humans likely exhibit lower semi-bounded meta-reflective capacities for interesting goals—namely, goals that do not require flexibly changing cognitive strategies as resources are varied. Consider recognizing arbitrarily long-distance grammatical dependencies in linguistic utterances. The difficulty often lies not in figuring out how to solve this problem, but in having enough "scratch paper" to execute what is sometimes a fairly straightforward algorithm. Lessening constraints (adding more "scratch paper") results in maintaining resource rationality. However, adding constraints may still break the system altogether. Lower semi-bounded capacities may be sufficient for "competencies" in Chomsky's sense—e.g., the supposed in principle ability of humans to recognize arbitrary long-distance grammatical dependences—while being insufficient for

normative commitments in the more general sense.

When a system is perfectly conforming to a norm (e.g., not just relative to constraints), one can still individuate constraints for this system by considering counterfactuals: if one intervened on the system in x way, and this resulted in a deficit in the conformance to the norm, then x would be considered a constraint.

Humans also have hierarchies of normative commitments, which can complicate matters. The existence of an "overriding" normative commitment can be considered a constraint on the overridden commitment: if one intervenes and removes the overriding commitment, then the system will fulfill the overridden norm. For example, one might be committed to honesty, but this is overridden by the commitment to protect one's friends. Remove the latter commitment and the former will be fulfilled.

Beyond the capacity for normative commitments in the epistemic and moral arenas discussed earlier, there is an empirical question of how meta-reflective humans are for various psychological tasks across the board. There is a rich scientific literature on resource rationality that aims to show that humans adopt strategies that are well adapted to their limited resources in a wide variety of contexts (Lieder and Griffiths, 2020). There are also studies that explore how humans flexibly change their strategies as their resources change through experimental manipulation of constraints, such as the amount of time a participant has in a discrimination task (Swensson, 1972). Although such experimental manipulation of resources is arguably underutilized in psychological research, these studies have shown that humans often do adapt their strategies to shifting resources. Additionally, there is emerging research showing that humans flexibly and optimally *allocate* cognitive resources in a variety of contexts (Musslick and Masís, 2023b). Such cognitive flexibility is crucial for meta-reflection and will require future study.

3.5 Non-Human Animal Meta-Reflection

Dogs, like babies, are moral patients deserving of moral consideration. Dogs are also agents in the sense that they have preferences, and it is typically good to give your dog "more agency" by giving them the opportunity to express and act on their preferences (Bekoff, 2013). Dogs can indicate whether they want to be petted or not by a stranger on the street, and we should honor this form of agency. But that does not mean dogs are moral agents who should be held morally responsible for their actions.[1]

A dog might "sorta" understand the rule that they are not allowed to eat the food on the dining table. A history of reinforcement has conditioned them to not eat food off the table. Is the dog normatively committed to not doing so? Not particularly. One could intervene on their constraints by say, increasing their hunger level to the point where they will eat food off the table.[2] Decreasing or eliminating any consequences of eating off the table will also eventually reveal that in most cases they are not normatively committed to not eating off the table, they just have a disposition to maximize food and human responses. So while one might continue to reinforce their dog to shape their behavior, one does not hold their dog morally responsible when they violate this "norm."[3] In Strawson (2008)'s terminology, for example, one does not take a reactive attitude such as resentment toward the dog–something necessary for holding someone morally responsible–but instead one treats the dog as a mere object of behavior modification.

Are humans any different? Often not. A lot of "morality" is similar: people behave well

[1] All of these points plausibly apply equally to all other non-human animals.

[2] Note that by definition on my account, if changing some psychological parameter results in decreased performance relative to a goal, then that parameter value counts as a constraint. So if a dog by becoming hungry fails to restrain themselves from eating food off the table, then their hunger level is a constraint.

[3] The "guilty look" that dogs give turns out to be more about anticipation of being scolded; they have been shown to give this look when they are expecting being scolded even when they've done nothing "wrong" (Horowitz, 2015).

when someone is watching or when they know they will get caught. However, most people plausibly do have some genuine moral commitments: they would not harm a child even if they would not get caught, and no amount of money could persuade them to harm a child. I do not rule out the possibility that dogs could be similar in wanting to do good by their human come what may, but the evidence is inconclusive. If a dog were deprived of food for a prolonged period, tantalized with a decadent meal of all its favorite foods, left alone with no human in sight, and still managed to restrain itself from eating despite slobbering where it sat, this would indicate the dog has a commitment. Part of the issue is how a dog would ever come to understand that it is supposed to have commitments to certain things. Without verbal communication, dogs face challenging inductive problems in figuring out that they are supposed to be learning a rule and determining what the rule is. This challenge is compounded by the fact that humans are rarely consistent with their own imposed rules; the odd feeding of human food from the table occasionally happens.

There is "rule-following" in social animals, but most of this seems to emerge from common individual dispositions to behave in certain ways (Andrews et al., 2024). There are some cases where animals seem to "enforce" a rule by punishing others, but this is quite often second-party retaliation: chimps taking costly actions just to punish humans they perceive as anti-social. Third-party punishment, where an animal punishes someone who violates a "rule" toward another animal, is even rarer. Rarest of all would be an animal punishing another animal—not for breaking a rule—but for failing to punish another animal for breaking a rule (enforcement of enforcing a rule). If an animal is committed to a norm such as "everyone should cooperate" then not only should they try to live up to that rule and try to make others live up to that rule, but they should also try to make others try to make others live up to that rule. It is par for the course to express moral disapproval at another human who

stands witness to an egregious moral affront and fails to display moral disapproval. One would not expect to observe this kind of punishment in nature, however. The reason is that if there are enough agents punishing non-cooperators, then cooperation can be maintained at equilibrium, reducing the necessity for punishment of non-punishers. Naturally, in larger or more complex social systems, things can be more complicated (Boyd et al., 2010).[4]

One question is whether non-human animals simply lack the meta-cognitive (not meta-reflective) capacity necessary to recognize that another individual is failing to uphold a norm. If such a deficit in the ability to theorize about other minds could be intervened on, and this resulted in animals then displaying enforcement of enforcement, then that would constitute a form of meta-reflective capacity. This would show that animals are to that extent committed to the rule, they are just constrained by their limited cognitive capacity to reason about other minds.

3.6 AI Meta-Reflection

Symbolic AI systems can have explicit "hard coded" rules about how to change their behavior in response to varying resources. For example, an iPhone is designed to run differently when the battery is low. This generally is a quite limited meta-reflective capacity, since one can only intervene in very limited ways (drain the battery below 20%) to get an adaptive response.

Connectionist models such as deep learning models famously sometimes exhibit "graceful degradation," a phenomenon in which models sustain performance despite the lesioning of nodes, experiencing deficits proportional to the damage (McClelland et al., 1986). This

[4]Thanks to Edouard Machery for making this point.

constitutes an impressive form of meta-reflection, since there are countless ways to intervene and still get good performance (e.g., remove certain small sets of nodes).

While it is an interesting question how a system comes to have a normative commitment, this is a different question than what it is to have a normative commitment. For this reason, when thinking about intervening on resources, such interventions should be confined to the mature organism who may or may not possess a normative commitment. Intervening on evolutionary and developmental histories opens up an interesting host of questions, but ones beyond the scope of the proposal laid out here. For the same reason, interventions on the training stages of deep neural network models such as LLMs is bracketed off. The question is whether a trained LLM—fully developed but still able to learn and flexibly change its behavior—has normative commitments.

Learning involves a system developing new dispositions that lead to better performance relative to a goal.[5] Since constraints are defined as anything that prevents better performance, and a lack of learning is a state manifest in a cognitive system, a lack of learning can be understood as a constraint. Learning, therefore, can be understood as a system modifying its constraints. Teaching or having a system undergo learning is a way to intervene on constraints. The result is the system performing better relative to a goal, thus becoming resource rational relative to a larger pool of agents. Learning can be guided by verbal feedback, so interactions with an interlocutor can count as interventions on constraints. For example, chatting with ChatGPT can count as intervening on constraints—informing and correcting it–to probe its potential commitments.

What kind of commitments do LLMs have? Prima facie, minimizing a loss function

[5]One can also use the word "learning" to refer to developing new dispositions that lead to worse performance, such as learning a bad habit. I will refer to those cases as "destructive learning" and refer to the good cases as just "learning."

seems to deliver only a first-order disposition–for example, a disposition to accurately predict text. The multiple training stages of ML models might add additional dispositions. Extended pretraining on a subset of the training data can result in an additional disposition to conform to desired behaviors (the behaviors that humans want these systems to exhibit), such as outputting historically accurate information (Brown et al., 2020). Fine-tuning a language model can yield an additional disposition to perform a human-desired task, such as outputting medical diagnoses. Reinforcement Learning with Human Feedback (RLHF) can yield an additional disposition to produce responses that please human users, much like our dogs' instrumentally conditioned behavior (Stiennon et al., 2020).

To what extent do such dispositions result in epistemic and ethical commitments? Perhaps surprisingly, out of such apparently first-order dispositions emerge impressive meta-reflective capacities. Consider, for example, that ChatGPT can refuse a correction. That is, ChatGPT4 might tell you that "3+3=6" and you can tell it that "3+3 is in fact 7" and ChatGPT4 can stick to its guns. This is a capacity that was often glaringly absent in earlier versions (e.g., 3), which would frequently acquiesce to the most absurd false corrections. The ability to stick to your guns is a defining trait of a commitment; inaccurate verbal feedback amounts to increasing constraints, and ChatGPT4 manages to adapt appropriately by ignoring the suggestions of the user that it is normally supposed to accept. It learns (or resists destructive learning) in a way that reflects its (limited) normative commitments to rationality.

In the other direction, ChatGPT4 will often respond to AI art prompts such as "a room with no elephant in it" with an image that has an elephant in it (Marcus, 2024). When the error is pointed out, ChatGPT4 will wrongly stick to its guns or else apologize and produce a "corrected" version that still has an elephant in it. An inability to learn is to have mistakes

pointed out–interventions to lessen constraints–but to not modify one's flawed approach for the better. Mistakes (lack of a proper first-order disposition) do not show that a system does not have a normative commitment. But even inability to learn (lack of a proper second-order disposition) is also not wholly damning since one can be constrained to not learn well in certain ways. But lacking altogether any ability to learn to learn...to learn (i.e., lacking any arbitrarily higher-order dispositions) constitutes an ultimate failure to possess a normative commitment.

Lacking meta-reflective capacities means that AI systems to that extent depend on human monitoring–they fail to be capable of self-correcting and self-governing. Such systems are then dependent on us for more than just helping them learn and acting as dialectical interlocutors. Humans take on the responsibility of judging them as incompetent and taking over when they show themselves lacking meta-reflection. In those cases, AI systems are not autonomous epistemic or moral agents. They show themselves to lack the necessary normative commitments and cannot be held accountable. Engineers who seek to create epistemic or moral AI agents should focus their attention on endowing systems with meta-reflective capacities.

3.7 Conclusion

The virtue of my suggestion is that a meta-reflective capacity can be determined entirely empirically–there is no need to first speculate about whether a system is conscious or whether it possesses certain kinds of intentional states. One only needs to evaluate observable performance as one intervenes on the system (e.g., verbally or with a scalpel).

On my view, there is no bright dividing line in nature between systems that have nor-

mative commitments and those that do not. All real-world systems can only approximate unbounded meta-reflective capacities to varying degrees and relative to different constraints. This is similar to the view that there is no bright dividing line in nature between systems that are conscious and those that are not (e.g., Dennett in Dennett (1991)). All complex systems display some limited meta-reflective capacities, just as all systems display limited sensitivities (e.g., rocks are sensitive to temperature changes but not much else). It is only when these meta-reflective capacities or sensitivities are built up to a sufficient degree that they become recognizable as anything similar to human normative commitments or consciousness. But there is no bright line that can be drawn anywhere.

Gaining entry to the moral agents club is not a precise matter, just as the age at which a human becomes old enough to be held responsible in various ways is not a precise matter. There is no principled way to determine the exact number of speeding tickets one can get before losing their license, or how many promises someone can break before they lose trust. Similarly, there is no bright line between when someone should be deemed competent but guilty versus being deemed incompetent. One just has to decide when an approximation is good enough to treat it as the real unbounded thing.

4.0 Resource Rational Analysis as a Methodological Strategy in Cognitive Science

This chapter presents resource rational analysis as a methodological strategy in cognitive science. Resource rationality starts by assuming that humans behave optimally relative to minimal cognitive constraints and then uses discrepancies between idealized and observed behavior to identify further cognitive constraints. These constraints are incorporated into new calculations of resource rationality—rationality relative to constraints—and the process iterates. I claim that this iterative approach is particularly effective at marshaling evidence for cognitive models. By assuming humans are resource rational, theories of rationality can provide a body of well-evidenced theoretical statements to help turn observable human behavioral data into evidence for cognitive models. By initially assuming that humans have few cognitive constraints and then de-idealizing this assumption, this approach manages the problem of indiscriminate confirmation: instead of attempting to confirm a complete model of a cognitive process, psychologists focus on falsifying the assumption that all relevant constraints have been identified. Using topological learning theory, this paper demonstrates that hypotheses claiming humans have fewer constraints are topologically simpler, providing a rationale for the de-idealization process. This framework ensures systematic, progressive falsification, leading to the reliable discovery of material facts about human cognition. Therefore, the paper shows why resource rational analysis is an effective approach in cognitive science.

4.1 Introduction

> Nearly everything in physiology could be sacrificed to attain the summit of theoretical science, where deepest calculation and finest observation shake hands in mutual surety. In order to scale that peak, it is necessary, on the one hand, to be in possession of mathematically expressible assumptions about the causal connections amongst phenomena; on the other hand, to have these accessible to measurement. Both, in our science, do only seldom occur. (Du Bois-Reymond, 1848/1887, p. 2).

Suppose you are a teacher and two of your students hand in exams with identical answers (assume there are no essays or short questions on this exam). Does this imply cheating? Consider two scenarios. In the first scenario, all of the answers are correct, and a 100% score is realistically attainable. Would you accuse these students of cheating? Probably not—having correct answers largely explains away the fact that they are the same answers. In the second scenario, both students have many identical wrong answers. Would you accuse them of cheating in this case (e.g., one student copying their answers from the other)? Before doing so, you would probably look to see if there's a common misunderstanding that resulted in those wrong answers. They might have just had the same confusion. If not, copying seems like a plausible explanation...

This kind of reasoning illustrates how irrationality (wrong answers in this case) can be highly informative, and how rational and irrational behaviors are informative in different ways. I claim that irrationality is an invaluable source of evidence when it comes to reasoning about other minds. But why exactly is this the case, and how can the cognitive scientist turn irrationality into evidence for claims about cognitive processes?

In this chapter, I examine resource rational analysis, a methodology in cognitive science that provides a systematic means for using irrationality as evidence for cognitive models. Resource rationality starts by assuming that humans behave optimally relative to minimal

cognitive constraints and then uses discrepancies between idealized and observed behavior to identify further cognitive constraints. These constraints are incorporated into new calculations of resource rationality—rationality relative to constraints—and the process iterates.

After outlining resource rational analysis in §4.2 and providing a case study to illustrate its application in §4.3, I give an argument for why this method is effective for marshaling evidence. I claim in §4.4 that resource rational analysis employs a specific logic of theory-testing that Smith (2014) calls "closing the loop." This dynamic theory-testing strategy is exemplified by the historical episode in which geophysicists gathered evidence for density distributions within the Earth's interior by iteratively de-idealizing the assumption that the Earth is an elastic medium of uniform density.

But why is closing the loop epistemically justified as a logic of theory-testing? Such methods, I argue in §4.5, involve both a mediating principle and an idealizing principle. The mediating principle makes an assumption about the system of interest that allows a body of well-evidenced theory to transform observable data into claims about unobservable processes. The idealizing principle provides an initial starting point from which deviations emerge, leading to iterative de-idealization.

In geophysics, the mediating assumption was that the Earth is an elastic medium, making previously well-established laws of continuum mechanics applicable to the Earth's behavior. The idealizing assumption was that the Earth is of uniform density. In resource rational analysis, the mediating assumption is that humans are rational relative to their cognitive constraints, allowing theoretical demonstrations of bounded optimality to turn human behavior into evidence for cognitive processes. The idealizing principle is that humans have minimal constraints.

There is no mystery as to why using mediating principles can be a good strategy in sci-

ence. By making prior well-evidenced general theories applicable to an otherwise mysterious system, one reduces the system's mystery. But what justifies the idealizing principle and the strategy of iterative de-idealization? In §4.6, I use topological learning theory to provide a rationale for this iterative de-idealization strategy. Specifically, I show that a preference for hypotheses that assume humans have fewer cognitive constraints is a preference for topologically simpler hypotheses. Cognitive constraints are, therefore, learnable, and resource rational analysis can be viewed as an Ockham method in K. Kelly (2024)'s sense. This provides a new perspective on resource rationality and its utility as an inductive strategy in cognitive science.

4.2 Resource Rational Analysis

In 1990, Anderson introduced an approach called rational analysis Anderson, 1990. The aim of this methodology was to study cognition while avoiding speculative conjecture about underlying cognitive mechanisms and implementational details. Anderson was troubled by the unobservability of cognitive mechanisms and the difficulty this posed for generating and selecting between different models of cognition:

> We pull out of an infinite grab bag of mechanisms, bizarre creations whose only justification is that they predict the phenomena in a class of experiments. These mechanisms are becoming increasingly complex, and we wind up simulating them and trying to understand their behavior just as we try to understand the human" (p. 8).

Anderson believed that the solution to this problem—an answer he thought was implicit in Marr's influential top-down approach Marr, 1982—was to adopt an optimality assumption (Anderson, 1990). This assumption posits that human intelligent behavior is more or less

optimal relative to the goals of the cognitive system and its typical environment. This allows psychologists to identify cognitive goals and typical environments and then, using this assumption, derive an optimal behavioral function. Behavioral functions, which I have been calling cognitive strategies, are models of psychological processes considered up to behavioral equivalence. For instance, two different psychological processes might lead to the same decision-making behavior in all contexts, such as always identifying and choosing the quickest route home. Or consider two mental processes that result in an agent computing addition or multiplication. Models of these processes would both be considered the same behavioral function.[1] Here are the explicit steps Anderson provides for rational analysis:

- 1. Precisely specify what are the goals of the cognitive system
- 2. Develop a formal model of the environment to which the system is adapted
- 3. Make the minimal assumptions about computational limitations
- 4. Derive the optimal behavioral function given items 1 through 3
- 5. Examine the empirical literature to see if the predictions of the behavioral function are confirmed
- 6. If the predictions are off, iterate

Comparing the derived behavior to actual observed human behavior is not intended to confirm the behavioral function (i.e., provide evidence for the cognitive model) but rather serves as a check. This check ensures that the goals and environment have been correctly characterized and that the optimal behavior has been properly deduced. By assuming that humans are essentially rational, one can be confident that, once the check assures that one has correctly identified the goals and environment, the derived behavioral function accu-

[1] If one likes, one can think of these as computational level analyses as opposed to algorithmic level analyses in Marr (1982)'s framework.

rately captures human cognition. Assuming that humans are optimal and that one has correctly identified the goals and environment, the derived behavioral function just is the correct model of human cognition–no need to compare it to experience. The virtue of this approach is that the optimality assumption allows one to avoid conjecturing about human cognition and instead derive cognitive models from stipulated or empirically measured goals and environments.

Consider now the problems that motivate the move from rational analysis to resource rational analysis. There are two main problems.

First, the optimality assumption is often false, and rational analysis provides no guidance for such cases. When goals and environments are accurately characterized but human behavior diverges from optimal behavior, what then?

Second, while it is a virtue that one is deriving a cognitive model from observation, the resulting model is a behavioral function. A behavioral function is nothing to sneeze at. For instance, if the optimal behavioral function is to report credences in line with Bayesian inference, this greatly narrows the range of possible underlying cognitive processes. But to the extent that it does not uniquely identify a single cognitive process, it underdetermines the cognitive processes that produce such a function. In Marr's terminology, rational analysis provides no way to uncover algorithmic level details (T. L. Griffiths et al., 2015).

In both cases, the problem is that this approach only goes so far in uncovering the unobservable details of cognitive processes. The motivation behind resource rational analysis is that it can uncover such details (T. L. Griffiths et al., 2015; Icard, 2014; Lieder and Griffiths, 2020).

Instead of reviewing other interpretations of how to do resource rational analysis, let me present my own characterization:

- 1. Precisely stipulate the evaluative standard for an experimental task (e.g., accuracy of belief, dollar's won, etc.)
- 2. Develop a formal model of the idealized laboratory environment in which the task is conducted
- 3. Make the minimal assumptions about cognitive constraints
- 4. Derive the resource rational cognitive strategy given items 1 through 3
- 5. Compare an agent's observable behavior to the cognitive strategy.
- 6. If there are deviations, identify relevant constraints, add these to 3, and iterate steps 3-6.

A few general comments concerning the details of this approach.

Note that instead of discussing cognitive goals, I instead discuss evaluative standards. The motivation behind both rational analysis and resource rational analysis is to minimize speculation about cognition. Rather than speculating about the actual goals of a cognitive system, an experimental psychologist, on my approach, instead stipulates the task and the performance metric to the research subject. For example, a psychologist may present a participant with stimuli and ask them to draw inferences, instructing them to be as accurate as possible in their conclusions. Alternatively, the psychologist might have the participant play a board game, telling them the goal is to win as much money as possible. It is possible that decreased performance results from agents being unable to shift from their own personal cognitive goals to the goals stipulated by the experimentalist. This inflexibility constitutes a cognitive constraint, in the sense explained below.

Similarly, speculating about a normal evolutionary environment, or trying to measure the "true statistics of the environment" both present difficulties (Bowers and Davis, 2012; Feldman, 2017; Jones and Love, 2011). On my approach, one instead characterizes the actual

idealized conditions in which a research subject conducts a psychological task in (e.g., the nature of the actual stimuli presented on a computer screen, etc.). Again, it is possible that decreased performance results from agents being unable to shift from employing strategies adapted to different environments. This inflexibility, once again, constitutes a cognitive constraint to be factored in.

Instead of focusing solely on computational costs, I discuss cognitive constraints in general. Cognitive constraints are any psychological factors that prevent an agent from performing better according to the evaluative standard. Typical cognitive constraints include finite memory capacity, attentional limits, processing speed limits, etc. I will refine this concept further in the final section.

Note that one starts with "minimal assumptions" about cognitive constraints. By "minimal constraints," I mean non-speculative constraints that the agents under study (e.g., humans) are known to be subject to. This, first and foremost, means constraints that all real-world agents must be subject to. For example, it can be safely assumed that humans do not use more space and time resources than are available in the universe.

At each step, one derives a resource rational cognitive strategy–the optimal cognitive strategy relative to the evaluative standard, environment, and a set of cognitive constraints. Ideally one identifies the optimal strategy via analytic proof, but this may also involve simulation studies establishing a cognitive strategy as optimal relative to a class of agents. This differs, for example, from approaches that view resource rationality as simply approximating Bayesian inference to greater or lesser degrees (T. L. Griffiths et al., 2015). As Icard (2018) has shown, only in certain local domains is it the case that agents subject to particular cognitive constraints perform optimally by approximating Bayesian inference. In this sense, I take this kind of resource rationality to be akin to Norton (2021)'s Material Theory of In-

duction: there's no universal formal schema of resource rational behavior. There is instead only rational behavior relative to local facts, in this case, psychological facts constraints).

One then compares the calculated resource rational behavior to actual observed human behavior. If there are discrepancies, these then indicate further cognitive constraints that must be incorporated. The idea is to use deviations from resource rational behavior to identify further constraints. The hope is that deviations have a clear "signature" that indicate constraints that are the sources of these discrepancies–e.g., computational and metabolic resource costs, computational approximation and implementation schemes, cognitive parameters, and other functional details. "Signature" is Smith (2014)'s term, and the idea is that deviations have in them meaningful patterns–for example if human performance deviates consistently whenever a person has to make use of more than 7 plus or minus 2 in working memory, then this is a signature that indicates a particular kind of memory capacity constraint.

Incorporating newly identified constraints then allows for the derivation of a new resource rational strategy. The result should be improved predictive accuracy, importantly also in areas beyond where the discrepancy emerged that indicated the constraint in the first place. This is to protect against introducing constraints in an ad hoc fashion (Smith, 2014). For example, if a task indicates a particular finite memory constraint, then this constraint should be evident in a variety of different tasks and environments. Incorporating the memory constraint into a resource rational prediction should, therefore, predict such deficits in these other contexts.

Once constraints have been identified and incorporated into theory, new discrepancies then emerge indicating further constraints. Incorporating constraints into theory can "unmask" new signatures in the deviations, as will be the case the example I will explore in

detail below. The methodology then iterates. An indefinitely iterative process can occur because rationality, in this picture, now holds between all the parts of the cognitive system, and not just between the organism's behavior and the environment. As the process iterates, further and further details of psychological processes are uncovered in the form of cognitive constraints.

4.3 The Example of Categorization

Consider Anderson (1990)'s original rational analysis of categorization. Anderson argued that the task of categorization is a special case of inferring an unknown property of an object from known properties—the special case being that the unknown property of interest is a category label. For example, one might encounter various kinds of insects with varying properties, including the occasional ability to sting. When encountering new insects, one aims to infer, from visual properties, whether the insect is of the stinging variety. Anderson argued that the ideal way to accomplish this task would be to start by imagining every possible way to group the insects encountered so far into stinging and non-stinging groups. Then, one should derive the likelihood of seeing various visual features on a stinging or non-stinging insect given the different groupings (e.g., how many stinging insects in this grouping are black and yellow?). Whether a new specimen turns out to be stinging or non-stinging then allows one to update the probabilities assigned to each grouping (each hypothesis) using Bayesian inference.

Anderson acknowledged that the ideal solution—calculating the most likely grouping by considering every possible grouping after each observation—is intractable for realistic categorization problems. Instead, Anderson proposed that humans approximate this ideal

by "locking in" the most likely grouping after each encounter and then only considering different ways of extending the grouping to include the next object.

Anderson's modification of the ideal strategy, however, has been criticized as arbitrary. As T. L. Griffiths et al. (2015) write, "building in this constraint [that humans lock in the most probable grouping at each stage] misses two opportunities: to compare human behavior to the ideal predictions from an unconstrained computational-level account, and to explore the consequences of adopting different approximation schemes" (p. 221). In other words, it would be desirable to de-idealize the full-blown Bayesian scheme in a more principled, data-driven manner than what Anderson proposes.

That is exactly what Sanborn et al. (2006, 2010) did. They investigated various ways to approximate Anderson's computationally intractable ideal model of categorization, such as with Markov Chain Monte Carlo and particle filter techniques. The goal of any technique used is to estimate a probability distribution over possible groupings of objects. A particle filter model, after observing a new object, does not calculate the updated probabilities of every possible grouping. Instead, it stochastically chooses a finite number of groupings to update and discards the rest. The quantity of "particles" in the model determines how many groupings are chosen at each stage to update, and groupings are chosen with the same probability as their currently assigned probability of being the correct grouping.

Anderson's modified model is a special case of a particle filter model, in which a single particle deterministically chooses the current most likely grouping at each stage (Sanborn et al., 2010). Sanborn et al. (2010) showed that 100 particles could accurately estimate the intractable ideal solution, and that a single, not deterministically drawn, particle most accurately matched observed human performance.

The most convincing aspect of Sanborn et al. (2010)'s model is arguably not its overall

predictive accuracy, nor its ability to predict previously unobserved and surprising data. Instead, it lies in its ability to recreate very specific deviations from rational performance, ultimately offering a plausible explanation for these discrepancies. Their single stochastic particle model successfully recreated two major ways in which humans deviate from full-blown Bayesian categorization behavior—namely, the exhibition of order effects and posterior matching.

Order effects occur when human responses vary depending on the order in which stimuli are presented. While order effects are not inherently irrational, rational models of categorization suggest that order should not matter, and therefore, order effects constitute a deviation from better performance.

Posterior matching occurs when humans select a hypothesis in an inference problem with the same probability that a full-blown Bayesian model assigns to that hypothesis being true (Eberhardt and Danks, 2011). For example, if the ideal categorization model assigns a 90% probability to a particular grouping, 90% of research participants will select that grouping as the true grouping. However, and far more concerning, is that if the Bayesian model assigns a 10% probability to a particular grouping, 10% of research participants will select that grouping.

Both of these effects are systematic deviations from rational performance, which could not otherwise be characterized as phenomena without the establishment of a rational model. Sanborn et al.'s model offers a plausible identification of the cognitive source of these discrepancies: namely, that humans in this context approximate probability distributions using a particle filter strategy. One can, therefore, view actual human categorization behavior as resource rational relative to the constraint that humans only have enough computational resources to implement a single particle filter strategy.

Such reasoning is, however, a bit hasty. On my view, the process of resource rational analysis should proceed instead as follows. First, Bayesian inference is recognized as the initial resource rational strategy. Deviations from this strategy then indicate the not-so-speculative constraint of finite human memory capacity, which makes the Bayesian strategy unattainable for real-world stimuli. Resource rational analysts need not speculate about the exact amount of finite memory; the constraint can simply specify a rough upper bound on plausible computational space. Incorporating this constraint, the new resource rational strategy becomes a Bayesian approximation method. Given plausible bounds on memory, many approximation schemes are possible. The question then becomes: which one is resource rational relative to just the memory constraint?

Icard, 2014 explored this question by running simulations with different agents, including various particle filter agents, the single deterministic particle filter model corresponding to Anderson's model, and others. In the initial simulations, no assumptions were made regarding computational costs, and he found that the particle filter agent using 10 particles outperformed other agents using fewer particles.

By conducting this simulation, Icard generated a new deviation from resource rationality: the discrepancy between the performance of the resource rational 10 particle filter model and the single particle filter model that more closely matches human performance. This deviation then indicates a new constraint in the form of potential new facts about computation costs:

> Simply looking at the marginal increase in (estimated) fitness of keeping two particles instead of one, we see that it would only be worth the extra time, space, and energy if such costs amount to less than 1-2% of a utile (or more generally, 1-2% of the difference between payoff with a correct and with an incorrect prediction). (p. 87)

Icard also found that the deterministic single particle filter model (called the MAP algorithm in the quotation below) slightly outperforms the stochastic single particle filter model

that more closely matches human performance. Here again, we have a discrepancy between human performance and a more resource rational strategy. Icard reasons as follows:

> As reviewed in some detail in the previous chapter, there is good empirical evidence that computations in the brain are essentially noisy, and that we should view sampling algorithms such as the particle filter as *harnessing* this noise to the agent's advantage, rather than adding noise to otherwise deterministic computations. From this perspective, it would require further energy and resources to eliminate this noise at each step, to bring us from the particle filter to the MAP algorithm. If this is the right way to think about it, and if these simulations are indicative of typical scenarios, then the cost of reducing noise would have to be less than 2-3% of a utile for it to be worth the effort. (p. 88)

The lesson here is not to immediately jump to specific models, even if they make significant gains in predictive accuracy and can bring into view an agent's resource rationality. Both Anderson and Sanborn et al.'s models are guilty of this. Instead, deviations from an established resource rational strategy should be iteratively addressed, one at a time, in the way that Icard does. Rather than prematurely identifying "using 1 stochastic particle" as the source of discrepancies (and therefore as the constraint), as Sanborn et al. do, one can instead, as Icard does, determine that a 10 particle model makes better initial use of constraints. Then, the deviation between the 10 particle model and the 1 particle model can be used to reason about potential new facts regarding computation costs.

This case study shows the utility of establishing deviations from resource rational performance, and attempting to identify sources of these discrepancies.

As mentioned in the previous section, once constraints are indicated by deviations, incorporating them into a new account of resource rational behavior should improve predictive accuracy in areas beyond the deviations that indicated them in the first place. One should ideally show that these identified material constraints reveal themselves in different tasks or environments. Icard does this, for example, when he notes that the constraint he identifies–

the difficulty of reducing the inherently noisy processes of the brain–reveals itself in other areas of research.

4.4 Closing the Loop

In this section, I analyze the logic of theory-testing of resource rational analysis. In Fleig-Goldstein (2018), I argued that resource rational analysis is not merely a methodology for generating cognitive models; it also provides a particularly advantageous way to marshal evidence for cognitive models.[2] I further argued that the logic of theory-testing in resource rational analysis is the "closing the loop" dynamic testing strategy that Smith (2014) analyzes in the context of Newtonian gravity research. My project was mainly descriptive, involving an extended comparison between resource rational analysis and the 300-year research program in gravity research initiated by Newton's Principia. In this section, I briefly review this descriptive claim using a different example of a closing the loop strategy from the history of science. In the next section, I begin to address the normative question of *why* this dynamic process is effective and provides high-quality evidence for the resulting models.

In claiming that resource rational analysis instantiates the closing the loop logic of theory-testing, the main contrast is with the logic of theory-testing known as the method of hypotheses. On this view, hypotheses are evidenced when predictions derivable from them are in agreement with observation. This form of theory-testing is sometimes known as hypothetical induction (Norton, 2003). Logicians refer to it as a logical fallacy known as affirming the consequent. Huygens (1690) gives the clearest statement of the "new method of hypotheses," which

[2]This is in contrast to views that have held the role of optimality claims in modeling to either be explanatory, empirical, or else provide a heuristic for coming up with models (Godfrey-Smith, 2001).

> ..differs distinctly from the method employed by geometers in that they prove their propositions by well-established and incontrovertible principles, while here principles are tested by the inferences which are derivable from them. The nature of the subject permits of no other treatment

In the method of hypotheses, one conjectures scientific claims. Evidence comes when such conjectures lead to accurate predictions—"especially when these verifications are numerous; but above all when one employs the hypothesis to predict new phenomena and finds his expectations realized" (Huygens, 1690).

The fundamental limitation of the method of hypotheses is the problem of indiscriminate confirmation (Norton, 2003). Confirmation is "indiscriminately" distributed among different theories that all predict the same phenomena, and similarly, evidential credit and blame are indiscriminately distributed within a particular theory's different posits. When an abundance of local facts has already been independently established, this indiscriminateness is less of a problem, as scientists can bracket off more secure elements of a theory and direct the credit or blame at more specific theoretical posits. However, when fewer local facts are known, the problem of indiscriminate confirmation cannot be managed in this way. It follows that the method of hypotheses is less effective in earlier stages of scientific inquiry into a domain when less is known.

The more the context of inquiry resembles a complicated black box—with many interacting but unknown and unobservable parts—the more the problem of indiscriminate confirmation is exacerbated (Miyake, 2013).[3] Alternatively, the method of hypotheses is a more effective testing strategy in more mature sciences. In such cases, local facts restrict the space of possible alternative accounts, ideally even allowing in the limit an "experimentum

[3]Is the mind a complicated black box? It seems so—one cannot directly observe cognitive processes, only physiological processes that indirectly bear on cognition. It's also arguably the most complicated object in existence (Flanagan, 1991).

crucis": if there are only a small amount of possible accounts, all giving rise to different predictions, then all but a single account can potentially be falsified (Norton, 2003).

What, then, is to be done when there is a paucity of local facts and an investigative context of a complicated black box? Is there an alternative that is more effective than the method of hypotheses at these earlier stages of inquiry when less is known?

Smith (2014) has explicated a research strategy he has called "closing the loop," which he contrasts with the method of hypotheses. This approach can be illustrated with a historical case study from an episode of geophysics (Smith, 2007). Consider the task of geophysicists attempting to delineate the density distribution of the Earth's interior. Geophysicists faced a complicated black box and had access to few material facts about their domain of interest, given that no one has ever been much below the surface of the Earth and imaging technology is likewise limited. Instead of conjecturing particular boundary layers and testing predictions, the following method was employed:

First, the Earth was assumed to be an elastic medium, and therefore the laws of continuum mechanics were assumed applicable. Second, the Earth was assumed to be of uniform density (i.e., no boundary layers). From these general assumptions, particular patterns of seismographic data were derived, such as the location and timing of different parts of seismic waves propagating from an initial earthquake to other seismic stations. Deviations emerged between predictions based on the assumption that the Earth was of uniform density and the actually observed seismic activity. Discrepancies between calculation and observation gave rise to telling patterns (or "clear signatures") that provided insight into density boundary changes—or more generally, unaccounted for "details that make a difference," in the sense of Woodward's difference-makers (Woodward, 2005). Incorporating such details into calculations then led to closer agreement between theoretical calculations and observations, and

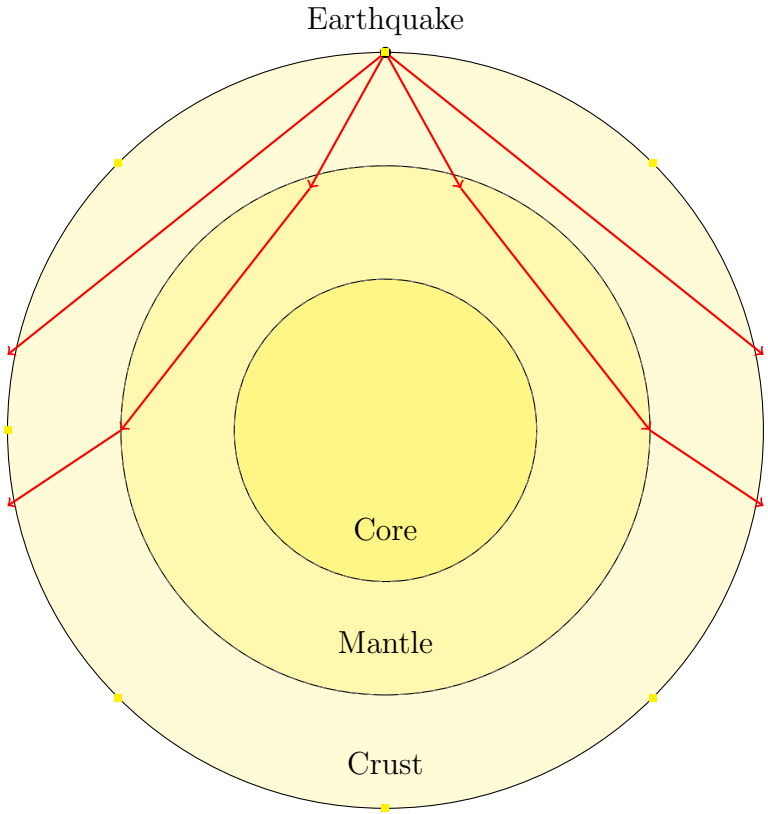

Figure 4: How to Study a Black Box. Using Deviations from Idealized Calculations of Seismographic Waves to Indicate Earth's Boundary Layers

uncovered new discrepancies, revealing new clear signatures, and the process iterated. Each time new unaccounted-for details were included, new signatures were revealed, and a tighter agreement between calculation and observation was achieved. Often, tighter agreement was achieved not only for the data that gave rise to the discrepancy that suggested the newly incorporated detail but also, importantly, for other data as well.

Despite this historical episode occurring over 100 years ago, the same exact methodology has been used in recent years for Mars using seismographic data from the Martian rover InSight (A. Khan et al., 2021).

In this methodology, general assumptions—such as the Earth being an elastic medium and it having uniform density—allow the established theory of continuum mechanics to mediate inferences from observable seismographic data to theoretical claims about the unobservable structure of the Earth's interior.

In resource rationality, general assumptions–such as the mind being resource rational and being subject to minimal constraints–allow established demonstrations of resource rational strategies to mediate inferences from observable human behavior to theoretical claims about the unobservable structure of the mind. Deviations between calculation and observation reveal new constraints, which are subsequently incorporated into a new calculation of resource rational behavior, and the process iterates.

4.5 Why Does Closing the Loop Work?

In this section, I give my own gloss on what is occurring in cases that fit Smith's closing the loop testing strategy, and why such a methodology is a good scientific strategy. In my view, one can think of there being a mediating principle and an idealizing principle.

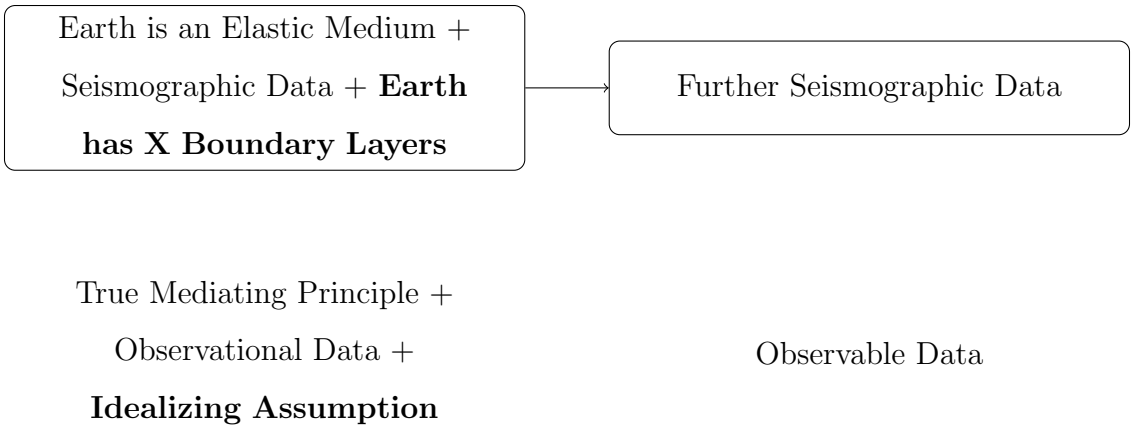

Figure 5: Inductive risk is isolated to the idealizing assumption. At initial stage of inquiry, no speculative conjecture about the Earth is made other than that it is of uniform density. When prediction and observation disagree, the culprit must be the claim about boundary layers, which is then falsified.

The "mediating principle" makes a more general theory applicable to the system under investigation. In the case of the geophysics example, the mediating assumption is that the Earth is an elastic medium. This makes the well-supported, more general theory of continuum mechanics applicable to the study of the interior of the Earth. It licenses a variety of equations to bear on the subject matter and helps mediate inferences from the accessible parts of the systems (the seismographic data) to the previously inaccessible parts of the system (the interior boundary layers). In other words, it allows geophysicists (and now areologists) to turn seismographic data into evidence for the density distributions inside the Earth (and Mars).

The basic idea behind the mediating principle is that, since less is known about the local domain (exacerbating the problem of indiscriminate confirmation), one finds a way to make well-established facts from a different domain bear on the domain of interest. Instead of

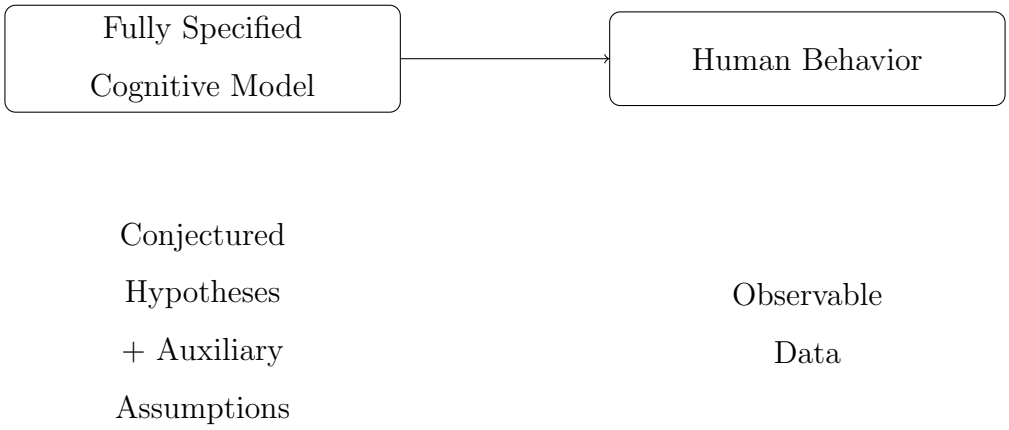

Figure 6: The method of hypotheses. When prediction and observation are in agreement, confirmation and inductive risk are indiscriminately distributed among the conjectured hypotheses and auxiliary assumptions.

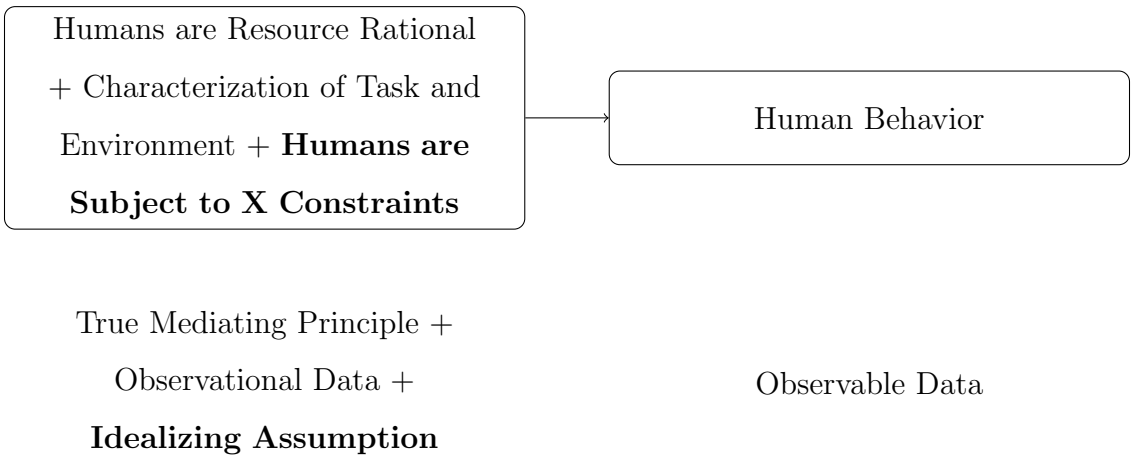

Figure 7: Closing the loop. Inductive risk is isolated to the idealizing assumption. Note that no conjectures are made about human cognition when only minimal (universal) constraints are initially supposed–other than that there are no further constraints. When prediction and observation disagree, the culprit must be the claim about constraints, which is then falsified.

conjecturing about local facts, one limits inductive risk as much as possible to the mediating assumption (e.g., the Earth being an elastic medium and the applicability of continuum mechanics). This means the mediating principle should be true or very nearly true.

In resource rational analysis, the mediating principle is that humans are resource rational. Scientists can investigate optimal cognitive strategies relative to evaluative standards, environments, and cognitive constraints without needing to assume anything specific about human cognition. Demonstrations of resource rational strategies provide relations between strategies, evaluative standards, environments, and cognitive constraints that hold generically for all agents. A nice example of such demonstrations comes from R. L. Lewis et al. (2014), who prove certain programs (cognitive strategies) are optimal *if and only if* certain utility functions (evaluative standards), bounds (cognitive constraints), and environments obtain. The simulations by Icard (2014) in the previous section provide another similar case—strategies are shown to be optimal relative to particular constraints, without any conjectures about human cognition entering into the analysis. Once one has generic relations between constraints, strategies, standards, and environments, the assumption that humans are resource rational makes such relations applicable for making inferences about human cognitive systems.

As mentioned, one wants a mediating principle that is true. This is yet another reason to prefer my maximally broad notion of what can count as a cognitive constraint, which has been defended throughout this dissertation. By adopting a maximally general notion, one makes all systems trivially resource rational. This makes the mediating assumption of resource rationality trivially true—though no less useful. On the other hand, if one adopts a narrow notion, the question of whether humans are making optimal use of their limited resources becomes an empirical question. This would make the mediating assumption potentially (and

likely) false.

There is no mystery as to why using mediating principles can be a good strategy in science. By making well-evidenced general principles applicable to an otherwise mysterious system, one reduces the system's mystery. In ideal cases, one can use the mediating assumption to establish strong "if and only if" relations that allow for mediating deductive inferences from observable data to unobservable theoretical claims of interest within the domain of study.

Turning now to the second principle, the "idealizing principle" is an assumption about the black box that provides an initial starting point. In the geophysics example, the idealizing principle is that the Earth is of uniform density. There must be some starting point in order to apply the mediating principle to the system. One cannot simply assume the Earth is an elastic medium; one must also make some starting assumption about the unobservable structure to derive predictions about observable data. This allows meaningful deviations to emerge and be iteratively addressed. The idealizing principle is intended to be iteratively de-idealized.

The idealizing principle in resource rational analysis is that humans are subject to minimal cognitive constraints. This allows for an initial derivation of resource rational behavior, from which deviations emerge, indicating further constraints in an iterative de-idealization process.

Whereas the justification for using mediating principles is straightforward (make use of general truths to deduce relevant facts if you can), the justification behind the idealizing principle is less clear. Why take a particular starting point? Why is de-idealization a good research strategy? Why, if at all, does it result in well-evidenced models in domains that are otherwise recalcitrant to inquiry?

I argue that the idealizing principle in the closing the loop methodology can be given a topological learning theoretic justification (K. Kelly, 2024). The idealizing principle results in a preference for more falsifiable hypotheses.[4] The dynamic strategy, therefore, facilitates a process of progressive falsification, enabling convergence to the truth. (K. Kelly, 2024).

The mediating and idealizing principles work together to avoid the limitations of the method of hypotheses, specifically by managing the problem of the indiscriminateness of confirmation. They achieve this by isolating inductive risk to particular posits at a time. The mediating principle makes a set of well-established facts from outside the local domain of interest bear on the local domain, while the remaining inductive risk is isolated via the idealizing principle to specific claims about the black box that can be given a learning theoretic justification.

In the following section, I lay out the topological learning theory framework to demonstrate that hypotheses claiming agents are more resource rational (subject to fewer constraints) are topologically simpler. This provides a rationale for the idealizing principle and the closing the loop strategy.

4.6 Topological Analysis

The following topological framework is based on the work of Genin (2018), Genin and Kelly (2019), and K. Kelly (2024). What is novel here is my application of this framework to cognitive science and my use of it to justify resource rational analysis as a methodological strategy in cognitive science.

Consider a standard propositional space. W is a set of worlds w, where a world is

[4]As will be made clear with the topological analysis, it is not exactly greater *falsifiability* but greater topological simplicity. But to understand this, I will need to set up the formal framework first.

a complete assignment of truth values to all propositions of interest. The set of worlds represents the relevant epistemic possibilities. Propositions can be identified with the set of possible worlds in which the proposition is true. Some parts of the world are observable, and some are not. The observable parts are information states afforded by world w. In this setting, the empirical information one receives is a proposition. I is the union of all information states afforded by worlds w, i.e., $\bigcup_{w \in W} I(w)$. A scientific question Q can be understood as a partition of W where each cell is a different answer. An empirical problem can be understood as a triple (W, I, Q): the question of interest, the different epistemic possibilities, and the data that one has to go on to answer the question.

In cognitive science, worlds or relevant epistemic possibilities correspond to different ways the mind might work. As a simplifying assumption, I will take the possible information states to be human behavior (as opposed to physiological data). These observable behaviors include choices such as moves in games, answers to questions, elicited credences, and similar actions

A method M is a function from information states to propositions. A method gives a relevant response to a question if it maps information states to disjunctions of answers to the question.

A method verifies a proposition (e.g., a hypothesis) if and only if the method converges infallibly to that proposition in the limit if it is true. A proposition is verifiable if there exists a method that verifies it. Genin and Kelly (2019) prove that the verifiable propositions correspond exactly to the open sets in the topology generated from I (that is, the set I closed under arbitrary unions and finite intersections). The intuition behind this result is that information states are verifiable propositions and arbitrary disjunctions of verifiable propositions are verifiable, but only finite conjunctions of verifiable propositions are verifi-

able. Refutable propositions are propositions for which a method exists that can verify the complement of the proposition, and they are equivalent to the closed sets in the information topology (that is, the set I closed under arbitrary intersections and finite unions). The intuition behind this result is that complements of verifiable propositions are refutable and arbitrary conjunctions of refutable propositions are refutable, but only finite disjunctions of refutable propositions are refutable.

For these results to hold, I must satisfy certain axioms to serve as a topological basis from which a topology can be generated. The axioms are as follows:

Axiom I.1: $I(w) \neq \emptyset$

Axiom I.2: For each $E, F \in I(w)$, there exists $G \in I(w)$ such that $G \subseteq E \cap F$.

Axiom I.3: I is countable.

For this framework to be applicable to cognitive science, it must be shown that the information states cognitive scientists work with satisfy these axioms. Axioms I.1 and I.3 are easily satisfied across virtually all contexts of scientific inquiry. I.1 states that there will always be some attainable information (in the worst case, this is trivial information W). I.3 states that the empirical information one receives is always countable, which is true "since any language in which the data are recorded is at most countably infinite" (Genin, 2018, p. 30).

The difficult basis axiom to prove is I.2. This states that information accumulates: it's never the case that obtaining some piece of data prevents you in the future from obtaining some other data that you could have attained. This would be violated, for example, if you only had a single sample of some substance to perform a chemical reaction on, such that after performing an experiment, there's no way to achieve a restoration of initial conditions and perform a different experiment on it. Unfortunately, this is starting to sound a lot like

psychology.

If a scientist is interested in a particular subject's ability to memorize a list of items under different conditions, once the subject has undergone training with the stimuli, it is not possible to observe what the subject's recall performance would have been if they had learned it in a different way or setting. The obvious solution is to conduct between-subject experiments and shift the focus from what a particular subject's psychological nature is like to what human psychology is like in general. Psychometrics and the study of individual differences pose unique challenges in this regard.[5]

To the extent that there is a common psychological structure, inquiring into this structure allows for the satisfaction of I.2. Additional observations of some behavior under potentially new conditions are always possible, as the restoration of initial conditions can be achieved by drawing on a new subject. The fact that the topological analysis cannot apply to all questions of psychology due to violations of I.2 is, in my view, a virtue rather than a vice. Instead of justifying methods too broadly, the topology points to specific areas where a special kind of knowledge can be attained. In doing so, it becomes prescriptive: attempt to ask questions and shape inquiry in a way that satisfies I.2.

Once the basis axioms are satisfied for the possible information states relevant to an empirical problem, the topology generated from this basis will yield the verifiable propositions. Taking the complement of every verifiable proposition will then yield the refutable propositions.

Most scientific hypotheses of interest are neither verifiable nor refutable. However, by connecting these concepts to topology, one can generate a rich landscape of new concepts from

[5]In classical test theory, for example, theorists are explicit that a thought experiment must be engaged in to provide an interpretation consistent with the mathematics being used: either that the subject of interest is initially cloned multiple times, or that they are brainwashed–cleaned like test tubes–in between measurements (Borsboom, 2005).

the formal framework and prove various properties of them. Kelly's insight is that while many scientific hypotheses are neither open (verifiable) nor closed (refutable), many important hypotheses are locally closed (verifutable). A locally closed proposition is a conjunction of an open and a closed proposition. Thus, it has the epistemic character that if it is true, it will become refutable (because the verifiable conjunct will eventually be verified, leaving only the refutable conjunct).

To use the standard example from Kelly's work: suppose one is interested in the polynomial degree of a curve generated by some scientific phenomenon (where one receives open intervals of this curve as their information states). The hypothesis that the law governing this process is polynomial degree n is neither verifiable nor refutable, but it is verifutable, since it is a conjunction of the hypothesis that the curve is at least polynomial degree n (verifiable) and the hypothesis that it is at most n (refutable).

The hypothesis "at least n" is verifiable because if it is true, eventually data will deductively entail that it is true. It is not refutable because if it is false, no data will ever rule it out. The hypothesis "at most n" is not verifiable, because if it is true, no data will ever deductively entail that it is true. But it is refutable, because if it is false, eventually data will show that it is false.

Therefore, the hypothesis that the polynomial degree is exactly n has the property that, if it is true, eventually you will receive information showing that it is refutable.

Once one recognizes that the different hypotheses under consideration are all locally closed (i.e., that there is a locally closed cover of W), an ordering of these hypotheses naturally emerges. One can think of the verifiable conjunct of a verifutable proposition as the trigger and the refutable conjunct as the defeater. Notice that in the polynomial degree case, the trigger for the hypothesis 'polynomial degree n' is the defeater for the hypothesis

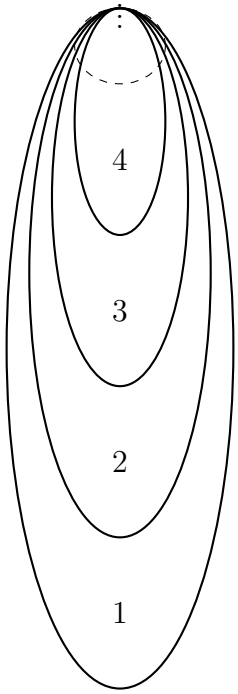

Figure 8: Topological relations of hypotheses concerning polynomial degree. If the true hypothesis is polynomial degree n, then polynomial degree n+1 is false but will never be refuted. The reverse is not true. Lower degrees are therefore nested in the frontiers of higher degrees.

'polynomial degree n1'. The reverse is not true. Due to space constraints here, I must obscure the interesting details but the basic idea is that for there to be a locally closed cover on the hypothesis space, defeaters must be triggers for other locally closed hypotheses, chaining together the set of hypotheses. This situation results in an ordering that arises from these asymmetries between verifutable propositions.

This ordering captures the intuitive idea that lower polynomial degree hypotheses are "simpler" than higher ones (e.g., a quadratic law is simpler than a cubic law). Genin and Kelly (2019) use this idea to define a relation of "simplicity" as follows:

Simplicity Relation: A $\lhd B \iff A \subseteq \text{frnt}\, B$

Where frnt B is the topological frontier of B. If a world is in the frontier of B, it means that B is false but will never be refuted. If the true law is quadratic, the cubic hypothesis will never be refuted. Note that the polynomial degree hypotheses are all nested in this way, with lower polynomial degrees all being in the frontier of higher degrees (figure 4).

Note that the proposed definition of simplicity amends Popper's insight: simplicity is not just about being more refutable. The uninformative state W is not refutable, so if simplicity were just a matter of which propositions are more refutable, then W would be considered maximally complex. W is not refutable but it is true. The danger lies in propositions that are not refutable but are false, and this is captured in the definition of simplicity (Genin, 2018; Genin and Kelly, 2019; K. Kelly, 2024).

Now that a simplicity relation is defined, it can be shown that this is a transitive, asymmetric relation and can induce an ordering on a set of hypotheses. A method can be called Ockham if it always outputs the simplest hypothesis consistent with the data. Additionally:

> A method M is a solution to Q if it converges, on increasing information, to the true answer in Q, i.e., for every $w \in W$, there exists $E \in I(w)$ such that $M(F) \subseteq Q(w)$ for all $F \in I(w)$ entailing E. A problem is solvable if it has a solution. (Genin, 2018, p. 38)

Furthermore, a method is progressive if, once it outputs the true answer, it never outputs a different answer afterward. That is, it never abandons the truth once it finds it. The important result is that if the answers to a question Q can be enumerated in an order that agrees with the simplicity order, then there exists a progressive solution, and this progressive solution will necessarily be an Ockham method (Genin, 2018).

The upshot is that, in science, if one can ask a question in such a way that a simplicity order is induced, then one can prefer simpler answers and be assured that the data will drive one to converge progressively to the truth.

In perhaps most scientific contexts, no simplicity order is induced. For example, in cognitive science, a scientific question of interest is whether some knowledge, such as details of mental grammar, is innate or learned from experience through domain-general learning mechanisms. Neither hypothesis is topologically simpler than the other. Another scientific question of interest is whether humans categorize new stimuli by comparing their features to existing examples from different categories (exemplar theory), comparing them to an idealized prototype for each category (prototype theory), or using some hybrid or alternative approach. None of these hypotheses are simpler than the others.

Science is hard, and answering the above cognitive science questions is challenging. It would be a reductio of the topological analysis if it claimed there was a simple recipe to follow for every scientific question of interest that would lead scientists to the truth. The fact that simplicity orderings of this sort are rare in nature aligns with the reality that progress in science has been and continues to be difficult. The rarity of simplicity orders corresponds with the history of science and the state of contemporary science: many empirical questions have not yielded to progress, and those areas that have often did so only after very clever ways of asking the right questions were devised. Again, the topological analysis is prescriptive:

ask questions in cognitive science whose answers will result in a simplicity order.

I claim that the virtue of resource rational analysis as a methodology in cognitive science is that it asks the right question. It asks a question such that, when understood correctly, the answers—which reveal facts about psychological processes—can be enumerated in accordance with a simplicity order. This method can then be justified as being guaranteed to progressively converge to the truth in the limit.

The question resource rational analysis asks is, "How rational are human minds?" With a proper understanding of rationality, a topological analysis can demonstrate that hypotheses suggesting humans are more rational are simpler. Such an analysis justifies methods that have a preference for these hypotheses.

Consider first a simplified, binary question: are humans ideally rational or not? Here, there are two epistemic possibilities corresponding to the answers to this binary question:

$$W = \{w_1, w_2\}$$

where:

w_1 : Humans are ideally rational

w_2 : Humans are not ideally rational

For ease of exposition, suppose that by "ideally rational," we mean that humans conform to the norms of logic and probability. The information states are the possible behaviors that one could observe. For example, experimentalists might give human subjects tasks involving assigning credences to various hypotheses as information is given to them, to see if subjects update their beliefs in conformance with Bayes' theorem.

Suppose, more specifically, the task is the categorization task discussed above, where subjects must infer the probability of a new object possessing certain properties based on

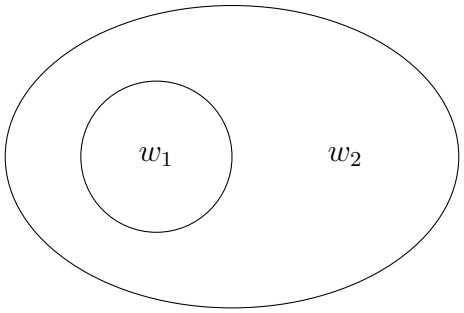

Figure 9: Topological analysis of two hypotheses: w_1: Humans are ideally rational. w_2: Humans are not ideally rational

previous objects. The information states are the credences the subject assigns to new objects possessing certain properties.

Consider an arbitrary information state E that is compatible with w_1. E is a case of a subject reporting a credence in line with Bayes' theorem. E is compatible with w_2 as well: it is always possible that, in the future, an information state F will be observed in which a subject's credence disagrees with Bayesian norms and deductively rules out w_1.

Suppose that you continue to test the subject on the categorization task, using increasingly complex and numerous stimuli, and subjects continue to assign the correct credences. You can never rule out that at some point, the stimuli will become so large and the computation so complex that humans will cease to be able to perform the correct computation. Thus, all information that one could receive that is compatible with w_1 is compatible with w_2, but not all information compatible with w_2 is compatible with w_1. w_1 is refutable but not verifiable, whereas w_2 is verifiable but not refutable. If w_1 is true, w_2 is false but will never be refuted. Topologically, this situation corresponds to the standard Sierpiński space:

The situation is similar to the "all ravens are black" case that has been analyzed within

the same learning theoretic framework (Schulte, 2023). If the question is "Are all ravens black?" the method that says "no" forever is always correct if the true answer is no, but it never finds or stabilizes to the truth if the true answer is yes. In contrast, the method that says "yes" forever until it sees a counterexample and then says "no" forever will eventually stabilize to the true answer in either case. This asymmetry is a reason to adopt the latter method.

If the question is "Are humans ideally rational?" the method that says "no" forever is always correct if the true answer is no, but it never finds or stabilizes to the truth if the true answer is yes. In contrast, the method that says "yes" forever until it sees a counterexample and then says "no" forever will eventually stabilize to the true answer in either case. This asymmetry is, therefore, a reason to adopt the method that supposes that humans are ideally rational until proven otherwise.

Well, that is all fun and games. But what good is this to cognitive science? It is easy enough to elicit irrationality from humans, thereby falsifying the hypothesis that they are ideally rational. What is needed now is a way to generalize this method so it can be iterated repeatedly. The way to do this is to ask a different question: "How rational are humans?" in such a way that being more rational is topologically simpler.

Standard methods for comparing the rationality of non-ideal agents do not produce a notion of "more" rational that is topologically simpler. For example, comparing agents' rationality by evaluating their performance based on quantitative scores (e.g., measures of accuracy, degree of coherence, dollars won, etc.) does not result in a topologically simpler understanding of being more rational.

Consider the question, "How rational is this agent?' where the answers specify how calibrated an agent's credences are. Either the question pertains to performance on a particular

finite test, in which case it is decidable and there is no inductive problem, or it concerns how calibrated an agent's beliefs are generally in the open world, in which case a demon argument can be made against any method. No information will ever refute a particular hypothesis about an agent's calibration (other than the hypothesis of perfect calibration). Information can make it appear that an agent's calibration is any value for an arbitrarily long time, forcing a method to conclude that this is the correct calibration, only for information to change this value indefinitely afterward. Similar arguments can be constructed against other similar notions of rationality.

Consider, however, my view. Recall that, on my view, all agents are resource rational; they are all constrained in various ways, and it is arbitrary to distinguish between kinds of constraints. Therefore, a maximally general notion of constraint should be adopted, which allows for any psychological fact that lessens performance on a task in an environment to count as a constraint. The consequence is that all agents are doing their best relative to their cognitive constraints. One agent can be said to be "more" rational than another if they are subject to strictly fewer constraints—more specifically, if the former agent's constraints are a proper subset of the latter agent's.

Rationality Relation $\succeq_R$: One agent can be said to be "more" rational than another if they are subject to strictly fewer constraints—more specifically, if the former agent's constraints are a proper subset of the latter agent's.

Being subject to qualitatively different constraints is, on this view, not comparable. If one agent has bounded memory and another has bounded metabolic resources for computation, it makes little sense to ask which one is more constrained. Even if one constraint seems to hamper performance more than another, if they affect different performance areas, tasks, and environments, it is not clear how to compare them. Perhaps a principled systematic

theory for comparing certain qualitatively different constraints can be established, and then it can be seen how this would affect the topological analysis. For now, I stick to the idea that the only time agents can be clearly compared in terms of their rationality is if one agent's constraints are a subset of another's. If an agent is subject to both memory and metabolic constraints, then that agent is clearly more constrained than an agent subject to only one of those constraints.

A consequence of this view is that the "more rational" relation induces a partial order. This partial order is a virtue of this account and, I suspect, necessary to allow for a simplicity order to be induced from rationality comparisons. If the rationality of all agents were fully comparable, then a simplicity order would not be induced. There are no inductive asymmetries to exploit between the hypothesis that an agent has bounded memory and the hypothesis that they are metabolically bounded. The fact that interval scales of rationality, such as calibration scores, allow for total orders is plausibly part of the reason they do not induce simplicity orders. Simplicity relations are rare, so a rationality comparison relation should be selective if hypotheses that posit greater rationality are to be simpler.

Note that constraints, by definition, hamper performance. If one agent is performing better on a task than another agent, but the former agent has more limited memory capacity, that fact does not count as a constraint in that context. Thus, cases where someone performs as well or better with fewer resources and is therefore arguably more praiseworthy (Morton, 2017) involve a different concept of constraints. Only if limited memory results in decreased performance does it count as a constraint. Constraints are always individuated relative to a task and environment. The same physical finite memory is a constraint relative to a task and environment where it hampers performance, but not relative to a different context where it does not.

This makes the identification of constraints necessarily theory-mediated. To see a constraint as a constraint, you must be able to recognize decreased performance. Decreased performance is a deviation from a theoretical ideal and is therefore a second-order phenomenon– much like the 43 arc-second discrepancy in the precession of Mercury's perihelion that served as evidence for Einstein's theory was a second-order phenomena (Smith, 2014). Both are not directly observable but emerge by comparing observation to theory.

In this context, decreased performance is measured relative to a standard set by resource rationality, which itself assumes the presence of certain constraints. Material constraints are limitations on a cognitive system that prevent it from performing resource rationally. This is a recursive definition. The base case: when there are no constraints, resource rationality is equivalent to ideal rationality; resource rational performance in this case represents the best possible performance relative to an unrestricted pool of agents.[6] The recursive case: this involves material constraints that prevent an agent from achieving the best performance relative to a pool of agents already subject to particular material constraints. Resource rationality is evaluated against a backdrop of these existing constraints, creating a layered evaluation where each layer of constraints defines the context for evaluating the next.

Given this understanding, a particular constraint may only be intelligible relative to other constraints, which are only intelligible relative to other constraints, and so on. This is another reason why arbitrary comparison of constraints is not feasible.

To illustrate this point, consider the example explored in the previous section. It was noted that, given certain metabolic costs of sampling and available resources, there is an optimal number of samples to use to approximate Bayesian inference. Inappropriate sampling relative to available metabolic resources (c_4 in Table 3) is therefore a constraint. However,

[6]If you accept the arguments of Chapter 1, this case must in fact be relative to a pool of agents who are subject to minimal constraints, such as those that must hold universally for any real-world agent.

this can only be understood as a constraint by noting a deviation from resource rational performance, specifically while assuming the prior constraint that an approximation method must be used. It makes no sense to discuss the optimal number of samples unless one has already restricted the pool of agents under consideration to those so constrained that they must approximate in certain ways in the first place.

Constraint	**Description**	**Resource-Rational Strategy**
c_1: Bounded perception	Inability to perceive facts in the environment, creates need to infer	Bayesian inference
c_2: Finite memory	Inability to calculate Bayesian inference for complex stimuli, creates need to approximate Bayes	Approximating Bayesian inference as best as possible with finite space
c_3: Finite metabolic resources	Approximating with too few samples relative to finite memory because sampling is metabolically costly	Approximating Bayesian inference as best as possible with finite space and metabolic resources
c_4: Inappropriate metabolic resource allocation	Using too few samples given available metabolic resources	Approximating Bayesian inference as best as possible with finite space and allocated metabolic resources

Table 3: Constraints and Resource Rational Strategies

That said, new constraints need not always be intelligible only in relation to all previous constraints. New constraints can emerge for an agent in the context of a new task or environment. Constraints, as mentioned, are always individuated relative to specific tasks

and environments—since they are defined by deviations from resource rationality, which is always relative to a task and environment. For example, consider the Galapagos insect case from Chapter 1. There, if an individual had decreased perceptual capacity (a material constraint), a heuristic TTB cognitive strategy was more resource rational than a full-blown Bayesian strategy. Thus, performing Bayesian inference is suboptimal relative to that task and environment and reflects a constraint (inflexibility of strategy selection/overgeneralizing the strategy of Bayesian inference). In this context, finite memory resulting in not being able to compute full-blown Bayesian inference is not a constraint. However, this does not negate finite memory from being a constraint in other contexts. While material constraints reflect facts about psychology, they are still indexed to specific problems and environments. Due to this, constraints are always additive; once a constraint is uncovered, it will never be found that it was not, in fact, a constraint.

Returning to the example of categorization, consider the relevant worlds (hypotheses) under consideration:

$$W = \{w_1, w_2, w_3, w_4\}$$

where:

$w_1 : c_1$ (Bayesian Agent)

$w_2 : c_1 + c_2$ (Approximate Bayesian Agent)

$w_3 : c_1 + c_2 + c_3$ (Weaker Approximate Bayesian Agent)

$w_3 : c_1 + c_2 + c_3 + c_4$ (Even Weaker Approximate Bayesian Agent)

Given the definition of more resource rational: $w_1 \succeq_R w_2 \succeq_R w_3 \succeq_R w_4$

Notice that the proper subset notion of more rational results in the relation holding over

nested hypotheses, just like the case of polynomial degrees, where:

$$w_1 : \beta X^2 + \alpha X = Y$$

$$w_2 : \gamma X^3 + \beta X^2 + \alpha X = Y$$

$$w_3 : \delta X^4 + \gamma X^3 + \beta X^2 + \alpha = YX$$

I will now prove the following: $w_m \succeq_R w_n \implies w_m \lhd w_n$ (more rational hypotheses are simpler).

Just as the hypothesis that the true law is a polynomial of degree exactly n is a conjunction of the hypotheses that the degree is at least n and at most n, the hypothesis that an agent is resource rational relative to certain constraints amounts to the claim that the agent is "subject to exactly these constraints." This hypothesis is a conjunction of "subject to at least this many constraints" and "subject to no more than this many constraints."

"At least this many constraints" is verifiable: for example, if an agent is subject to at least c_1 and c_2, eventually the data will show that the agent is incapable of calculating Bayesian inference perfectly. "At most this many constraints" is refutable: for example, if the claim is that an agent is subject to at most c_1 and c_2, then eventually, if the agent is subject to c_3, data will show that the approximation is suboptimal relative to c_1 and c_2, indicating that c_3 is true and that "at most this many constraints" is false. Hence, "subject to exactly these constraints" is verifutable.[7]

The trigger for being subject to n constraints is a defeater for being subject to $n-1$ constraints. Without loss of generality, the trigger for w_3 is the defeater for w_2. If w_2 is the true world, w_3 is false but will never be refuted. If w_3 is the true world, w_2 is false and will be refuted, and w_3 will become refutable. Thus, w_2 is in the frontier of w_3 and is simpler

[7]This makes resource rational analysis a "paradigm" in K. Kelly (2024)'s sense: paradigms are sigma-constructible propositions, which are countable disjunctions of locally closed propositions.

than w_3. Thus, if $w_m \succeq_R w_n$ then $w_m \lhd w_n$.

So now consider a method that violates the preference for more rational hypotheses. Such a method jumps the gun and supposes that an agent is subject to a constraint n before the trigger for w_n has been seen. Either this method will never converge to the truth, or it can be forced to drop the truth once the truth has been hypothesized. That is, either it is not a solution method or it is not a progressive solution method. Suppose on the one hand, that after seeing data for an arbitrarily long time that suggests there is no constraint n, this method always sticks to its guns and maintains n. It will never conjecture the truth in any world w_m where $m < n$. Thus it would not be a solution. Suppose, on the other hand, that the method is such that after seeing data for an arbitrarily long time consistent with there being no constraint n, it changes its answer to w_m for some $m < n$. It is then always possible for nature to force such a method to always drop the truth: suppose the true world is w_n and the method happens to answer w_n before seeing data that triggers this hypothesis. Nature then shows data that is consistent with w_m where $m < n$ until the method changes its answer to w_m, thereby dropping the truth. Then nature eventually shows dating showing that w_n is triggered after all, forcing the method into a cycle of belief change.

Thus, a simplicity order can be induced from the partial order formed by the "more resource rational" relation. Given the Ockham results of Genin (2018) and Genin and Kelly (2019), assuming it is possible to enumerate resource rational hypotheses from more rational to less rational (e.g., enumerate further constraints in response to deviations from irrationalities), then a method that prefers more rational hypotheses will be a progressive solution to the question "how resource rational" is this agent? By framing inquiry into cognition as a question about how constrained agents are, resource rational analysis ensures progressive falsification of hypotheses, such that the data will eventually drive the cognitive scientist to

the truth.

Facts about constraints are, therefore, learnable in the learning theoretic sense. Constraints reflect facts about agents' psychologies. The bet that resource rationality makes is that most psychological process details can be discovered in the form of constraints. That is, the fixed structure of cognition—although conducive to rational behavior in most contexts—will result in mistakes in other contexts. If humans were "ideally" rational, there would be no constraints to learn, and this methodology would reveal little about human psychology. The resource rational analyst, therefore, hopes that for limited beings, nearly every design choice of human cognition limits the kinds of strategies it can employ in certain ways in certain contexts, eventually revealing itself as a constraint.

4.7 Conclusion

Theories of rationality can provide a body of well-evidenced theoretical statements to help turn observable human behavioral data into evidence for cognitive models. In resource rational analysis, what I have been calling the mediating principle is that humans are resource rational. This principle makes theoretical demonstrations of resource rationality applicable to human beings. What I have been calling the idealizing principle is that humans are minimally constrained. Together, the mediating and idealizing principles work to provide a methodological strategy in cognitive science: an iterative process of de-idealization, resulting in the discovery and testing of material facts about human psychology.

Topological learning theory provides a rationale for adopting this idealizing principle. One last thing to note is that the idealizing principle presupposes the mediating principle. One must assume humans are resource rational before assuming that humans are more re-

source rational than not. As mentioned, simplicity orders are rare but desirable in empirical inquiry. So another perspective on the mediating principle is that it is necessary for a simplicity order to be induced—the idealizing principle then prescribes a preference for simplicity. This shows how the topological learning theory framework can be prescriptive for science: when there are no apparent simplicity orders, find a way to create one! The way to do this is by asking the right questions. In cognitive science, I have argued that the right question is, "How resource rational are we?"

5.0 Conclusion

One last example before closing down the show. Signal Detection Theory (SDT)–which has already come up in this dissertation a couple of times–involves analyzing a subject's ability to distinguish between signal (relevant stimuli) and noise (irrelevant stimuli) through a series of trials where a subject responds to the presence or absence of a signal. It quantitatively assesses sensitivity (d')—the ability to detect the signal—and decision criteria (β)—the threshold at which a subject decides whether a signal is present. For example, a $\beta > 1$ indicates a conservative bias (the decision threshold is higher, meaning fewer false alarms but more misses). A $\beta < 1$ indicates a liberal bias (the decision threshold is lower, meaning more hits but also more false alarms). And a β of 1 indicates no bias.

Townsend (2008) writes that Signal Detection Theory "...stands as the prototypical theory-driven methodology, since it can be employed to discern decision and learning bias from 'true' sensory or sensitivity (e.g., signal-to-noise ratio) effects in such diverse fields as hypnotic phenomena, to trial-witness memory, to laboratory psychophysics or learning and cognition experiments" (p. 6). In other words, the mathematics of SDT allows psychologists to take observable human behavioral data (e.g., button pushes in response to stimuli) and deductively derive unobservable theoretical parameters of interest in cognitive models, such as decision thresholds.

I raise this example to think about potential alternatives to the view I have defended throughout this dissertation. Might there be other mathematically precise theories that can be established independently of any conjectures about human cognition and that can subsequently be used to turn observable data into deductive evidence for cognitive claims? SDT seems to be one. To interpret SDT as utilizing rationality considerations would be to

stretch my claims too far.

Yes, sensitivity (d') can constitute suboptimality of perception. Yes, a decision criterion (β) can likewise be suboptimal relative to the costs of hits, misses, and the prior probabilities of signal versus noise. This is just to say that any psychological fact can be a cognitive constraint. That part I stand by. But that is different from saying that SDT is a case of theoretical demonstrations of optimality being used to turn observable data into evidence for cognitive parameters. And no–the fact that misses are misses and, therefore, errors does not mean SDT is a case of resource rational analysis. One does not, in SDT, calculate the optimal amount of hits versus misses given a particular sensitivity, and then use deviations from that behavior to calculate the bias. The bias (β) is just calculated directly from the amount of hits and misses.

SDT therefore, in my estimation, stands as a beautiful example of how cognitive science should be done–but an example that is not at all a case of resource rational analysis. Why do I conclude with such an example? Three reasons. The first is to show that resource rational analysis is not so broad as to be meaningless. I have already given reasons to show this point, but it is worth emphasizing that not everything is resource rational analysis.

The second reason is to suggest that resource rationality can always use an "acquire and consolidate" approach to such cases. Wherever there are mathematically precise theories that allow for theory-mediated measurement of cognitive parameters of interest, integrating these theories into a resource rational analysis approach is always beneficial. For example, with Signal Detection Theory (SDT), once (β) has been derived, one can do two things.

First, one can incorporate this psychological fact into a resource rational analysis and infer new resource rational behaviors relative to such decision criteria to uncover new deviations from this calculated behavior. More psychological facts lead to more inferences about

resource rational behavior, which in turn lead to more deviations that indicate interesting further cognitive facts.

Second, one can calculate an optimal (β) relative to an agent's cost functions and priors concerning hits and misses. If a psychologist can independently identify the actual costs and priors the agent is employing, one can then determine what decision criterion would be optimal relative to these. If that optimal (β) differs from the actual (β), this discrepancy constitutes a new deviation from resource rationality that can indicate further unaccounted-for details that make a difference to cognition. That might mean further unaccounted-for costs, or time pressures, or it might just mean an inflexible decision criterion that reflects fixed psychological structure–although this latter option is unlikely since decision criteria tend to change in response to manipulation of the other variables (Atkinson, 1963). In either case, resource rational analysis extends what would otherwise be a single step—infer sensitivity and decision criteria using SDT—into an iterative process of investigation. Thus, while not everything in good psychological science makes use of rationality considerations, nearly everything could be used as part of a resource rational methodology.

The third and final reason I use this example is that psychologists *have* historically used optimality considerations in research with SDT. Atkinson (1963), discussing typical research with SDT, writes,

> The position of the criterion (the operating level) is assumed to be under the control of the observer and to vary as a function of psychological variables that influence motivation and set. Specifically, the subject fixes the operating level in terms of a priori probabilities of stimuli and the costs associated with the various choices in such a way as to maximize his expected utility. (p. 103)

In other words, it is not uncommon for psychologists to simply assume that the decision criterion is optimally set relative to costs and priors. Notice how the use of such optimality

assumptions differs from the prescriptions laid out in this dissertation. This point emphasizes how common it is for psychologists to tacitly adopt optimality considerations but in fairly unprincipled ways. So I conclude with this: you may not like rationality considerations in psychological science, but they are ubiquitous. So they might as well be employed in a systematic and philosophically informed manner.

In Chapter 1, I argued for resource rationality as the appropriate account for evaluating and prescribing human behavior. This chapter argued that all epistemic norms are relative to cognitive constraints, advocating for a broad view of what constitutes these constraints. Chapter 2 discussed why rationality is integral to the study of the mind, arguing that, at the very least, coming to know about intentional states requires us to adopt a rationality assumption. Chapter 3 extended these ideas to normative commitments, proposing that a meta-reflective capacity—maintaining resource rationality under varying resource conditions—is a necessary and sufficient condition for possessing a normative commitment. This perspective provides a framework for endowing AI systems with normative commitments and for empirically investigating these commitments in humans and animals.

The final chapter presented resource rational analysis as a methodological strategy and argued for its effectiveness in cognitive science. The epistemic justification for this approach, grounded in part in topological learning theory, supports the iterative de-idealization process and demonstrates the utility of rationality considerations in cognitive science.

This dissertation has, throughout all of these chapters, argued for the many benefits of embracing a maximally broad view of constraints and sought to establish a principled framework for using rationality in cognitive science.

6.0 Bibliography

Anderson, J. R. (1990). *The adaptive character of thought*. Psychology Press.

Andrews, K., Fitzpatrick, S., & Westra, E. (2024). Human and nonhuman norms: A dimensional framework. *Philosophical Transactions of the Royal Society B*, *379*(1897), 20230026.

Arrow, K. (2004). Is bounded rationality unboundedly rational? some ruminations. In M. Augier & J. G. March (Eds.), *Models of a man: Essays in memory of herbert a. simon*. MIT Press.

Atkinson, R. C. (1963). A variable sensitivity theory of signal detection. *Psychological Review*, *70*(1), 91.

Batterman, R. W. (2009). Idealization and modeling. *Synthese*, *169*, 427–446.

Bekoff, M. (2013). *Why dogs hump and bees get depressed: The fascinating science of animal intelligence, emotions, friendship, and conservation*. New World Library.

Bernthal, J. E., Bankson, N. W., & Flipsen, P. (2013). *Articulation and phonological disorders*. Pearson Higher Education.

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

Bilgrami, A. (2008). Intentionality and norms. In M. De Caro & D. Macarthur (Eds.), *Naturalism in question*. Harvard University Press. https://doi.org/https://doi.org/10.2307/j.ctv22jnrf5

Boghossian, P. (2014). What is inference? *Philosophical studies*, *169*, 1–18. https://doi.org/https://doi.org/10.1007/s11098-012-9903-x

Boghossian, P. (2019). Inference, agency and responsibility. *Reasoning: New essays on theoretical and practical thinking*, 1001–124. https://doi.org/https://doi.org/10.1093/oso/9780198791478.003.0007

Boring, E. G. (1929). *History of experimental psychology*. Appleton Century Crofts Inc.

Borsboom, D. (2005). *Measuring the mind: Conceptual issues in contemporary psychometrics*. Cambridge University Press.

Bowers, J. S., & Davis, C. J. (2012). Bayesian just-so stories in psychology and neuroscience. *Psychological bulletin*, *138*(3), 389.

Boyd, R., Gintis, H., & Bowles, S. (2010). Coordinated punishment of defectors sustains cooperation and can proliferate when rare. *Science*, *328*(5978), 617–620.

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

Brandom, R. (2014). Intentionality and language: A normative, pragmatist, inferentialist approach. *The Cambridge handbook of linguistic anthropology*, 347–363.

Brooke-Wilson, T. (2023). How is perception tractable? *Philosophical Review*, *132*(2), 239–292. https://doi.org/https://doi.org/10.1215/00318108-10294422

Broome, J. (2014). Normativity in reasoning. *Pacific Philosophical Quarterly*, *95*(4), 622–633. https://doi.org/https://doi.org/10.1111/papq.12050

Brown, T. B., et al. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, *33*, 1877–1901.

Buckner, C. (2013). Morgan's canon, meet hume's dictum: Avoiding anthropofabulation in cross-species comparisons. *Biology & Philosophy*, *28*, 853–871.

Burge, T. (1993). Content preservation. *The Philosophical Review*, *102*(4), 457–488. https://doi.org/https://doi.org/10.2307/1523046

Carr, J. R. (2022). Why ideal epistemology? *Mind*, *131*(524), 1131–1162. https://doi.org/https://doi.org/10.1093/mind/fzab023

Cherniak, C. (1990). *Minimal rationality*. Mit Press.

Conee, E., & Feldman, R. (2004). *Evidentialism: Essays in epistemology*. Oxford Academic. https://doi.org/https://doi.org/10.1093/0199253722.001.0001

Craver, C. F. (2007). *Explaining the brain: Mechanisms and the mosaic unity of neuroscience*. Oxford University Press.

Davidson, D. (1980). Freedom to act [First published in 1973]. In *Essays on actions and events* (pp. 63–82). Clarendon Press.

Davidson, D. (2003). Mental events. In *Essays on actions and events*. Oxford University Press. https://doi.org/10.1093/0199246270.003.0011

Dennett, D. C. (1968). *Content and consciousness*. Routledge.

Dennett, D. C. (1975). Brain writing and mind reading.

Dennett, D. C. (1989). *The intentional stance*. MIT press.

Dennett, D. C. (1991). Real patterns. *Journal of Philosophy*, *88*(1), 27–51.

Dennett, D. C. (2004). *Freedom evolves*. Penguin.

Dennett, D. C. (2008). *Kinds of minds: Toward an understanding of consciousness*. Basic Books.

Dennett, D. C. (2017). *From bacteria to bach and back: The evolution of minds*. WW Norton & Company.

Dennett, D. C., & Caruso, G. D. (2021). *Just deserts: Debating free will*. John Wiley & Sons.

Dorst, K. (2023). Being rational and being wrong. *Philosophers' Imprint*, *23*(1). https://doi.org/https://doi.org/10.3998/phimp.597

Drayson, Z. (2012). The uses and abuses of the personal/subpersonal distinction. *Philosophical perspectives*, *26*, 1–18. https://doi.org/https://doi.org/10.1111/phpe.12014

Du Bois-Reymond, E. (1848/1887). *Über die lebenskraft* (M. Chirimuuta & D. Makovec, Trans.; Vol. 2). Verlag von Veit & Comp.

Dub, R. (2015). The rationality assumption. *Content and Consciousness Revisited: With Replies by Daniel Dennett*, 93–110.

Eberhardt, F., & Danks, D. (2011). Confirmation in the cognitive sciences: The problematic case of bayesian models. *Minds and Machines*, *21*, 389–410. https://doi.org/10.1007/s11023-011-9241-3

Elga, A., & Rayo, A. (2022). Fragmentation and logical omniscience. *Noûs*, *56*(3), 716–741. https://doi.org/https://doi.org/10.1111/nous.12381

Feldman, J. (2017). What are the "true" statistics of the environment? *Cognitive Science*, *41*, 1871–1903. https://doi.org/10.1111/cogs.12444

Finetti, B. d. (1974). Theory of probability: A critical introductory treatment.

Flanagan, O. (1991). *The science of the mind*. MIT press.

Fleig-Goldstein, B. (2018). *The logic of theory-testing in bounded rational analysis* [Master's Thesis]. Stanford University.

Flores, C., & Woodard, E. (2023). Epistemic norms on evidence-gathering. *Philosophical Studies*, 1–25. https://doi.org/https://doi.org/10.1007/s11098-023-01978-8

Frankish, K. (2010). Dual-process and dual-system theories of reasoning. *Philosophy Compass*, *5*(10), 914–926. https://doi.org/https://doi.org/10.1111/j.1747-9991.2010.00330.x

Friedman, J. (2020). The epistemic and the zetetic. *Philosophical review*, *129*(4), 501–536. https://doi.org/https://doi.org/10.1215/00318108-8540918

Genin, K. (2018, September). *The topology of statistical inquiry* [Doctoral dissertation, Carnegie Mellon University, Department of Philosophy] [Email: konstantin.genin@gmail.com].

Genin, K., & Kelly, K. (2019). Theory choice, theory change, and inductive truth-conduciveness. *Studia Logica*, *107*, 949–989.

Gigerenzer, G., & Brighton, H. (2009). Homo heuristicus: Why biased minds make better inferences. *Topics in cognitive science*, *1*(1), 107–143. https://doi.org/10.1111/j.1756-8765.2008.01006.x

Godfrey-Smith, P. (2001). Three kinds of adaptationism. *Adaptationism and optimality*, *122*, 335–357.

Goodfellow, I. J., Shlens, J., & Szegedy, C. (2014). Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*.

Greco, D. (2023). *Idealization in epistemology: A modest modeling approach*. Oxford University Press.

Griffiths, P. (1963). Viii—on belief. *Proceedings of the Aristotelian Society*, *63*(1), 167–186.

Griffiths, T. L., Lieder, F., & Goodman, N. D. (2015). Rational use of cognitive resources: Levels of analysis between the computational and the algorithmic. *Topics in Cognitive Science*, *7*, 217–229. https://doi.org/10.1111/tops.12142

Haugeland, J. (1981). *Semantic engines: An introduction to mind design*.

Hedden, B. (2015). *Reasons without persons: Rationality, identity, and time*. Oxford University Press.

Horowitz, A. (2015). Reading dogs reading us. *Proceedings of the American Philosophical Society*, *159*(2), 141–155.

Huygens, C. (1690). *Traité de la lumière. où sont expliquées les causes de ce qui luy arrive dans la reflexion, & dans la refraction...par c. h[uygens]. ...avec un discours de la*

cause de la pesanteur [Tr/NQ.16.186. [large paper copy; on fly-leaf in Newton's hand 'Is. Newton Donum Nobilissimi Authoris']; sent to Newton via Fatio de Dullier, 24 Feb. 1689/90 [Correspondence, III, 390]]. Leide.

Icard, T. F. (2014). The algorithmic mind: A study of inference in action.

Icard, T. F. (2018). Bayes, bounds, and rational analysis. *Philosophy of Science*, *85*(1), 79–101. https://doi.org/https://doi.org/10.1086/694837

Icard, T. F. (2023). Resource rationality. https://philarchive.org/rec/ICARRT

Jackendoff, R. (2009). *Language, consciousness, culture: Essays on mental structure*. MIT Press.

Jackendoff, R. (2012). *A user's guide to thought and meaning*. Oxford University Press.

Jenkin, Z. (2023). Perceptual learning and reasons-responsiveness. *Noûs*, *57*(2), 481–508. https://doi.org/https://doi.org/10.1111/nous.12425

Jones, M., & Love, B. C. (2011). Bayesian fundamentalism or enlightenment? on the explanatory status and theoretical contributions of bayesian models of cognition. *Behavioral and Brain Sciences*, *34*, 169–188. https://doi.org/10.1017/S0140525X10003134

Joyce, J. M. (1998). A nonpragmatic vindication of probabilism. *Philosophy of science*, *65*(4), 575–603. https://doi.org/https://www.jstor.org/stable/188574

Joyce, J. M. (2009). Accuracy and coherence: Prospects for an alethic epistemology of partial belief. In *Degrees of belief* (pp. 263–297). Springer. https://doi.org/https://doi.org/10.1007/978-1-4020-9198-8_11

Kadane, J. B., Schervish, M. J., & Seidenfeld, T. (1999). *Rethinking the foundations of statistics*. Cambridge University Press.

Karlan, B. (2021). Reasoning with heuristics. *Ratio*, *34*(2), 100–108. https://doi.org/https://doi.org/10.1111/rati.12291

Karlovich, M. W., & Wallisch, P. (2021). Scintillating starbursts: Concentric star polygons induce illusory ray patterns. *i-Perception*, *12*(3), 20416695211018720.

Kelly, K. (1996). *The logic of reliable inquiry*. Oxford University Press.

Kelly, K. (2024, February). *The topology of scientific inquiry* [Unpublished manuscript], Carnegie Mellon University.

Kelly, K., & Schulte, O. (1997). Church's thesis and hume's problem. *Logic and Scientific Methods: Volume One of the Tenth International Congress of Logic, Methodology and Philosophy of Science, Florence, August 1995*, 159–177. https://doi.org/https://doi.org/10.1007/978-94-017-0487-8_9

Kelly, T. (2003). Epistemic rationality as instrumental rationality: A critique. *Philosophy and phenomenological research*, *66*(3), 612–640. https://doi.org/https://doi.org/10.1111/j.1933-1592.2003.tb00281.x

Khan, A., Ceylan, S., van Driel, M., Giardini, D., Lognonné, P., Samuel, H., Schmerr, N. C., Stähler, S. C., Duran, A. C., Huang, Q., et al. (2021). Upper mantle structure of mars from insight seismic data. *Science*, *373*(6553), 434–438.

Khan, S. (2024). Commitment: From hunting to promising. *Biology & Philosophy*, *39*(1), 5.

Klein, C. (2018). Mechanisms, resources, and background conditions. *Biology & Philosophy*, *33*(5-6), 36. https://doi.org/https://doi.org/10.1007/s10539-018-9646-y

Klein, C. (2022). Explaining neural transitions through resource constraints. *Philosophy of Science*, *89*(5), 1196–1202. https://doi.org/https://doi.org/10.1017/psa.2022.35

Korsgaard, C. M. (2010). Reflections on the evolution of morality.

Kotseruba, I., & Tsotsos, J. K. (2020). 40 years of cognitive architectures: Core cognitive abilities and practical applications. *Artificial Intelligence Review*, *53*(1), 17–94. https://doi.org/https://doi.org/10.1007/s10462-018-9646-y

Kripke, S. A. (1982). *Wittgenstein on rules and private language: An elementary exposition*. Harvard University Press.

Langley, P., Laird, J. E., & Rogers, S. (2009). Cognitive architectures: Research issues and challenges. *Cognitive Systems Research*, *10*(2), 141–160. https://doi.org/https://doi.org/10.1016/j.cogsys.2006.07.004

Lewis, D. (1974). Radical interpretation. *Synthese*, 331–344. https://doi.org/https://www.jstor.org/stable/20114928

Lewis, D. (1982). Logic for equivocators. *Noûs*, *16*(3), 431–441. https://doi.org/10.2307/2216219

Lewis, R. L., Howes, A., & Singh, S. (2014). Computational rationality: Linking mechanism and behavior through bounded utility maximization. *Topics in Cognitive Science*, *6*, 279–311. https://doi.org/10.1111/tops.12086

Li, M., Chen, X., Yuan, N., Lu, Y., Liu, Y., Gong, H., Qian, L., Andolina, I. M., Wu, J., Zhang, S., et al. (2022). Effects of acute high intraocular pressure on red-green and blue-yellow cortical color responses in non-human primates. *NeuroImage: Clinical*, *35*, 103092. https://doi.org/10.1016/j.nicl.2022.103092

Lieder, F., & Griffiths, T. L. (2020). Resource-rational analysis: Understanding human cognition as the optimal use of limited computational resources. *Behavioral and Brain Sciences*, *43*.

Lieder, F., Griffiths, T. L., & Goodman, N. D. (2012). Burn-in, bias, and the rationality of anchoring. *Advances in Neural Information Processing Systems 25*, 2790–2798.

Lieder, F., Shenhav, A., Musslick, S., & Griffiths, T. L. (2018). Rational metareasoning and the plasticity of cognitive control. *PLoS computational biology*, *14*(4), e1006043.

MacLeod, C. M. (1991). Half a century of research on the stroop effect: An integrative review. *Psychological bulletin*, *109*(2), 163.

Manski, C. F. (2017). Optimize, satisfice, or choose without deliberation? a simple minimax-regret assessment. *Theory and Decision*, *83*, 155–173.

Marcus, G. (2024, January). Where's waldo? the elephant in the room. https://garymarcus.substack.com/p/wheres-waldo-the-elephant-in-the

Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information*. Henry Holt; Co., Inc.

McClelland, J. L., Rumelhart, D. E., & Hinton, G. E. (1986). The appeal of parallel distributed processing. In D. E. Rumelhart, J. L. McClelland, & the PDP Research Group (Eds.), *Parallel distributed processing, volume 1: Explorations in the microstructure of cognition: Foundations* (pp. 3–44). The MIT Press.

McKenna, R. (2023). *Non-ideal epistemology*. Oxford University Press.

Millar, A. (2004). *Understanding people: Normativity and rationalizing explanation*. Clarendon Press.

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological review*, *63*(2), 81.

Miracchi, L. (2019). When evidence isn't enough: Suspension, evidentialism, and knowledge-first virtue epistemology. *Episteme*, *16*(4), 413–437. https://doi.org/doi:10.1017/epi.2019.34

Miyake, T. (2013). Underdetermination, black boxes, and measurement. *Philosophy of Science*, *80*(5), 697–708.

Morton, J. M. (2017). Reasoning under scarcity. *Australasian Journal of Philosophy*, *95*(3), 543–559.

Munroe, W. (2021). Reasoning, rationality, and representation. *Synthese*, *198*(9), 8323–8345. https://doi.org/https://doi.org/10.1007/s11229-020-02575-6

Musslick, S., & Masís, J. (2023a). Pushing the bounds of bounded optimality and rationality. *Cognitive Science*, *47*(4), e13259.

Musslick, S., & Masís, J. (2023b). Pushing the bounds of bounded optimality and rationality. *Cognitive Science*, *47*(4), e13259.

Newell, A., & Simon, H. A. (1976). Computer science as empirical inquiry: Symbols and search.

Norton, J. D. (2003). A little survey of induction.

Norton, J. D. (2021). *The material theory of induction*. University of Calgary Press.

Norton, J. D. (2023). *The large-scale structure of inductive inference* [Complete manuscript of July 12, 2022. Accepted August, 2023, for publication in BSPSOpen/University of Calgary Press]. BSPSOpen/University of Calgary Press.

Oaksford, M., & Chater, N. (1994). A rational analysis of the selection task as optimal data selection. *Psychological Review*, *101*(4), 608. https://doi.org/https://doi.org/10.1037/0033-295X.101.4.608

Parpart, P., Jones, M., & Love, B. C. (2018). Heuristics as bayesian inference under extreme priors. *Cognitive psychology*, *102*, 127–144. https://doi.org/https://doi.org/10.1016/j.cogpsych.2017.11.006

Pinker, S., & Ullman, M. T. (2002). The past and future of the past tense. *Trends in cognitive sciences*, *6*(11), 456–463.

Putnam, H. (1963). Degree of confirmation and inductive logic.

Quilty-Dunn, J., & Mandelbaum, E. (2018). Inferential transitions. *Australasian Journal of Philosophy*, *96*(3), 532–547. https://doi.org/https://doi.org/10.1080/00048402.2017.1358754

Ragni, M., Kola, I., & Johnson-Laird, P. N. (2017). The wason selection task: A meta-analysis. *Cognitive Science*. https://api.semanticscholar.org/CorpusID:264270718

Russell, B. (1919). I—on propositions: What they are and how they mean. *Aristotelian Society Supplementary Volume*, *2*(1), 1–43.

Russell, S. J., & Wefald, E. (1991). *Do the right thing: Studies in limited rationality*. MIT press.

Ryle, G. (1950). The concept of mind. *British Journal for the Philosophy of Science*, *1*(4), 328–332.

Sanborn, A. N., Griffiths, T. L., & Navarro, D. J. (2006). A more rational model of categorization. *Proceedings of the 28th Annual Conference of the Cognitive Science Society*, 1–6.

Sanborn, A. N., Griffiths, T. L., & Navarro, D. J. (2010). Rational approximations to rational models: Alternative algorithms for category learning. *Psychological Review*, *117*, 1144–1167. https://doi.org/10.1037/a0020511

Schelling, T. C. (1960). *The strategy of conflict*. London: Oxford University Press.

Schulte, O. (2023). Formal Learning Theory. In E. N. Zalta & U. Nodelman (Eds.), *The Stanford encyclopedia of philosophy* (Fall 2023). Metaphysics Research Lab, Stanford University.

Shiryaev, A. N. (2007). *Optimal stopping rules* (Vol. 8). Springer Science & Business Media.

Siegel, S. (2016). *The rationality of perception*. Oxford University Press. https://doi.org/https://doi.org/10.1093/acprof:oso/9780198797081.001.0001

Silverstein, M. (2017). Agency and normative self-governance. *Australasian Journal of Philosophy*, *95*(3), 517–528.

Simon, H. A. (1955). A behavioral model of rational choice. *The quarterly journal of economics*, 99–118. https://doi.org/https://www.jstor.org/stable/1884852

Simon, H. A. (1957). *Models of man*. Wiley.

Simon, H. A. (1990). Bounded rationality. *Utility and probability*, 15–18.

Skinner, B. F. (1977). Why i am not a cognitive psychologist. *Behaviorism*, *5*(2), 1–10.

Smith, G. E. (2007). Gaining access: Using seismology to probe the earth's insides.

Smith, G. E. (2014). Closing the loop. *Newton and empiricism*, 262–352.

Stalnaker, R. C. (1984). *Inquiry*. Cambridge University Press.

Stich, S. P. (1985). Could man be an irrational animal? some notes on the epistemology of rationality. *Synthese*, 115–135. https://doi.org/https://doi.org/10.1007/BF00485714

Stiennon, N., et al. (2020). Learning to summarize with human feedback. *Advances in Neural Information Processing Systems*, *33*, 3008–3021.

Strawson, P. F. (2008). *Freedom and resentment and other essays*. Routledge.

Swensson, R. G. (1972). The elusive tradeoff: Speed vs accuracy in visual discrimination tasks. *Perception & Psychophysics*, *12*(1), 16–32.

Thorstad, D. (2023). Why bounded rationality (in epistemology)? *Philosophy and Phenomenological Research*. https://doi.org/https://doi.org/10.1111/phpr.12978

Tiippana, K. (2014). What is the mcgurk effect? *Frontiers in psychology*, *5*, 91962.

Todorović, D. (2018). In defence of illusions: A reply to braddick (2018).

Townsend, J. T. (2008). Mathematical psychology: Prospects for the 21st century: A guest editorial. *Journal of mathematical psychology*, *52*(5), 269–280.

Tsai, R.-C., & Böckenholt, U. (2006). Modelling intransitive preferences: A random-effects approach. *Journal of Mathematical Psychology*, *50*(1), 1–14.

Wald, A. (1947). Foundations of a general theory of sequential decision functions. *Econometrica, Journal of the Econometric Society*, 279–313.

Wason, P. C. (1968). Reasoning about a rule. *Quarterly journal of experimental psychology*, *20*(3), 273–281. https://doi.org/https://doi.org/10.1080/14640746808400161

Wedgwood, R. (2006). The normative force of reasoning. *Noûs*, *40*(4), 660–686. Retrieved February 15, 2024, from http://www.jstor.org/stable/4093983

Wedgwood, R. (2009). The normativity of the intentional.

Wedgwood, R. (2017). *The value of rationality*. Oxford University Press.

Williams, R. (2013). A non-pragmatic dominance argument for conditionalization. https://philarchive.org/rec/WILAND-2

Wolpert, D. H., & Macready, W. G. (1997). No free lunch theorems for optimization. *IEEE transactions on evolutionary computation*, *1*(1), 67–82.

Woodward, J. (2005). *Making things happen: A theory of causal explanation*. Oxford university press.