UNIVERSITY OF CALIFORNIA
Los Angeles

Confidence and Objective Probability Signaling in Perceptual Systems

A dissertation submitted in partial satisfaction
of the requirements for the degree
Doctor of Philosophy in Philosophy

by

William A Kowalsky

2020

ABSTRACT OF THE DISSERTATION

Confidence and Objective Probability Signaling
in Perceptual Systems

by

William A Kowalsky
Doctor of Philosophy in Philosophy
University of California, Los Angeles, 2020
Professor C. Tyler Burge, Chair

It's familiar to us from our thinking and deliberating that we can more or less confident about the truth of a proposition. For instance, you might be more confident that it will be overcast tomorrow than you are that it will rain. States of confidence are most familiar to us at the level of our thought. Are there states of confidence below the level of thought? Are states of confidence assigned by our psychological sub-systems? Recent discussions in the philosophy of perception and philosophy of psychology raise the specific question of whether states of confidence are assigned in perceptual systems. A perceptual system can represent an object as red, but can it also assign a level of confidence to the object's being red?

This dissertation argues that perceptual systems do assign levels of confidence. To address the question, we must have some sense for what sort of state a level of confidence. We must explicate the notion of a level of confidence, in order to identify central psychological signatures and roles that such a state plays. Such signatures clarify what it would take for a perceptual state to count as a level of confidence. Having explicated the notion, we must then turn to empirical science. We must look carefully at perceptual psychology and allied fields to see whether perceptual systems in fact have states that satisfy the signatures. I follow this methodology.

In explicating levels of confidence, I draw on both commonsense understanding and normative disciplines that invoke levels of confidence, such as decision theory and formal epistemology. I identify two core signatures of levels of confidence. One signature is roughly that increasing levels of confidence in p tend to lead to an increased *reliance*, by subsequent psychological processes, on the truth of p. The other signature is that increasing levels of confidence in p tend to be formed on the basis of information that better *supports* the truth of p.

I also argue, on empirical grounds, that perceptual systems have sensory capacities for signaling objective probabilities. Most saliently, they have capacities for tracking the objective probability that the perceptual representations they produce are veridical. A perceptual representation as of *red* may be objectively more likely in one circumstance than in another to be veridical. Perceptual systems have capacities for signaling this probabilistic difference specifically. I argue for such states by considering experimental results in multimodal cue integration. I embed my account in broader accounts of objective probability and sensory signaling.

I then argue that the probability signaling states constitute levels of confidence in perceptual systems. The probability signaling states guide the use of perceptual representations in a way that satisfies the first signature. Because they track different probabilities of a representation's veridicality, they satisfy the second signature. I close by considering what other sorts of sensory phenomena might be explained in terms of confidence, by examining capacities for "approximate number representation."

The dissertation of William A Kowalsky is approved.

Gabriel Jae Greenberg

Joshua David Armstrong

Chris Hak Wan Lau

C. Tyler Burge, Committee Chair

University of California, Los Angeles

2020

To my mother and father,

with love and gratitude.

TABLE OF CONTENTS

Introduction **1**

1 Objective Probability **14**

1.1 A model of objective probability 15

1.2 The explanatory roles of objective probability in the special sciences 22

1.3 Some important probability concepts 29

1.3.1 Random variables and probability distributions 29

1.3.2 Instantial probabilities properties 35

1.3.3 Instantial distributional properties 41

2 Probability signaling in sensory systems **44**

2.1 The experimental set-up of Ernst and Banks (2002) 45

2.2 Probabilistic properties of the experiment 53

2.2.1 Production distributions 53

2.2.2 Posterior distributions given representational states 58

2.2.3 Probabilities of veridicality and natural analogues 61

2.3 The central findings of Ernst and Banks (2002) 64

2.3.1 The two central findings 64

2.3.2 Cue integration in natural analogues 68

2.4 A model of the cue integration process 71

2.4.1 How the findings constrain the cue integration process 71

2.4.2 A basic model of cue integration 79

2.4.3 The causal properties of noise-tracking states 80

2.4.4 Noise-tracking states as kinds of modes of presentation 85

2.4.5 Why aren't modes of presentation associated with more fine-grained conditional distributions? 90

2.4.6 Summary 92

2.5 Sensory signaling states 93

2.5.1 Sensory signaling as a natural kind of functional encoding 96

2.6 IPP Signaling States 103

2.6.1 The function of noise-tracking states 103

2.6.2 Candidate properties for signaling 108

2.6.3 Likelihood functions and reliability norms 113

2.7 Why probability signaling states are perceptual 119

3 Levels of confidence 123

3.1 Identifying the explicandum 124

3.2 Reliance 128

3.2.1 Confidence and reliance in the philosophical literature 128

3.2.2 Decision-theoretic reliance 134

3.2.3 Generalizing beyond the idealizations of decision theory 143

3.3 Support 154

3.3.1 Conditional support 155

3.3.2 Levels of Support and Levels of Confidence 162

3.4 Perceptual levels of confidence 170

4 Analogue magnitude representations and objective probability 175

4.1 The Behavioral Signatures of Numerosity Discrimination 176

4.2 The Log-Gaussian Model of the ANS 181

4.3 Psychological, semantic, and methodological preliminaries 186

4.4 Against numerosity semantics 192

4.5 Against partial numerosity 204

4.6 Ratio semantics 206

4.7 Comparative semantics 209

4.8 Conclusion 211

Bibliography 212

LIST OF FIGURES

1.1 An example of a probability distribution. 32

2.1 Example of a visual stimulus in the no-noise condition of Ernst and Banks (2002) 47

2.2 Example of a visual stimulus in the high-noise condition of Ernst and Banks (2002) 48

2.3 Example production distributions. 55

2.4 Posterior distribution given a perceptual state. 60

2.5 Graphical illustration of the central findings of Ernst and Banks (2002). 67

2.6 Unimodal and multimodal production distributions for three types of trial . . . 76

2.7 Multimodal compromise as a function of haptic precision, given visual precision 0.11. 83

2.8 Multimodal compromise as a function of haptic precision, given visual precision 1. 83

3.1 Payoff matrix for the Detective Case. 136

3.2 A general payoff matrix. 137

4.1 Same-different stimuli. 177

4.2 Behavioral results in numerosity same-different task. 178

4.3 Alternative stimuli from Nieder et al. (2002) 179

4.4 The Log Gaussian Model 184

4.5 Proposed mapping from ANS states to contents. 196

4.6 Results of simulation comparing discrete and analog models of numerosity discrimination. 198

4.7 Predictions for free-tapping experiments. 200

4.8 Model predictions for response counting experiments. 201

4.9 Illustration of content mapping under a ratio semantics. 209

ACKNOWLEDGMENTS

I would like to begin by thanking my advisor, Tyler Burge. The way Tyler does philosophy—in his writings, in seminars, in conversation—has been a constant inspiration to me. I aspire to his seriousness, patience, and rigor in philosophy. His work and his approach have helped shape nearly every part of my philosophical outlook. I am deeply grateful to have had Tyler as a teacher and advisor. It would be hard for me to overstate how much my intellectual development owes to his tutelage. I thank him for his guidance and encouragement.

I have learned many things from Gabe Greenberg. From his example, I learned the value of elegant and judicious use of formalism in philosophy, and I learned the value of reading widely and in places where people tend not to look. I have been trying emulate Gabe's incisiveness since day one of graduate school. But above all, I am grateful for his sheer enthusiasm in doing philosophy. If your current project feels like a deadened husk to you, puzzle through it with Gabe and watch it come alive again.

Josh Armstrong has long been a penetrating interlocutor, and I thank him for his insight and scrutiny. I aspire to produce philosophy that engages deeply and knowledgeably with empirical science, and I am grateful to Josh for providing such a splendid model for how to do that. I thank him also for the many suggestions he has made about my research, all of which turned out to be extraordinarily fruitful.

I thank Hakwan Lau for letting me hang around his lab during the 2015-2016 academic year. The discussions that took place in his journal club and lab meetings, and in particular Hakwan's commentary, helped me understand the way practicing psychologists think about things. I also thank Michael Rescorla for helpful conversations about the topics of this dissertation, and for his encouragement.

My philosophical perspective has been shaped by the faculty of the UCLA Philosophy Department in deep and myriad ways. I want to thank especially Sam Cumming, David Kaplan, Sherri Roush, and Sheldon Smith for philosophical conversation and for their general

good cheer. Seana Shiffrin has been a mentor to me in several ways, for which I am deeply grateful. Finally, I thank Gavin Lawrence for his reliability and steadfast support.

I cannot imagine a graduate student body more exhilarating or stimulating than the one I have found at UCLA. For their philosophical conversation, their in-the-hall badinage, and their friendship, I would like to thank Tristen Cardwell, Jason Carnell, Milo Crimi, Jenna Donohue, Andrew Flynn, Antti Hiltunen, Amber Kavka-Warren, Tanya Kostochka, Esther Nikbin, Torsten Odland, Alexi Patsaouras, David Pederson, Sam Pensler, Jacob Reid, Melissa Retkwa, Megan Robb, Ayana Samuel, Vaheh Shirvanian, Eric Tracy, Jordan Wallace-Wolf, and Aaron West. For their steadfast camaraderie, and for generally making my past eight years so enjoyable, I thank Ian Boon, Christian De Leon, Kim Johnston, and John Kardosh. You all know why. And I especially want to thank my Mind and Language buddies. Gabe Dupre is basically solely responsible for my entire outlook in the philosophy of science, and I have learned more about language from him than from anyone else—not least from his uncanny ability for a *bon mot*. Gabby Johnson is one of the clearest thinkers I know, and in conversation, collaboration, and correspondence with her, she has made me a clearer thinker too. She brings me back to good philosophy. Kevin Lande has been, to me, a paragon of every virtue I would like to have as a philosopher of psychology. He is a sensitive and insightful reader of psychology, and he has a knack for pulling deep and novel ideas from it. My friendship with each of you has brightened my life.

For helpful conversation, I would like to thank Jake Beck, Denis Buehler, Alexi Burgess, Katie Elliott, Genna Erlikhman, Steven Gross, Casey O'Callaghan, John Morrison, Brian Odegaard, Susanna Siegel, and Jona Vance.

For their tireless work in keeping our department humming, I would like to thank: Michelle Anderson, Brittany Bolden, Laura Clennon, Julian Fischer, Tanya Kim, Rachel Lee, Blake Livesay, Doug Meyers, and Jackie Perez.

Thanks to my old cronies for continued friendship, broader intellectual sustenance, and general good times. For this, I thank Tian Chen, Miguel Cutiongco, Valentin Fernandez, Andrea Friedman, Ruthe Foushee, Susie Kim, Spencer Mason, Alex Mogyoros, Suhas Rao, and David Zuluaga. And I extend my most heartfelt gratitude to Julian Arni, Siddarth Chandrasekaran, and Katie Dahlinghaus, for more than I can say.

Finally, I want to thank my family. I thank my aunt Joanne for her constant support. And I thank my mother and father. I cannot convey my gratitude for your love, support, and guidance.

VITA

2016	C. Phil in Philosophy, University of California, Los Angeles.
2012	A.B. in Philosophy, Harvard University.

INTRODUCTION

A familiar aspect of our cognitive lives is that we can be more or less confident about the truth of a proposition. You might be more confident that it will be overcast tomorrow than that it will rain. You might be less confident now than you were six months ago that the Democrats will win a Senate majority. You might be very confident that the theory of general relativity is approximately true. Our levels of confidence are often reflected in action. We may hesitate or hedge. Our levels of confidence are also often reflected in thought. We may form a contingency plan or withhold judgement. Often one's confidence in a proposition changes as one gathers new evidence bearing on its truth. The confidence one forms is subject to appraisal. I may be charged with being overly confident, or not confident enough. Like beliefs and desires, levels of confidence are familiar elements of our thought.

Levels of confidence are kinds of psychological states familiar to us at the level of beliefs and desires. Are there levels of confidence in perceptual systems? Consider, for example, seeing a cube through fogged glasses. You see it as a cube, but it is quite blurry. As a result of the blur, you might, in cognition, come to have only a moderate level of confidence that the object is a cube. Does your perceptual system attach a separate level of confidence to the object's being a cube—a state of confidence formed independently of the higher-level confidence you form? What conditions would have to be met for a state of the perceptual system to count as a level of confidence? Is there empirical evidence that such perceptual states exist? These questions are the topic of this dissertation. The questions are raised, for instance, by recent disputes about the interpretation of Bayesian models in perceptual psychology and about accounting for aspects of "imprecise" perceptual phenomenology.

Let me sharpen the question slightly. Some psychological states are partly type-identified semantically. Such states are about some subject matter. Psychological states that are type-similar semantically may differ along a further dimension. One could believe that it is raining, but one could also desire or suppose it. These states share a subject matter, but differ in their *orientation* towards that subject matter. For one, the success conditions of the states depend on the subject matter in different ways. If it is not raining, the belief fails as a belief, but the desire does not fail as a desire. That beliefs fail but desires do not when their semantic content is not true reflects a difference in orientation. Moreover, the causal roles of the three states differ. The belief may lead, by practical inference, to an intention to grab the umbrella; the supposition typically will not. As they appear in thought, levels of confidence appear to be type-identified in a similar way. A level of confidence is partly type-identified by what it represents, and partly type-identified by an orientation towards what it represents.

Similarly, some states of perceptual systems are representational, and may be classified by their orientations towards what they represent. Consider a visual percept, produced as the final result of visual processing, that represents a red cube. The state has representational content. Furthermore, the orientation of the state towards its representational content is, like belief, committal: the state undergoes a certain type of failure as a percept if the object is not a cube, or is not red. Our question is whether there are perceptual states that are oriented towards their representational contents in the ways that levels of confidence are oriented towards contents in thought. To the extent that a perceptual state is so oriented, it is confidence-like. If the similarities in orientation are fundamental enough, the state may be an instance of the kind—a genuine level of confidence.

Answering our question requires some understanding of what kind of psychological state a level of confidence is. Which functions and causal roles are typical of a level of confidence? Which are constitutive? Articulating signatures and constitutive conditions on levels of confidence is an *explicatory* project. Explicating the psychological kind will draw on and refine commonsense understanding of confidence. Explication will also draw on the theoretical understanding of confidence, as confidence is assumed and idealized by normative disciplines such as decision theory and formal epistemology. Given an explication of confidence, answering our question then requires *empirical* investigation. Once we have signatures or conditions on a state's being a level of confidence, we must assay empirical evidence to see whether there are perceptual states meeting the conditions or bearing the signatures. Perceptual psychology and allied fields provide well-confirmed and mathematically precise models of perceptual capacities. Any claim that perceptual systems assign levels of confidence must be grounded in the empirical facts about such systems, and must be consistent with the best explanations of such systems offered by the relevant sciences. We will consider results and explanations in perceptual psychology in some detail.

Chapter 3 pursues the explicatory project. I reflect on commonsense and theoretical understanding of confidence as a psychological state at the level of thought. I reflect on the roles that states of confidence play in commonsense psychological explanations and in the normative claims of decision theory and formal epistemology. I argue that different states of confidence in a given proposition p are, at a minimum, linearly ordered. The term "level" in "level of confidence" thus denotes, at least, a relative position in the order. I leave open whether different confidence states may have richer structural relations. I further leave open relations between confidence and belief. I leave open, for example, the relation between a level of confidence in p and a belief that p. I then argue for two signatures of levels of confidence. The first signature is that greater levels of confidence in p typically lead to a greater *reliance* on the truth of p by psychological processes that utilize the confidence. I define a general and abstract metric of the degree to which an output of a psychological process relies on the truth of p. Roughly, an output's reliance on p is determined by how much the output depends on p's being true in order to reap some benefit and to avoid some loss that would accrue if p were false. The metric of reliance applies to *practical* processes geared towards action and planning, as well as *theoretical* processes geared towards true belief and understanding. The notions of "benefit" and "loss" are defined abstractly in terms of the *functions* that various psychological processes and states have or contribute to. Such functions may be biological, practical, representational, or epistemic. My account of reliance may be seen as a substantial generalization of traditional decision-theoretic glosses of confidence as a betting dispositions.

The second signature is that greater levels of confidence in p are typically formed from bodies of information that support p's truth to a greater degree. I sketch a general notion of *support*, exemplified in familiar relations of evidential support. I focus on support relations between a body of information and a proposition that obtain when the body of information deductively entails the proposition or confers a certain objective probability on the proposition's truth. I point out that our levels of confidence in p often change as we gain further information that bears on p. Moreover, as our information comes to support p to a greater extent, our confidence in p often rises. I consider various kinds of failure that a level of confidence in p is subject to when it fails to match the strength of support afforded p by a body of information. I give reasons for thinking that matching a level of support is a *norm* on a level of confidence.

The empirical component of the project occupies Chapters 1 and 2. In Chapter 2, I argue that, on empirical grounds, human perceptual systems have capacities for tracking and signaling *objective probability* properties. In different kinds of situations, different objective probability relations obtain between distal properties—such as the possible colors of a surface—and sensory states—such as the possible representations of color. Given a particular representation of a specific color, a distal color may be objectively more likely in one kind of situation than it is in another. Given that the system has represented a surface as red, the surface may be more likely to *be* red under white light than red light. I argue that perceptual systems have states that are differentially produced in such situations. One kind of state is produced when one kind of probabilistic relation holds, and a different kind is produced when a different probabilistic relation holds. Further, such state have systematic psychological uses that are explained by the probabilistic properties of the situations that they are preferentially produced in. Thus, I argue, such states bear functional correlations with the probabilistic magnitudes. I call such states *probability signaling states.* I argue for the existence of states on the basis of experimental results in multisensory cue integration. I argue that such states are members of a broader class of sensory signaling states. Sensory signaling states include perceptual representations, but they also include sub-representational sensory states. States for signaling the orientation of an edge in a retinal image or for signaling a direction and distance of the shortest path to an ant's nest are examples of sub-representational sensory states.

Giving a precise account of probability signaling states requires having some understanding of what objective probabilities are. In Chapter 1, I articulate my understanding of objective probability. Objective probabilities are, in the first instance, relations between property types, relative to a background type of set-up. The outcome type *coming up heads* has objective probability 0.5 relative to a background set-up type that may specify, for instance, physical properties of the coin and characteristics of the flipping mechanism. Relative to a different set-up type, *coming up heads* may have a different objective probability. The objective probability of an event relative to a set-up has modal entailments about the relative frequency of that event within instances of the set-up. I briefly address common sources of skepticism about objective probability. I defend realism about objective probabilities by discussing the ways in which they are presupposed and utilized in explanations in the special sciences. I explain how my account avoids a problem—Humphrey's Paradox—facing many similar accounts, and I discuss the sources of probabilistic variability. I also articulate a notion of "single case" or *instantial* probabilities. Instantial probabilities are properties of token events and situations. Instantial probabilities enable us to count, for instance, not only how many actual flips of various coins came up heads, but also how many came up heads *with* a certain probability.

The account of objective probability enables us to pose and answer various fine-grained questions about probability signaling states. The account of instantial probabilities gives precision to the claim that a state "correlates with" a probabilistic property. The account naturally raises the question of the set-up types that individuate the probabilistic magnitudes signaled. Furthermore, we can imagine the sensory system placed in a novel environment, in which radically different probability relations obtain. In such an environment, probability signaling states may fail—they signal the presence of a probabilistic property that is not present. Such failure must be relative to some set-up type instantiated by the novel environment. What sorts of factors determine *which* set-up type determines the success or failure of a probability signaling state?

Our background model of objective probability also enables us to precisely state and evaluate different proposals about various *norms* on perceptual systems. Perceptual representational states function, in one way, to be veridical. The norms I consider are norms to be *reliably* veridical. Such norms must be relative to a set-up, and I discuss the factors that fix the set-up relative to which such norms are individuated. I thereby give an account of a perception-level analogue to what has traditionally been called the "reference class problem." Articulating such norms also helps us sharpen our account of probability signaling states. In the scenarios I discuss, each signaling state is differentially sensitive to multiple probabilistic properties that are metaphysically distinct but mathematically the same in structure. Which of these is the property signaled by the state? I propose that it is the property the tracking of which best explains how the state contributes to fulfillment of reliability norms on the perceptual system. I discuss these matters in Chapter 2.

At the end of Chapter 3, I argue that probability signaling states bear the signatures of confidence identified in Chapter 3. The states discussed in Chapter 2 signal probabilistic quantities systematically related to the probability that a perceptual representation is *veridical*. Greater probability signaling states indicate a greater probability that an attendant perceptual representation is approximately veridical. It thus satisfies the signature of support. The signaling state also determines the relative influence that a perceptual representation has on downstream processes. For instance, a visual probability signaling state partly determines the influence of an attendant visual representation of width on a downstream multimodal representation of width. Greater visual signaling states bring the multimodal representation closer to the visual representation. In this way, the signaling states satisfy the signature of reliance.

The question of whether perceptual systems assign confidence is raised by several recent discussions. Most saliently, it is raised by debates about the interpretation of Bayesian models in perceptual psychology. Literal construals of such models appear to require that perceptual systems have confidence-like states, but it is hotly contested to what extent Bayesian models should be literally construed, if at all.

We can introduce the notion of a Bayesian model by way of a metaphor. We imagine a hypothetical agent, which we call an *ideal observer*. We assume the ideal observer has a sensory system and must execute some task. For example, suppose that there is a slanted surface in front of the observer. The surface may have one of many different possible slants. The slant of the surface, together with other properties of the environment, causes proximal stimulation of the observer's sensory receptors. From the proximal stimulation, the observer's sensory system produces a representation of the surface's slant that is as accurate as possible.

The ideal observer's sensory system is stipulated to contain several types of states. Most centrally, the sensory system assigns various *subjective probabilities* to various possibilities. For instance, the system assigns a subjective prior probability to each particular slant S, intuitively understood as reflecting the system's anticipation that the slant will be S before it has received any stimulation. The system also assigns subjective conditional probabilities. It assigns, for instance, a subjective probability to a proximal stimulation of type P under the supposition that the slant is S. In the ideal observer, a subjective probability is a perceptual state that is partly type-identified by a representational content, and partly type-identified by a magnitude that is measured by some real number in the unit interval $[0, 1]$.

Suppose that the ideal observer receives proximal stimulation P on a trial. The sensory system produces its representation of slant by first assigning a subjective posterior probability to each slant S given that the proximal stimulation is P. Let the system's subjective prior probability in property X be $c(X)$ and let its subjective conditional probability in X given Y be $c(X|Y)$. On a trial in which the proximal stimulation is P, the system assigns a posterior probability $P(S|P)$ to each slant S. Each such posterior probability is computed by the system from its other subjective probabilities, according to Bayes' Theorem:

$$c(S|P) = \frac{c(P|S)c(S)}{c(P)}$$

The final percept of slant is then determined from the posterior probabilities so assigned to different slants. Commonly, the final percept will be as of that slant assigned greatest subjective posterior probability.

Ideal observers provide models of actual perceptual systems. The performance of a perceptual system may match that of the ideal observer when both are set the same kind of task in the same sorts of conditions. The ideal observer thus permits compact articulation of the input-output laws relating actual sensory stimulations and actual perceptual representations. Since ideal observers are all described in the same Bayesian framework, they permit comparison and unified explanations of different sensory processes. When the subjective probabilities of an ideal observer match environmental objective probabilities, they help explain why perceptual systems follow the particular input-output laws that they do. The Bayesian framework funds an analysis of input-output laws as statistically optimal or sub-optimal. Quite often, actual sensory systems are found to perform optimally, given the probabilities of their natural habitats. In such cases, the Bayesian analysis enables us to see the input-output laws as the consequence of factors (evolutionary, developmental, etc.) that drive sensory systems towards optimality within their niche.

For these and other reasons, Bayesian models have been enormously fruitful in perceptual science. The success of such models naturally raises a question about their *truth*. To what extent are Bayesian models approximately true descriptions of perceptual processes? To what extent do perceptual systems in fact have capacities similar to those possessed by the ideal observer? At one extreme, one may be a *radical instrumentalist* about Bayesian models. Such a view holds that, despite their success, Bayesian models are not even approximately true descriptions of perceptual systems. Perceptual systems have nothing like a capacity for attaching subjective probabilities to possibilities, and their transitions are in no way like those postulated by a Bayesian model. At the other extreme, one may be a *radical realist*. Such a view holds that an empirically adequate Bayesian model is an entirely and literally true description of part of a perceptual system. Perceptual systems really do assign subjective probabilities—conditional and unconditional—to various possibilities, and those subjective probabilities interact in the same ways they do in the ideal observer. Between these two extremes, moderate instrumentalist and moderate realist positions are possible.

In most research traditions labeled "Bayesian," subjective probability is a posited psychological state understood as an idealization of a level of confidence or uncertainty. Founding texts of Bayesian decision theory give intuitive glosses on subjective probability as a level of confidence, uncertainty, or degree of belief. For instance, the decision theory of Savage (1954) assumes that subjective probability "measures the confidence that a particular individual has in the truth of a proposition" (p. 3). A popular textbook in the Bayesian approach to statistical inference tells us that "the essential characteristic of Bayesian methods is their explicit use of probability for quantifying uncertainty in inferences based on statistical data analysis" (Gelman et al. 2004, p. 3). Such glosses are commonplace in perceptual science as well. For a representative example, Ma (2019) says that the subjective probabilities about a slant S cited in our Bayesian model "should be interpreted as a degree of belief in a value of S" (p. 165).

Thus, radical realists must show at least two things. First, they must show that perceptual systems have states that are structurally like subjective probabilities. There must be perceptual state-types with rich enough structure to warrant ascription of real-valued magnitudes between 0 and 1. The state-types must form a collective so that their magnitudes sum to 1, and generally satisfy the other axioms of probability. Second, the radical realist must show that these perceptual states are levels of confidence—as opposed to some other kind of degreed state. The second claim is required because states' having a structure formally describable with probabilities does not entail that the states are levels of confidence. Block (2018) considers a model in which psychological magnitudes are postulated and labeled degrees of "cortical uncertainty." Block argues that "what [the model-builders] call 'cortical uncertainty' is equally well described as 'degree of cortical competition'" (p. 6). Block argues for this as part of a more general defense of instrumentalism about Bayesian models. A realist about the model Block discusses must explain why the postulated psychological magnitudes are levels of confidence rather than (as Block suggests) levels of competition.

Of course, one may be a realist about Bayesian models without being a radical realist. Rescorla (2019) is one example. Rescorla defends a realist attitude about Bayesian models, but expresses skepticism that perceptual systems assign the full suite of subjective probabilities that are ascribed to the ideal observer. He suggests that realism about Bayesian models requires only that the perceptual system have "credal states." States corresponding to individual subjective probabilities are credal states, but they are not the only kinds. He suggests that states encoding the parameters of a probability distribution could count as credal states. He also suggests that a disposition of the system to produce certain kinds of states with certain frequencies (as in a "sampling" account) could count as a credal state. Even so, credal states, by way of their connection with subjective probability, are understood as being close in kind to levels of confidence.

The argument of this dissertation contributes to the dispute over Bayesian models in two ways. First, the explicatory project provides a minimal account of what makes a state "credal." It thereby offers clarification about a necessary condition for realism about Bayesian models. Second, it argues that perceptual systems have such credal states, and that those states take part in transitions that are explicable as implementations or realizations of the rules specified in the Bayesian analysis. Thus, the argument of the dissertation may be read as defending a moderate realism about Bayesian models.

The question of whether perceptual systems assign confidence is also raised by recent debates about conscious perceptual *experience*. Morrison (2016) argues that phenomenally conscious perceptual experience sometimes involves the assignment of levels of confidence. He focuses on perceptual experiences that are indistinct or imprecise, such as one I might have when I cannot quite tell whether the tablecloth in the candlelit room is crimson or scarlet. He argues that, in such a case, my visual experience involves representational contents for the two mutually incompatible colors. Furthermore, the experience assigns weights to each content. These weights, Morrison argues, are levels of confidence in the contents. Morrison argues for such experiential confidences abductively, claiming that they would best explain the cognitive level of confidence I would form in thought if I "completely trusted" the experience. Similarly, Munton (2016) argues that levels of perceptual experiential confidence are needed to explain the different degrees of justification that some experiences provide for beliefs. I am sympathetic to the position that both authors take. The arguments of this dissertation help buttress and refine the position.

This dispute centers on perceptual *experience*, which is understood to be a psychological state that is necessarily phenomenally conscious. My discussion of perceptual confidence does not focus on experience. It focuses on the capacities of perceptual systems, leaving aside the question of the relation between those capacities and conscious perceptual experience. In principle, perceptual experience could assign levels of confidence to representational contents even if perceptual *systems* did not. In principle, the source of the experiential confidence might be post-perceptual but pre-cognitive. For instance, Carey (2009) introduces a notion of a *core cognition* module. It is possible that some such modules are sufficiently domain-general to count as post-perceptual, yet do not involve *propositional* representations—hence, as I use the term, pre-cognitive. A core social cognition module, for instance, may be post-perceptual, and yet exclusively utilize non-propositional tree-like representations (see Camp 2009). Perhaps a similar module is the source of levels of confidence assigned in perceptual experience.

Prima facie, however, it is quite plausible that, if there are perceptual experiential confidences, they arise from the perceptual system. Inasmuch as experiential confidence reflects aspects of experience like blur or fuzziness, the confidence pertains to matters that fall within the domain of perceptual systems. The likelihood that there are experiential levels of confidence would be severely diminished if perceptual systems never assigned levels of confidence. My argument for levels of confidence in perceptual systems blocks this potential problem for experiential confidences. Perceptual systems have the kinds of capacities that are plausibly required for perceptual experience to assign confidences.

My account of levels of confidence in perceptual systems is salutary for experiential confidences in a further way. Morrison (2016) explicitly assumes that, if an experience assigns a level of confidence to a representational content, it must also assign a level of confidence to a distinct and mutually incompatible representational content. As Morrison says, "you can perceive multiple possibilities without simultaneously perceiving each possibility as actual" (p. 24). Such a view has been seen as problematic. Bayne (2010) argues against the existence of perceptual experiences that have, at one time, incompatible representational contents. And Siegel (forthcoming) argues that, even if perceptual experience can represent multiple possibilities simultaneously, it cannot assign probabilistically structured weights to those possibilities.

The assumption that a single experience can represent incompatible possibilities is not, however, necessary for perceptual experience to assign a level of confidence to a representational content. I argue that, when cues about an environmental property are available to multiple modalities, each modality produces a representation of the property and a unimodal level of confidence in that representation. In the model I focus on, a unimodal level of confidence in a content does not entail distinct levels of confidence in incompatible contents. On that model, a visual system may produce a representation as of 45° slant, and assign a level of confidence to it; it does not also attach a separate level of confidence to a representation as of 46°. Such levels of confidence may provide the source for experiential confidence, without entailing that experiences present incompatible possibilities.

The literature on experiential confidence suggests a link between confidence and perceptual states that are intuitively indistinct or imprecise. Chapter 4 explores this link further. I focus on sensory capacities for discriminating aggregates of items on the basis of their numerosities—roughly, the cardinalities of the aggregates. Psychologists postulate an *approximate number system* that underlies numerosity discrimination. The system comprises a quasi-continuum of states that bear different profiles of differential sensitivites to various numerosities. The states are produced stochastically. I argue that no state of the approximate number system actually represents cardinalities or ranges of cardinalities. Rather, I argue, the states of the system represent only ordinal relations between aggregates. The system may represent one aggregate as being more numerous than another, without representing either as having a specific cardinality. Perception of aggregate quantity in such cases is, intuitively, indistinct. But here, the indistinctness does not reflect levels of confidence in representations of numerosity. Rather, it resides in the systems' representing a determinable-level property without representing any of its determinates.

CHAPTER 1

Objective Probability

Historically, there have been three main options for interpreting the formal apparatus of probability theory. A *subjective* interpretation sees probabilities as psychological kinds. Probabilities might *be* a mental state, or they may be idealizations or measurements of aspects of mental states. A *logical* or *evidential* interpretation sees probabilities as abstract relations between propositions or sets of propositions. An *objective* interpretation sees probability as a property of actual systems, though not just psychological ones. Objective probabilities may be properties of physical or biological systems.

Subsequent chapters of this dissertation extensively utilize a notion of objective probability. This first chapter describes that notion. In §1, I describe a model of objective probability. My model is guided by the use of objective probability notions in scientific explanation and prediction. The model raises a spate of questions that would be pressing if it were proposed as a full-dress metaphysics of objective probability. But I do not attempt such metaphysics in this dissertation. Some philosophers and scientists, however, treat objective probability with a certain level of suspicion. In §2, I attempt to allay some of that suspicion, by describing the explanatory roles that objective probability plays in the special sciences. I do this by reflecting on an explanation from evolutionary biology. In §3, I develop several probability concepts that I will use throughout the dissertation, including concepts of *random variable*, *distribution*, and others. The model I give in §1 sees objective probability as a relation between repeatable *types*. However, we sometimes ascribe probabilities to token events and individuals. Such "single-case" probabilities will also be needed in later chapters. In §4, I give a model of such single-case probabilities, which I call instantial probability properties.

1.1 A model of objective probability

Intuitively, a standard quarter flipped in a standard way has a 50% probability of landing heads up. Suppose one bends the coin so that it is convex, with heads facing "outwards." Flipping the bent coin in the standard way will reveal a bias towards landing heads. After repeated flipping, one may conclude that the coin has a 60% probability of coming up heads. Now suppose one flips a standard coin with the mechanism built by Diaconis et al. (2007). If the machine's parameters are set correctly, the coin will land heads every time. In that case, the coin has a 100% probability of landing heads.

The probabilities at issue are properties of a physical system—the coin, together with the flipping apparatus. An assertion of "the coin has n% probability of landing heads" can, in some contexts, be taken as an assertion about one's state of uncertainty about the outcome. But often it is most naturally taken as a commentary about the coin and how it is flipped. To mark that the probabilities are understood as properties of systems, I follow tradition and call them *objective probabilities.*

In each of the three cases, an objective probability attaches to the event-type *coin landing heads up.* The objective probability that attaches to the event-type differs in the three cases. The difference in probability is due to differences in the type of process that leads to the coin facing heads or tails. Evidently, the objective probability of an outcome depends on the type of process that might produce the outcome.

These examples suggest that objective probabilities are relations between event-types and process-types.[1] The first example suggests an objective probability relation of 50%

[1] I will assume that there are infinitely many objective probability relations, one for each real number in the unit interval $[0, 1]$. Moreover, I will assume that the relations themselves have whatever properties are needed to ground arithmetical statements involving them. For instance, if one event has probability 0.1 and another has probability 0.2, I assume that probability relations have, metaphysically, whatever properties

that holds between the event-type *coin lands heads up* and some type that classifies the flipping process—perhaps a type that specifies physical symmetries in the coin and flipping mechanism. The second example suggests an objective probability relation of 60% that holds between the event-type and a different characterization of a flipping process—one that specifies a certain degree of bend in the coin to be flipped. One may think of a set-up type as a property corresponding to a description of a kind of process. Following Hacking (1965) and others, I will call the background process-type relative to which objective probabilities hold a **set-up type**, or sometimes simply a set-up or "chance set-up." If the objective probability of event-type A is s relative to set-up E, I'll write $P_E(A) = s$. I'll often drop the sub-script when context makes the set-up obvious.

A set-up type will often entail that certain types of objects or processes be present in any instance of it. A set-up may entail that any instance of it contain some coin and some flipping mechanism. Or, a set-up type may entail that any instance of it contain a perceiver of a certain species and some surface located in front of the perceiver. Additionally, a set-up may entail that the objects have one of several possible properties in any instance of it. For example, a set-up may require that the perceiver's sensory system be in one of several states in any instance. Set-ups also specify a certain kind of causal structure. It might require, for instance, that the surface cause, via reflection of light, some internal sensory state in the perceiver.

Set-up types may themselves be partly individuated by objective probabilities. For instance, a coin may be flipped with many different initial velocities. Any token flip of the coin occurs with a specific velocity. A set-up type may specify that the initial velocities are produced with certain objective probabilities. Consider set-up type T. T specifies that a physically symmetric coin will be flipped by a mechanism in calm atmospheric conditions, and will land on a flat surface. Additionally, T specifies that the mechanism flips with an ini-

are required for it to be literally true that the probability of the latter event is twice as great as that of the former. Essentially, I am assuming that they are magnitudes, in the sense of Peacocke (2015).

tial velocity that is chosen uniformly at random from a fixed range of velocities V. Suppose we build a flipping mechanism whose initial velocity is chosen from V by a random number generator. Our flipping mechanism, together with its surrounding environment, has the structure specified by T. Token coin flips by the mechanism are instances of the process-type T.

Now consider a person who has trained themselves to be able to flip a fair coin so that it always lands heads.[2] Suppose the person flips a fair coin while intentionally and successfully exercising their ability to induce heads. This token flip is not an instance of the set-up type T. The initial velocity to the flip was supplied by a process that does *not* select initial velocities uniformly at random from V. The initial velocity was supplied, rather, by a process that selects, with overwhelming probability, initial velocities that heads.

In set-up type T, the range of velocities V is an **exogenous variable**. The specific velocity values in the range are **exogenous values**. Exogeneous values are event-types that potentially make a difference to the way a process unfolds, but whose causal antecedents are left unspecified by the set-up. Possible initial conditions for a process will generally be exogenous values. If a set-up specifies that an individual exogenous value is supplied with probability p, I'll call p an **exogenous probability**.[3]

A set-up type may specify individual probabilities for each exogenous value. Our example T did this. But a set-up type may only specify higher-level features of exogenous probabilities. For instance, consider a set-up B that specifies that any two adjacent velocity values v_i and v_j for flipping a coin are chosen with very similar objective probabilities. That is, the probability that v_i is chosen cannot be substantially greater or less than the probability that v_j is chosen. Suppose three people toss a coin in succession. None of them

[2] According to Landhuis (2004), Persi Diaconis trained himself to be able to flip a fair coin in this way.

[3] In allowing exogenous probabilities to help individuate set-up types, I draw inspiration from Strevens (2006)'s notion of a probabilistic network. Further inspiration comes from a scheme for set-up type individuation implicit in work on causal graphical models (see Spirtes et al., 2000; Pearl, 2000). The term "exogenous variable" derives from that literature.

imparts imparts velocities in V with equal probabilities. Each produces certain velocities with higher probabilities than other velocities. One tends to flip with lower velocities, another with medium velocities, and the third with higher velocities. But none of the three people is precise enough to flip adjacent velocities with drastically different probabilities. So each of the three token flips instantiates the set-up type B. Interestingly, B's higher-level constraint on velocity probabilities is sufficient for the probability of heads to be 50% in B.[4]

A single token process will instantiate multiple set-up types at once. Suppose a coin is flipped by our mechanism that chooses initial velocities uniformly from some range. Suppose that, on that occasion, it is flipped with initial velocity 0.1 m/s. The token flip instantiates our set-up T. It also instantiates a set-up type that specifies that the coin is flipped with initial velocity 0.1 m/s. And it instantiates a set-up type that is like T, except that the set-up specifies that velocity is chosen uniformly and *specifically* by a particular random number generation algorithm. The token instantiates all three set-up types. Not all need be of equal explanatory significance.

Objective probabilities correspond to long-run relative frequencies. Suppose a system has, at the relevant level of description, the causal structure identified by set-up type E. The system may be "run" many times. That is, the system may undergo or take part in several sequences of events, each of which is an instance of the process-type characterized by E. By dint of its having the causal structure of E, the system is disposed to, over many "runs," produce an outcome A with a relative frequency that is approximately equal to the objective probability of A (relative to E). This claim deserves substantial clarification, but some such connection is presupposed in virtually all scientific uses of objective probability.

[4] Strictly speaking, my description of B is too sketchy to rigorously evaluate the probability of heads in B. I intend the example to be a simplified illustration of the phenomenon that an outcome can have a stable probability under a wide range of different initial condition distributions. For instance, see Strevens (2006, Ch. 2), building on classical work by Poincare and others. Probability of *red* on a balanced roulette wheel is 0.5, regardless of the croupier that spins it. Different croupiers impart initial velocities to the wheel with different probability distributions. But given the physical dynamics of the wheel, one can show that, if the distribution over velocities is sufficiently "smooth," the probability of *red* will be approximately 0.5. My condition in the text about "adjacent velocities" is intended to be suggestive of this smoothness constraint.

Observed frequencies provide evidence for claims about objective probabilities, and objective probabilities are parts of explanations of observed frequencies. Since I am not attempting a full-dress metaphysics of objective probability, I will not attempt to further explicate the connection between probability and frequency.[5]

It will often aid understanding of claims about objective probabilities to illustrate them with related claims about frequencies. I will assume that objective probability relations have a certain kind of modal entailment. Consider the class of all actual and possible tokens of a set-up type E. Suppose the objective probability of outcome A is s relative to E. Then, within the class of possible E-tokens, the (limiting) relative frequency of A-tokens is s. I'll also speak of that limiting relative frequency as a proportion. Such relative frequencies within modal space are expected to correspond to relative frequencies within sufficiently many actual repetitions of E. A full metaphysics of objective probability would demand further explication of this modal entailment, but our purposes are served without such explication.

The objective probabilities discussed so far have been *unconditional* probabilities. There are also *conditional* probabilities. Consider a set-up F, specified as follows. There are two coins in a bag. One is fair, but the other has a 75% chance of coming up heads. I reach into the bag and take one of the coins, with equal probabilities. I then flip the chosen coin, where it lands on the table. Relative to F, the event-type *the coin on the table faces heads up* has probability 62.5%. But *given* that the fair coin is selected, the flipped coin lands heads with probability 50%. In other words, the conditional objective probability (relative to F) that the coin on the table faces heads up *given* that the fair coin was chosen is 50%. To introduce notation, if the conditional probability of A given B is s relative to E, I will write

[5] Attempts to relate repetition, frequency, and objective probability typically focus on Bernoulli's Theorem. Applied to the current case, the theorem suggests the claim that, if the probability of A in E is s, then as the number of repetitions of E grows to infinity, the limiting relative frequency of A is "almost surely" s. The notion of "almost certainly" requires clarification. Mathematically, it refers to the fact that the set of infinite sequences of E with a different limiting relative frequency has measure 0. But that measure requires a physical interpretation. See Strevens (2011) for an attempt to ground such "almost certainly" clauses.

$P_E(A \mid B) = s$. For those unfamiliar with this notation, remember that the *given* information goes on the right of the vertical bar. Again, I will drop sub-scripts where context makes the set-up obvious.

Unconditional objective probabilities reflect tendencies of a type of process to result in a certain type of outcome. Conditional objective probabilities reflect the impact that a process' having unfurled one way has on its tendency to unfurl in some other manner. There are two general ways that an instance of F could unfurl: either the fair coin is selected, or the biased coin is selected. Subsequent developments in the process depend on which of these two ways the process went. The system has an equal tendency to go either way. The unconditional probability of heads reflects both the dependence of the outcome on either coin's having been selected, as well as the tendency of the system to choose either coin. But when the fair coin has been selected, the possibilities for the process' subsequent development have been restricted. Certain trajectories of the chosen coin through the air will be impossible, or perhaps will require enormously exogenously unlikely atmospheric conditions. The conditional probability summarizes the impact of this restriction on an outcome of a certain type.

There is not, however, any temporal or causal order built into conditional probability relations. Just as there is a probability of heads given that the fair coin was selected, there is a probability that the fair coin was selected given that the coin landed heads. Relative to F, the conditional probability that the fair coin was selected, given that the coin landed heads, is 40%. Here, the conditional probability reflects the a tendency of a process-type to *have had* unfurled in a certain way, given that it had later unfurled in some other manner.

Both kinds of conditional probability may be given the same gloss in terms of frequencies. If the conditional probability of A given B (relative to F) is s, then within the class of possible F-tokens that are *also* B-tokens, the limiting relative frequency of A-tokens is s. Suppose F is repeated many times. That is, I repeatedly reach into the bag, select a coin, and flip it. Within the class of these actual flips that ended with a coin landing heads, the relative frequency of times that I will have selected the fair coin will (most likely) be approximately 40%.[6]

As I have been tacitly assuming, objective probability relations satisfy the axioms and theorems of probability theory, relative to a set-up. For example, the Law of Total Probability states that, for any event-types A and B, $P(A) = P(A \mid B)P(B) + P(A \mid {\sim}B)P({\sim}B)$ (where all probabilities are relative to the same suppressed set-up). The law is a theorem of probability theory. Consequently, the law constrains the possible objective probability relations that a set-up may bear to various types. The law rules out objective probabilities that are, in any case, intuitively impossible. For instance, A cannot be 50% likely if A is 10% likely given B and B is guaranteed by the set-up.

This completes my basic model of objective probability. I turn now to briefly defending realism about my model, and hence about objective probabilities as described by that model.

[6] This approach to objective conditional probabilities, in which the condition is temporally after the event receiving the probability, is inspired by Gillies (2000)'s propensity theory. Like Gillies' account, my model avoids Humphrey's Paradox. Humphrey's Paradox is simply the observation that objective probabilities cannot straightforwardly be identified with dispositions to cause, since conditional probabilities are temporally "reversible" whereas causal relations are not. My model remains broadly propensity-like in spirit, however. Unconditional probabilities mark dispositions of certain types of systems to unfurl in certain kinds of ways. Conditional probabilities mark constraints that some parts of the process have on other parts. "Forward" conditional probabilities mark the causal influence that a part of the process has on subsequent parts. "Backward" conditional probabilities mark the fact that an outcome can only be produced in certain kinds of ways. There are possible ways the process could have gone that would not have resulted in the given outcome. Those ways are ruled out, or made less likely, by the supposition that the outcome *did* occur. But the relations between outcomes and possible processes leading to them are still grounded in the dispositions of the system to unfurl in various ways.

1.2 The explanatory roles of objective probability in the special sciences

To begin, I will describe an explanation from evolutionary biology.[7] The gene that codes for hemoglobin in humans comes in one of two variants or *alleles*. Call the variants S and A. Within West African populations, it is observed that the ratio between the two alleles is constant across generations. For example, suppose that within the current generation of West Africans, 83% of all hemoglobin genes in the population are A and 17% are S. The observation is that in subsequent generations, the A allele will continue to account for 83% of all hemoglobin genes and the S allele for 17%.

Each human cell contains two copies of every gene. Accordingly, with respect to the hemoglobin gene, a human may be *homozygous* (both copies of the gene are of the same variety—either both S or both A) or *heterozygous* (one copy is S and the other copy is A). Humans that are SS homozygotes are substantially more susceptible to sickle-cell anemia—an often fatal disorder. This susceptibility is due to the fact that a single copy of the S allele causes sickling of blood cells, and two copies cause sickling severe enough for anemia. In light of its deleterious effects, one might expect that selection pressures would drive the S allele out of existence over generations. And yet the S allele occurs in the population with a frequency that is stable across generations. What explains the stable frequencies of alleles across generations, and why do those frequencies have the particular values they have?

Evolutionary biology offers an explanation. Each genotype (i.e. one of the three possible combinations of the two alleles) has a certain level of *fitness*. More precisely, the fitness of a genotype is the *objective probability* or *chance* that an individual with that genotype in

[7] My discussion of this explanation is based largely on Sober (1984, Ch. 1). I elide several details that are of importance to evolutionary biology, but that are immaterial to my aim here, which is to illustrate the kind of role that objective probabilities play in evolutionary theory, and indeed, in the special sciences more generally.

the relevant kind of environment survives to adulthood and reproduces.[8] Let us suppose that the AA genotype has a fitness of 0.85. In other words, the objective probability that an individual with the AA genotype survives to adulthood is 0.85. This probability reflects several factors relevant to survival, including the likelihood of encountering threats in the environment, the severity of those threats, and the capacities that an AA homozygote has for coping with them. An SS homozygote in the same environment contends with the same threats an AA homozygote does, and also with an additional one: an increased susceptibility to anemia. Suppose that this additional threat lowers the chance that an SS homozygote survives—and hence the fitness of the SS genotype—to 0.65. The AS heterozygotes, on the other hand, are not similarly susceptible to anemia. Additionally, their slightly-sickled hemoglobin confers an increased resistance to malaria—a significant threat in many parts of West Africa. Accordingly, we may suppose that the fitness of heterozygotes is 0.9—-slightly fitter than the AA homozygote, and much fitter than the SS homozygote.[9]

With these fitness values in hand, we may now explain the stability of allele frequencies over time. We begin by fixing the relative frequencies of the three genotypes in an initial population of zygotes (i.e. in an initial population of fertilized eggs, each of which may or may not survive to adulthood). These initial relative frequencies can be arbitrarily fixed, subject to general genetic constraints—they will be washed out. We can imagine for example that each genotype occurs with equal frequency in the initial pool of zygotes. Suppose that the zygote pool is very large, so that each genotype has many instances in it. As noted,

[8] The notion of fitness I describe in the text is known as a viability fitness. Evolutionary theory also countenances a notion of reproductive fitness, which is often understood to be the expected number of offspring an organism of a certain genotype will generate in the relevant kind of environment. Reproductive fitness is also understood in terms of objective probabilities: expected number of offspring is an average of each particular offspring number, weighted by the chance that the organism has that particular number.

[9] The absolute fitness values I ascribe here are arbitrarily chosen. In fact, the explanation does not require that fitness values have any particular absolute values. Stable presence of the S allele is predicted so long as the heterozygote has the greatest survival chance. The particular stable-state frequencies of the alleles requires that the survival probabilities fall into certain ratios. In order to ultimately explain the stable allelic frequencies described above, all that is required is that the heterozygotes be approximately 1.38 times fitter than SS homozygotes and approximately 1.05 times fitter than AA homozygotes. Such ratios are still, however, ratios between quantities grounded in objective probabilities of survival and reproduction.

each zygote has a particular chance of surviving to adulthood, depending on its genotype. Since we have a large population of zygotes, each of which exhibits the same chance of survival (and whose survivals are assumed to be independent of one another), Bernoulli's theorem applies.[10] By Bernoulli's theorem, the proportion of zygotes of a certain genotype that develop and survive to adulthood is expected to be approximately equal to the survival probability (the fitness) of that genotype. For example, we expect that approximately 65% of SS homozygotes in the population survive to adulthood. We assume that, once all the initial zygotes have either died or reached adulthood, random mating takes place among the surviving adults and produces a second generation of zygotes, in accordance with Mendelian principles.[11] Under these assumptions, the relative frequencies of genotypes among the reproducing adults determines the frequencies of genotypes in the next generation of zygotes. This process is then iterated: a certain proportion of the second-generation zygotes of each genotype survives to adulthood in accordance with the survival probabilities, at which time the surviving zygotes randomly mate, etc. It is assumed that the fitness of a genotype is constant across generations.

It can be shown that, in such a process, the frequency of alleles reaches a stable equilibrium. Moreover, the stable frequency of a genotype is a function only of that genotype's fitness—that is, of its survival probability. Given the particular fitness values described two paragraphs above, the frequency of A alleles is expected to reach 83%, and stay there

[10] Bernoulli's theorem is an instance of a "law of large numbers." Suppose there are n instances of some "experiment," each of which results in "success" with probability p and "failure" with probability 1-p. Suppose also that each experiment is probabilistically independent (a condition that is satisfied if, but not only if, their outcomes are uncorrelated). In the n instances, there will be some particular frequency f of "successes." Bernoulli's theorem informally states that, as n increases without bound, the frequency f and success probability p are arbitrarily close, with probability 1. Furthermore, the theorem enables us to say (again informally) that, if n is large, the probability that f and p diverge by some fixed amount is very low. Thus, if the population is large enough, it is extremely likely that the proportion of surviving genotypes is very close to the genotype's survival probability. For a more precise presentation of these mathematical results and illuminating philosophical discussion, see Mayo (1996, pp. 169ff.).

[11] "Random mating" means that probability of choosing a particular genotype to mate with is approximately equal to the relative frequency of that genotype at the time of mating, and is not dependent on e.g. the fitness of that genotype. Although I focus on fitness as my exemplar of objective probabilities in evolutionary biology, this assumption of random mating is another instance.

indefinitely. Thus, we have explained the observation we initially sought to explain.

This explanation gives only a small taste of the range, detail and variety of explanations present in evolutionary biology, but it nevertheless illustrates the role that the §1 model of objective probability plays in scientific explanations.

Foremost, the explanation is best construed as taking objective probability to be a relation between repeatable types, relative to a set-up. To claim that heterozygotes have a fitness of 0.9 is to claim that a certain kind of process—crudely, *development of an AS heterozygote in West Africa*—results in a certain kind of outcome—*reaching adulthood*—with a probability of 0.9. The same process-type is related to a different kind of outcome—*death before adulthood*—with a probability of 0.1. Each token development of an AS heterozygote in the environment is an instance of the former process-type, and it results in one of the two types of outcome.

One could offer a different interpretation of the objective probabilities in the explanation as relating *tokens*. On this interpretation, an objective probability is a relation that holds between a token instance of zygote development and a token outcome event. To say that an AS heterozygote has a fitness of 0.9 would then be to say that, for each token instance of AS zygote development, its probability of resulting in survival to adulthood is 0.9. Moreover, this interpretation of the probabilities coheres with the invocation Bernoulli's theorem, which requires only that each instance in the population have the same objective probability.

The token interpretation, however, does not obviously do justice to the role of objective probability in the explanation. In the explanation, ascription of a survival probability to a genotype functions to specify a *kind* of causal structure and to mark each possessor of the genotype as being embedded in that kind of causal structure. In specifying a kind of causal structure, an ascription of survival probability identifies certain causal factors as bearing on survival. A survival probability in part reflects the potentiality that a zygote has, by dint of its genotype, to become a creature with certain sorts of capacities (e.g. to resist malaria) and vulnerabilities (e.g. to become sickle-celled). Such potentialities and capacities are causal factors that bear on survival. A survival probability also reflects certain causal factors in the environment that bear on survival, such as the prevalence of malaria. Ascription of survival probability marks the survival of each possessor of the genotype as being subject to the same causal factors. It thereby groups the possessors of the genotype as being similar in explanatorily relevant ways. The causal structure is specified at a relatively high level of abstraction, leaving open many details. As a result, the specified causal structure does not *guarantee* any particular outcome. Nevertheless, the probability reflects the relative importance or impact of the specified causal factors on survival outcomes, whilst allowing for various unspecified causal factors to differently impact each token process, perhaps critically (imagine that a lightning strike kills one of the zygotes).

Viewing the objective probabilities as relating types accommodates the explanatory function of survival probabilities to subsume a class of individuals under a common high-level causal structure. The process-type that is one relatum of the objective probability is at least partly determined by the high-level causal structure identified in the explanation. By contrast, viewing objective probabilities as relating tokens or instances of processes does not easily accommodate the explanatory function of survival probabilities. On such a view, ascription of the same survival probability to each token zygote development would be without explanation. Each token process is suffused with detail and instantiates indefinitely many types at once. Ascription of a probability to a token process *tout court* fails to discriminate between aspects of the token process that bear on the outcome and those that don't, and so *eo ipso* fails to highlight the relevant causal structure that is, in fact, highlighted in the explanation.

The explanation of equilibrium allelic frequencies with which we began also illustrates the connection between objective probabilities and relative frequencies. The explanation proceeds by assuming that a generation constitutes a sample from the relevant fitness distributions, and so that relative frequency of survivors in a generation is approximately equal to the probability.

As discussed above, the ascription of a survival probability in the explanation presupposes some specification of the set-up to which it is relativized. In actual scientific explanations involving survival probabilities, the specification of the set-up is often somewhat vague and open-ended. Above, I crudely described an AS heterozygote's survival probability as attaching to the process-type *development of an AS heterozygote in West Africa.* This typing is too crude, however. As noted, a set-up is taken to specify causal factors, at a certain high level of abstraction, relevant to the outcome in question. A fuller specification of the set-up might be *development of an AS heterozygote in an environment in which malaria is prevalent and treatment of sickle-cell anemia is very hard to access.* But this still underplays the richness of causal factors presupposed in the ascription of the probability. The process of zygote development is assumed to be of a certain standard kind, as might be explained in detail by a molecular or developmental biologist. Certain standard causal processes of embryo development, gestation, and so on are assumed to operate. If, counterfactually, heterozygote chromosomes were present in an organism with very different development patterns, the chance of survival may be very different. On the environmental side, malaria is assumed not only to be prevalent, but to operate via standard causal processes of transmission, as might be studied by entomologists or epidemiologists.

So the objective probability is tacitly understood to attach to a relatively specific kind of set-up, which may be only partly specified or partly specifi*able* in the explanation. On the other hand, the set-up cannot be understood too finely, since it must be generic enough to be plausibly instantiated by each member of the population in question. Instantiation of the process-type cannot require, for instance, that a heterozygote grows up far away from wetlands. If it did, the heterozygote's malarial resistance would be otiose, thus negating (ceteris paribus) any fitness advantage over the AA homozygote.

As in the above explanation, much about the background set-up is left tacit. But the set-up presupposed in one explanation can become an object of theoretical inquiry in its own right. The ascription of fitness in the above explanation was somewhat cavalier. The numerical value of the fitness may have be estimated from observed survivorship frequencies (e.g. by consulting hospital records). A general understanding of the environment at issue may suggest that the fitness differences so estimated have something to do with different relations to malaria and anemia. Beyond this, a more precise specification of the set-ups for genotype fitnesses was not provided. But consider theoretical work such as Cody (1966). Different birds in different sorts of environment have different average clutch sizes (i.e. they lay different numbers of eggs in any one brood). Fitness values may be estimated, and given a loose explanation in terms of generic features of environments. Cody presents a principled way, following the modelling work of Macarthur and Levins (1967), to specify the relevant environmental types more precisely. Cody provides theoretical descriptions of environments whose parameters are the different sorts of causal factors that might affect the reproductive fitness of bird species (see note 2 above). Such parameters include factors such as clutch size, climate stability, degree of competition for food, and so on. These theoretical descriptions enable relatively precise type-identification of environments. Cody also describes principles that enable one to infer long-run facts about bird populations in environments typed-identified with particular values of these parameters. The theoretical model enables us to explain why certain kinds of birds with specific clutch sizes dominate in certain types of actual environment. Such a model characterizes the causal structure of the set-up more precisely.[12]

[12] In my discussion of objective probability in evolutionary theory, I have elided two major questions in the philosophy of biology. One major question is whether the reference to objective probabilities in evolutionary explanations is eliminable. A classic alternative is to see the actual population frequencies as explanatorily sufficient in themselves, without needing to see them as arising from repeated chance processes. In response to this alternative, three things may be said. First, the actual population frequencies are seen as arising from the interplay of particular causal factors. Ascriptions of chance mark those causal factors as relevant; the mere existence of actual frequencies does not. Second, evolutionary biology marks a fundamental distinction between evolution by selection and evolution by random genetic drift. A classic philosophical account of the distinction is that a change in gene frequencies is a result of selection when the gene frequencies align with survival probabilities, and is a result of drift when the gene frequency fails to align with the survival probabilities (due to "random sampling error"). It follows from this account of the selection/drift distinction that a distinction between chance and the frequencies they generate is fundamental to evolutionary theory. However, it is hotly contested in philosophy of biology whether this account of the selection/drift distinction is correct, and even whether such a distinction is fundamental in evolutionary theory. Third, as noted by

I take the references to objective probabilities in explanations, such as the one just discussed, to warrant realism about objective probabilities.

1.3 Some important probability concepts

In §1, I described objective probabilities as relations between set-ups and event-types. From this basic class of relations, more complex probabilistic relations may be defined. In this section, I define those concepts. I illustrate them with examples involving perception, to introduce the reader to the context in which we will be using these concepts throughout.

1.3.1 Random variables and probability distributions

Suppose that, relative to set-up E, X is a set of state- or event-types $x_1, x_2, \ldots$ which are mutually exclusive and exhaustive; that is, exactly one state-type in X will be instantiated in any instance of E. I will call X a *random variable*. To illustrate, consider a set-up in which a creature looks at scene with a slanted surface. The slant of a surface is the degree to which it is tilted away from the creature. In any instance of the set-up, the surface must have a specific slant property s from the set S of possible slants, and it cannot have more than one. Thus, relative to the set-up, S is a random variable. More precisely, we may measure the slant of a surface by its degree of rotation about an axis that is parallel to the horizon from the creature's perspective. If the surface is frontoparallel (i.e. roughly parallel to the creature's face), we say that the slant is $0°$. If the surface is inclined away, its slant

Mills & Beatty (1979), statistical significance tests are run when fitness values are estimated from data about actual frequencies of survival. This already implies that fitness is not the actual frequency, and that the actual frequency is data for estimating something else—namely a chance. The other major question I elide is whether fitness values are themselves explanatory, and if so, in what way. I am sympathetic to the position defended by Sober (1984, Chs. 1 & 2). Fitness values are not themselves causes. They do, however, mark relevant causal structure, and so are part of causal explanations. Additionally, fitnesses serve a unifying explanatory function. The use of objective survival probabilities enables us to see what is in common between all populations in which deleterious alleles persist, abstracting away from the particularities of the organisms and environments involved.

is 45°, and if the surface is inclined towards the creature, its slant is -45°. Thus, the set of possible slants that a surface may have is modeled by the interval S = [-90°,+90°]. This interval of slants is a random variable.

The term "random variable" comes from probability theory, but I abuse the notation somewhat. In probability theory, a random variable is a *function* from the set of possible maximally-specific outcomes of a process (a "sample space") to the real numbers. In our example, a random variable would map a token outcome of the perceptual set-up to real number s if and only if the slant of the surface was $s°$ in the token. A random variable thus does partition the set of tokens into mutually exclusive and exhaustive types. It also associates those types with numbers. The quantitative structure induced by a random variable enables meaningful definitions of quantities such as averages. If a set of exclusive and exhaustive types has a natural metric structure (as the set of slants S does), I'll assume that averages and other such quantities are taken with respect to that natural metric structure. If a set of exclusive and exhaustive types does not have a natural metric structure, I will call it a "random variable," but will not make reference to quantities such as averages on the set.

A random variable X may be a finite set, or it may be a discretely or continuously infinite set. I will follow standard probability theory notation and use "$X = x_1$" to refer to the event-type x_1 (while also making explicit that it is associated with the random variable X). Thus, we may write $P_E(X = x_1) = s$ to mean that the probability of x_1 in E is s. Similarly, $P(A \mid X = x_1) = s$ means that the probability of A, given that x_1 is instantiated, is s. The notion of a random variable, and its attendant notation, will be very useful for us as we proceed. We will be concerned in large part with aspects of probability distributions (to be defined shortly), and distributions are most perspicuously described in the language of random variables. I'll also utilize the above notation to refer to instantiations of properties in instances of a set-up. If e is a token event, and the type x_1 is instantiated in e, then I will say that $X = x_1$ in e.

The random variable notation can be used to express conditional probabilities. Recall our perceptual set-up type. Suppose that, relative to the set-up, the creature's perceptual system will represent the surface as having some slant. Specifically, the system will *attribute* a slant property to the surface. This attribution could be accurate or inaccurate. Following the terminology of Burge (2010), I will call a perceptual state that functions to attribute a property (such as slant) to a surface a *perceptual attributive.* Furthermore, I will refer to perceptual attributives and other representational states using an underlining convention. For instance, I will use "$\underline{45^\circ}$" to refer to a perceptual attributive that attributes a slant of 45° to a surface. The convention is that the linguistic expression underlined suggests the representational content of the attributive. Let $R =[\, \underline{-90^\circ}\, , \underline{+90^\circ}\,]$ be the set of visual slant attributives.[13] R is a random variable. We may suppose that, if the surface has slant 45°, there is a 0.2 conditional probability that a representation as of 45° slant will be produced. In our notation, we may write: $P(R = \underline{45}^\circ \mid S = 45^\circ) = 0.2$.

Predominantly, we will be concerned with random variables that correspond to a continuum of stimulus magnitudes (such as length, distance, or slant) or to a continuum of representations of such stimulus magnitudes (such as representations of length, distance, or slant). We will also be concerned with random variables expressing various ancillary conditions of sensory scenarios, such as degrees of fog, blur, garbling, gaze time, etc.

A single type will bear conditional probability relations to all elements of a random variable, relative to a set-up. For instance, any individual slant-type s in S will bear conditional probability relations to every representation type in R. The probability of $R = r_1$ given S $= s$ will be well-defined, the probability of $R = r_2$ given $S = s$ will be well-defined, and so on. The conjunction of all these individual conditional probabilities is a **conditional**

[13] On Burge (2010)'s use of the term "attributive," an attributive identifies not just the property represented by a perceptual state, but also its *mode of presentation.* The idea is that, just as "Hesperus" and "Phosphorus" can co-refer and yet have different psychological significance, so too can perceptual states pick out the same property but do so with different psychological significance. One attributive for *red* may be systematically produced under white-lighting conditions, and another may be systematically produced under blue-lighting conditions. The states involve different attributives. For now, however, I ignore these fine-grained differences. So a type in R (say, $\underline{+45}^\circ$) is understood as a type shared by all attributives that represent the property 45°, whatever their modes of presentation.

probability distribution over $\boldsymbol{R}$ given $\boldsymbol{S = s}$ (relative to some set-up E). I will use the notation $\mathbb{D}_E[R \mid S = s]$ to denote the distribution over R given that $S = s$ (relative to set-up E).

A distribution over a random variable X is typically specified by a **distribution function** that gives, for each element of X, the probability of that element (potentially given some conditions). Such functions can be graphed. For instance, Figure 1.1 below depicts a hypothetical conditional distribution over representations in R, given that the slant is $S = 45°$.

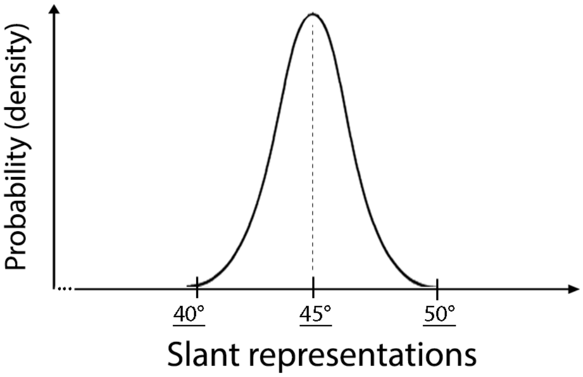


Figure 1.1: An example of a graph that summarizes the probability distribution over R given $S = 45°$.

Each point on the x-axis corresponds to one of the representation-types in R. For instance, the point labeled '$\underline{45°}$' on the x-axis corresponds to the attributive type that indicates the slant property 45°. The height of the curve above a representation-type r depicts the probability that r occurs, given that S = 45°. As we can see, given that S = 45°, the most probable representation is as of 45°, with representations on either side tapering off in probability gradually and symmetrically.

Strictly speaking, the value assigned by a distribution function to a representation-type r is not a probability but a probability *density.* The set R is modeled as being isomorphic to some interval of real numbers. The *probability* of any individual member of R is 0. The distribution function assigns a non-zero probability *density* to each member of R. Probability densities may be used to give the probability that a continuous magnitude (such as slant representations, as we have idealized it) is within an arbitrarily small range. If $f(r)$ is the probability density assigned to r by the distribution function f, then the probability that the representation produced is in the small interval $[r-\varepsilon/2\ ,\ r+\varepsilon/2]$ is approximately $f(r)\times\varepsilon$, with the approximation getting better as ε becomes smaller.

All that being said, I will mostly slur the distinction between probabilities and probability densities for expositional convenience. I take it that, ultimately, the use of real numbers to model sets of properties is an idealization. There are finitely many possible perceptual representations of slant. Distribution functions assigning densities will approximate the probabilities that each of the finitely many representational states have.

In addition to conferring a probability (density) on each representation, a distribution also confers probability on a representation's falling within some *interval* of representations. For instance, the probability that the representation is in the interval [$\underline{45}°$, $\underline{50}°$] (given S=45°) is given by the area under the distribution curve between those two points. In this case, it is nearly exactly half of the total area enclosed under the curve.[14]

The distribution illustrated in Figure 1.1 is an example of a particular kind of distribution known as a **Gaussian distribution**. Gaussian distributions are "bell-shaped," assigning a maximal probability to a central value and tapering off gradually and symmetrically around this central value. Gaussian distributions are pervasive in nature. They are ubiquitous in sensory and perceptual psychology. Gaussian distributions characterize the conditional response properties of many sensory states. The distribution illustrated in Figure 1.1. is a realistic distribution over slant representations, given a distal slant.

In what follows, various properties of distributions will be important. For one, a distribution $\mathbb{D}[R \mid S=45°]$ has a **mean** or **expected value**. This is an average of the representations in R, weighted by their probability of occurrence (given S = 45°). For another, the

[14] A Gaussian distribution, such as the one depicted in Figure 1.1, has a strictly non-zero value on all real numbers. Hence "almost exactly" half.

$\mathbb{D}[R \mid S = 45^\circ]$ has a **mode**; that is, a member in R that is most probable. In Figure 1.1, the mode is $\underline{45^\circ}$. Since the example is *Gaussian* , the mean and the mode coincide—in this case, both are a representation as of $\underline{45^\circ}$. The average representation produced, as well as the representation most frequently produced, across repeated instances of a slant with 45°, is approximately $\underline{45^\circ}$.

The distribution $\mathbb{D}[R \mid S = 45^\circ]$ also has a **standard deviation**. The standard deviation reflects the amount that the values of a random variable vary around the mean value. It is a numerical measure of how "spread out" the set of probability values are across the whole set R. Intuitively, it corresponds to the "width" or "fatness" of the curve. We may denote standard deviation with σ. Given the Gaussian distribution, σ denotes the size of the interval [$\underline{45}$°, $\underline{x}$°] of representations such that the probability that the produced representation is in that interval is 34.1%. Suppose the standard deviation in the example is $\underline{1}$°. Thus, with probability 34.1%, the produced representation will be in the interval [$\underline{45}$°, $\underline{46}$°]. Many representations on the "tails" of the curve have near-zero probability. Some representations near $\underline{45^\circ}$ have an appreciable probability. The standard deviation squared is called the **variance.** Standard deviation has the virtue of being in the units of the random variable (here, units of representational psychological magnitude). Variance is mathematically easier to work with. A large variance means that the distribution is spread out quite a lot, and that large deviations from the mean are relatively probable. A final quantity is the **precision** of a distribution. The precision of a distribution is the reciprocal of the variance. A larger precision means a "narrower" distribution, hence that the likely elements of R are clustered around the average. All three quantities—standard deviation, variance, and precision—are measurements of the "spread" or **dispersion** of a distribution.

Distribution properties such as the mean and variance can be thought of as abstract probability relations between types, grounded in base-level probability relations. For instance, we can think of the specific variance of the distribution for $\mathbb{D}_E[R \mid S = 45^\circ]$ as a many-place relation between the type S = 45°, each type $r \in R$, and the set-up type E.

Distributions have other properties besides mean and variance. For one example, a non-bell shaped distribution might skew more heavily above its mean than below. For another example, suppose that given initial velocity 0.1 m/s, the probability of heads is 0.6 and the probability of tails is 0.4. Where C is the random variable { *coin comes up heads, coin comes up tails* } and V is a random variable corresponding to initial velocity, the distribution C | V = 0.1 m/s has the property *making heads more likely than tails.* The distribution may share this property with other distributions, such as the distribution over C given that V = 0.2 m/s.

1.3.2 Instantial probabilities properties

So far, we have only been considering probability relations between types, and probability properties derived from these relations (e.g. variance). I turn now to articulating probability properties and relations that can hold of *tokens.*

We want to include a notion of probability properties/relations on tokens for two main reasons. First, we want to be able to capture probabilistic discourse that intuitively seems to be partly token-level. For one example, in both philosophical and scientific literature, one will sometimes read that a token event *makes probable* some other type of event. For instance, a token event of the surface being slanted 45° *makes probable* that the representation will be as of 45°. We want to understand what is meant by this better. More generally, some probabilistic discourse seems to pertain specifically to token events or situations, and seems to have truth-conditions that depend on the token in some way. Consider a submarine technician who, upon seeing a dot on his sonar screen, yells: "There's a 90% probability that the enemy is nearby!" This assertion seems intended to characterize the current nautical circumstance the crew finds themselves in. It is not intended merely as an ascription of a higher-order relation between abstract types. Moreover, the assertion's truth or falsity seems to hinge on the current circumstance in some way. The technician's intentions may determine a set-up type that the current environment is presupposed to instantiate, but which it may not. Even if the current environment is of the relevant set-up type, the ascription of the probability may be wrong; perhaps actually there is only a 70% probability that the enemy is nearby.

Second, we will later consider the idea that visual systems contain states that function to correlate with probabilistic properties, just as they contain states that function to register or correlate with the orientation of an edge at a location in the retinal image. Correlations are manifested in empirical statistical relationships between patterns of instantiation of the correlated properties by the objects or processes that bear the correlation. If an orientation detector state correlates with the orientation of an edge, then we would expect to see, empirically, that instantiations of the state *by the detector* tend to occur along with instantiations of an edge *by the retinal image.* If a sensory state correlates with a probabilistic property along these lines, the question arises: instances of the sensory state tend to occur along with *what* sort of property being instantiated by *what* sort of object? This gives us reason to articulate a notion of probability that applies to a single instance or case.

For instance, suppose a coin is flipped and it comes up heads. I say: “That flip had a 50% probability of coming up tails.” What sort of property is being ascribed to the token flip? I propose that some relational property is being ascribed. *Being 50 meters from the Eiffel tower* is a relational property. It is a property had by various objects, in virtue of their bearing a relation to something else. Similarly, I propose, the assertion ascribes a relation between the token flip, a set-up type, an event-type and a probability relation between the two types. The assertion serves to characterize the token process as an *instance* of some set-up type E. It therein relates the token and the type via the relation of *instantiation.* Furthermore, relative to that set-up type, the event-type *coming up tails* has probability 50%. So, by dint of its instantiating E, the token process is characterized as also bearing a probability relation of 50% to that event type—even if the coin in fact came up heads.

When a token process, situation or event x is typed according to its instantiation of a set-up type E and a probability s of an event-type P, I will say that the event-type P *has* *probability s* in token x (relative to E). (Alternatively, the type *had* probability s in the token; I will ignore any differences in tense here.) Supposing that the probability of type P relative to E is *s* (at the type-level), I will assume that a token *x*'s instantiating set-up type E is necessary and sufficient for P to have probability *s* in *x* (relative to E). Properties of the kind *event-type P having probability s in x relative to E* I will call **instantial probabilities**. I am conceiving of instantial probability properties as one-place properties that can be instantiated by token processes, events, or situations.

It's relatively clear what we are saying about an event token when we type it as an instance of a set-up type E. We are identifying the token as instantiating a certain kind of arrangement of entities with a certain causal structure. But what exactly are we saying about a token when we say that an event-type "has probability *s*" in it? What exactly does a token's relation to a set-up type, event-type and probability consist in? This question can feel more pressing given the following observation. As already noted, a token process can instantiate multiple set-up types that confer different probabilities on the same event-type. A token coin flip may instantiate a set-up type that makes heads 50% likely and another that makes heads 60% likely. One can get a sense of contradiction in hearing that heads has 50% probability and 60% probability in the very same process token (albeit in relation to different set-up types).

I take it that a token process's having an instantial probability characterizes some modal aspect of the token. Spelling this modal property out in any detail will quickly bring us into metaphysics that I mostly wish to avoid. To pursue the question briefly, however, one proposal is that heads' having 50% probability in a token flipping process *e* entails that, although the token flip was one in which the coin in fact came up heads, it *could* have been a token flip in which tails occurred. The numerical probability then characterizes a further modal property of this "could have been." The 50% probability marks that the various processes (type and token) that determine the coin's state have a certain symmetry property; their dispositions do not modally favor one outcome over another.

But it is unclear how to evaluate this proposal. On the one hand, it's easy to think that such modal statements as "the flip could have been one in which the coin came up tails" are necessarily false. The fact that the coin came up heads in *e* is plausibly essential to *e*. The coin coming up heads is partly what makes *e* the individual—the token process—that it is. On the other hand, a lot of our modal discourse seems to presuppose that processes *could* have had different outcomes than the ones they in fact had. After a drunk driver swerves narrowly to miss me, I exclaim "You could have killed me!" Perhaps such discourse does not ultimately serve to ascribe a *de re* possibility to the token process. But then what does it say, exactly? These are complex and delicate matters to resolve.

Nevertheless, structurally, ascription of an instantial probability to a token can be understood as placing the token within a certain class of possibilia. Heads being 50% likely in *e* (relative to E) entails that *e* is a member of the class of possible instances or repetitions of E. And it entails that the limiting relative frequency of token processes within the class that lead to heads is approximately 50%. Moreover, placing *e* within this class entails something about the way that *e* was generated, and the modal properties of that generation process. All E-instances are processes that begin with or involve some values being exogenously produced—a randomly generated number that determines initial velocity of the coin flip, for instance. Given the nature of E, the exogeneous values supplied to any E-instance must come from a process of a certain sort (e.g. a process that selects initial velocity randomly). Placing *e* within the class of E-instances thus marks it as the product of a certain type of process operating. In *e*, the coin was flipped with 0.3 m/s initial velocity in calm weather, let's say. The specific initial velocity was the outcome of a token process of random number generation; the particularities of the calm weather were the outcome of a token meteorological process. Given that *e is* a token of E, these exogeneous token processes are of various probabilistically-individuated types (e.g. they are instances of *choosing a number randomly* or *leading to calm weather with high probability*). Thus we can say: although on this occasion the instance of the exogeneous process-type generated a token coin flip in which heads came up, the exogeneous token is of a type that *could* have generated a token flip in which tails came up. But the exogeneous type also has a particular tendency in generating token coin flips; it tends to generate token flips leading to heads equally as often as it tends to generate token flips leading to tails.

Of course, this explication can feel unsatisfying. We earlier considered the oddity of saying that a token event e could have been a distinct token event *e'*. The exogeneous process token can no more have been a distinct token from the one that it is than the flip token could have been a token distinct from the the one that *it* is. So we're back where we started: what exactly is meant by saying that a different flip token *could* have been generated by the token exogenous process?

I raise these issues just to give a sense of what I am leaving unresolved. I assume that locutions like "That process could have resulted differently" or "That coin could have landed heads" are, often enough, literally true, whatever their metaphysical analysis. I assume that these locutions target some kind of modal structure, and that aspects of that modal structure can be characterized by probability relations. Ascriptions of set-up types to token processes relate the token to that modal structure. Ascriptions of instantial probabilities to or within token processes further characterize the kind of modal structure the token is related to.

Heads having 50% probability in a token e is an example of an unconditional instantial probability. I assume that tokens of set-up types also instantiate *conditional* instantial probabilities, again in line with the type-level probability relations fixed by the set-up. Return to our example of slant representation. Suppose that x is a token instance of the relevant set-up type: the slant of a surface produces a representation of slant. Suppose that the surface is slanted 45° in x. And suppose that, at the type level, the conditional probability of the representation $\underline{45}$° given a surface with 45° slant is 0.2. Then I will say that the slant of 45° *makes probable* the representation type $\underline{45}$° to degree 0.2 in x (relative to the set-up). Alternatively, I will say that slant of 45° *probabilifies* $\underline{45}$° to degree 0.2 in x. A token of the set-up possesses this property if and only if the surface is slanted 45° in the token.

Intuitively, possession of this probabilifying property marks the token as a particular kind of *sub-case* of a process-type. Of course, it marks the token as one in which the slant has 45°. But it also marks the the token as an instance of a type of process which unfolds a specific way along one dimension (in that the surface produced must have slant 45°) but which *could* have unfolded in different ways along another dimension (in that other representations could have been produced than the one that in fact was). It marks the token as belonging to a certain sub-class of possible instances of the set-up type, namely those in which 45° slant is produced. It thereby marks the token as having whatever commonalities are possessed by the possible tokens in that sub-class. In particular, it marks the token as involving a particular kind of process, the kind that produces 45° slant. Processes mutually constrain each other relative to set-up types. The slant production process' having resulted one way constrains how the other processes can unfold. In particular, it constraints the production of representations. It makes certain routes to a representational outcome rarer or harder. It favors other representational outcomes. These consequences on the production of representations is marked by the ascription of the probability conditional on the surface slant.

To extend our notation from before, for a token event x, we may write that $\mathrm{P}(\ \mathrm{R} = \underline{45}^\circ \mid \mathrm{S} = 45^\circ\ ) = 0.2$ in x (relative to a background set-up type). This is equivalent to saying that the slant's being 45° makes $\underline{45}$° probable to degree 0.15 in x.

The probabilification of one type by another in some token x is, as we might say, a non-monotonic relation. 45° slant might probabilify $\underline{45}$° to degree 0.2 in x, but 45° slant *and* the subjects' wearing prismatic lenses might probabilify $\underline{45}$° to degree 0.01 in x. The latter relation types x more narrowly than the former. It thereby types x as an instance of a more specific kind of E-process (where E is the set-up type). That more specific kind of E-process is one that makes production of $\underline{45}$° harder.

In general, I will assume that each type-level probability relation that is relative to set-up E corresponds to some instantial probability property that tokens of E may possess.

Officially, I have been characterizing instantial probabilities as one-place properties of tokens of set-up types. I characterized an instantial probability of 50% of heads as a property of the token flipping process, and not the constituent event that is the flip's outcome. But I will be lax about this grammar. I will sometimes say, for instance, that a surface's having slant 45° probabilifies a representation to some degree. The ascription is still relative to some token process that instantiates some set-up type. I bracket any metaphysical differences between the two.

1.3.3 Instantial distributional properties

As I said, random variables, their distributions, and properties of those distributions will figure centrally in our discussion, so I want to briefly discuss the instantial properties that correspond to these.

Suppose E is a set-up type, X and C are random variables, c is a type in C, and D is the distribution function for X given C = c relative to E. Suppose that *e* is a token of E in which c is instantiated. Then I will say that instantiation of the type c *induces distribution D over X in e* (relative to E). Alternatively, I will say that instantiation of *c fixes distribution D* over X in *e*. For instance, suppose that *e* is a token situation in which the surface has slant 45°. At the type-level, given a slant of 45°, the distribution over representations R is Gaussian with mean $\underline{45}$° and standard deviation of $\underline{1}$°. So, in *e*, the slant of 45° fixes a Gaussian distribution with mean $\underline{45}$° and standard deviation of $\underline{1}$° over R. In our notation, we may express this by saying that R | S = 45° has Gaussian distribution with mean $\underline{45}$° and standard deviation of $\underline{1}$° over R in *e.*

As at the type level, instantial distribution properties summarize a set of base-level instantial probability relations. To say that R | S = 45° fixes a Gaussian distribution with certain parameters is equivalent to saying that S = 45° probabilifies $R = r_1$ to degree p_1, and S = 45° probabilifies $R = r_2$ to degree p_2, and so on.

The above example of an instantial distribution property is one that fixes the distribution exactly. Instantial distribution properties can be individuated less exactly. Given the Gaussian that it fixes, it's also true that S = 45° fixes a standard deviation $\underline{1}$° over R in *e*. (This, again, may be read as shorthand for: the distribution over R, conditional on S = 45°, has standard deviation $\underline{1}$°, relative to E; and *e* is an instance of E in which S = 45°.) In our notation, we may say that R | S = 45° has standard deviation $\underline{1}$° in *e*. (It would be more fully correct to say that the *distribution* on R | S = 45° has standard deviation $\underline{1}$°, but I will sometimes drop the explicit reference to the distribution where convenient.)

This instantial distribution property—the slant of 45°'s fixing standard deviation of $\underline{1}$° over R relative to E—is instantiated by an E-token if and only if S = 45° in the token. We want a way, however, to succinctly state general similarities between E-tokens with respect to the distributional properties fixed in them. Suppose that *a* is an E-token in which S = 45° fixes variance 4 over R, and *b* is an E-token in which S = 50° fixes variance 4 over R. Both are cases in which the slant of the surface induces a variance of 4 over R. The specific way in which slant induces variance 4 differs in the two cases. In *a*, the slant induces variance 4 specifically by being 45°; in *b*, slant induces variance 4 specifically by being 50°. It may be that the full distributions fixed by S = 45° and S = 50° in the two cases are different—they may have different means, for instance. Nevertheless, at the relevant level of abstraction, they have the stated commonality in the variances over the distributions they fix.

Here we see three instantial distributional properties that an E-token may have or lack: being a situation in which slant of 45° fixes variance 4, being a situation in which slant of 50° fixes variance 4, and being a situation in which slant fixes variance 4. The former two properties are *determinates* of the *determinable* third property. The former two properties are specific ways of having the third property. Thus *a* and *b* both have the determinable property; in both cases, the slant (though different) fixes variance 4. They possess these properties in different determinate ways.

Suppose we have a third token situation, *c*, in which S = 55° fixes variance 5 over R.

Then *c* instantiates neither determinate property, nor the determinable property. It does, however, instantiate the determinable property of being a situation in which slant fixes variance 5 over R. And all three situations instantiate a property that is determinable to all the aforementioned distributional properties: being a situation in which S fixes variance some variance V over R. This highest-level determinate is instantiated by any possible token of the background set-up type E.

Notationally, I'll write that R | S has variance 4 in *e,* and mean by this that *e* has the determinable property of being a situation in whose specific slant value, whatever it is, fixes variance 4 over R. The convention is that, if, in an expression, a random variable appears as a condition in a conditional probability or distribution, but without a value specified, then the property denoted by the expression is determinable with respect to specific values of the random variable.

When a fully specific distribution is conditional on the values of multiple random variables, determinate/determinable relations can become more complex. This will be important for us later. To prefigure, suppose that the variability in slant representations is influenced by both distal slant S *and* the quality of the ambient lighting conditions. Suppose L is a random variable, such that L = b is a bad lighting condition and L = g is a good lighting condition. Thus, we can consider the probability of representations given that a surface has 45° slant and the lighting conditions are bad; that is, the distribution over R | S = 45°, L = b. Suppose R | S = 45°, L = b has variance 4. This is a determinate of the property R | S, L=b having variance 4. This latter property is had by those E-tokens in which the lighting is bad, and the lighting together with the slant (whatever it is) fix variance 4 on R. Both of these properties are determinate with respect to a higher level determinable: R | S, L fixing variance 4. This higher-level determinable property is had by any E-token in which any combination of specific S and L values fix a variance of 4 over R.

CHAPTER 2

Probability signaling in sensory systems

In this chapter, I argue that perceptual systems have states that function to discriminate and signal instantial distributional properties. Specifically, I argue that they have states that function to track high-level determinable distributional properties, such as *having a precision* of a certain magnitude. I argue mainly from the results of seminal papers in the psychophysics of cue integration, with special focus on Ernst and Banks (2002). §1 describes the experimental set-up of that paper. §2 describes the probabilistic structure, both assumed and measured, of the experimental set-up. I introduce the critical notion of a *production distribution*. I also describe non-laboratory scenarios that have the same relevant probabilistic structure as the laboratory settings of the paper. §3 describes the two central behavioral findings of Ernst and Banks (2002). §4 provides a model of the cue integration process. I argue that, given two empirically-supported assumptions about the cue-integration process, the perceptual system must have certain states that I call *noise-tracking* states. I provide a basic model of how noise-tracking states are produced and how they influence multimodal integration. §5 describes a general kind of sensory state that I call a *sensory signaling state*. I give examples and explicate the kind. §6 argues that the noise-tracking states meet conditions for being signaling states. I consider the function of noise-tracking states. I further investigate the specific distributional properties signaled by such states. I argue that such states signal a determinable dispersion-related property of *likelihood functions*. Finally, in §7, I explain why the noise-tracking states are capacities of perceptual systems, as opposed to post-perceptual systems.

2.1 The experimental set-up of Ernst and Banks (2002)

I will first give a brief summary of the experimental set-up and results of Ernst and Banks (2002). I will then proceed to describe the experimental set-up in more detail.

The overall purpose of Ernst and Banks (2002) is to investigate the process by which representations of an environmental property are produced from information in different modalities about that property. Subjects were presented with protruding bars of varying widths. They received both visual and tactile information about a bar's width. Unbeknownst to the subjects, however, the visual and tactile information conflicted; each suggested different widths, with a varying level of discrepancy between them. This conflicting information was achieved through the use of a virtual/augmented reality device. Subjects performed a discrimination task that required judging the width of a bar on the basis of the conflicting information received. Additionally, the visual information about the bar was garbled to varying degrees on different trials. Using subjects' responses, the experimenters estimated probability distributions for the subjects' multimodal width representations, given the conflicting visually- and haptically-specified widths and the degree to which the visual information was garbled. The multimodal representation produced was, on average, a compromise between the visually- and haptically-specified widths. When the visual information was more garbled, the multimodal representation was closer, on average, to the haptically-specified width.

Subjects also performed width discrimination tasks when only one of the two types of information was available. In the visual-only tasks, subjects had to judge the width of a bar given only visual information. In these tasks, the visual information was garbled to varying degrees across trials, in the same way that it would be in the multi-cue experiment. In the haptic-only tasks, subjects had to judge the width of a bar given only haptic information. Using subjects' responses in these tasks, the experimenters estimated the probability distributions over visual width representations, given the visual information and its level of garbling. Similarly, they estimates the probability distribution over haptic width representations, given the haptic information available.

The central finding of the paper is an observed quantitative relationship between the distributions of multimodal width representations and the distributions of visual-only and haptic-only width representations. The average multimodal width representation produced given the visually- and haptically-specified widths and a level of garbling is expressible as a weighted average of the average visual-only and haptic-only width representations produced given the same visually- and haptically-specified widths and the same level of garbling. The weights can be expressed as a simple function of the precisions of the visual-only and haptic-only distributions, again given the same visually- and haptically-specified widths and level of garbling. Furthermore, the precision of the multimodal width distribution in such a circumstance is a simple function of the precisions of the unimodal representation distributions in the corresponding visual-only and haptic-only circumstances. The authors point out that this quantitative relationship between distributions is what one would expect of a system that combined cues in conformity to norms of Bayesian statistical inference. Hence the title of the paper: "Humans integrate visual and haptic information in a statistically optimal fashion."

I will now describe the results of the paper in more detail.

Visual-only experiments

The primary task for subjects in Ernst and Banks (2002) is to discriminate the widths of virtually presented bars. Discriminations occur in one of three experimental conditions, distinguished by the sort of information about the bar's width that is available to the subject on a trial: only visual information, only haptic information, or both visual and haptic information.

Consider first trials in which only visual information is available. Visual-only trials are sub-divided into four sub-conditions, called **noise conditions**. In each noise condition, the visual stimulus is garbled by different amounts. Consider first the 0% noise condition, or *no-noise* condition. On trials in such a condition, the visual environment appears to the perceiver approximately as illustrated in Figure 2.1 (from a vantage point that is not the perceiver's own):

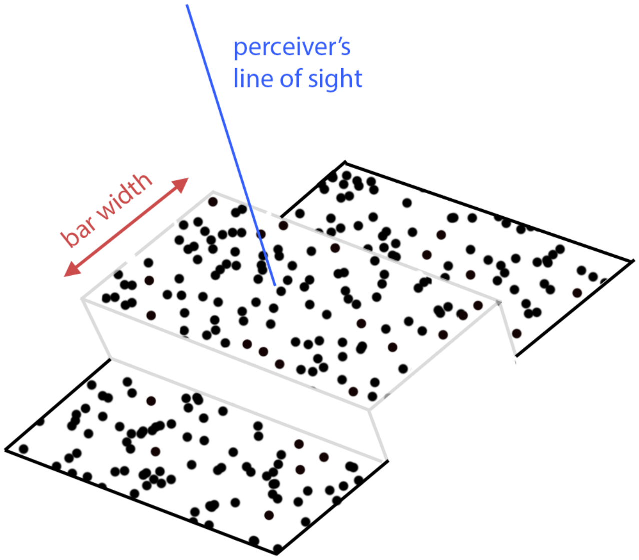


Figure 2.1: Illustration of an example scene as it appears to a subject in the no-noise condition of Ernst and Banks (2002). The illustration is from a perspective that is not the perceiver's own.

On this trial, it will appear to the perceiver that they are looking straight-on at a bar whose only visible surface is fronto-parallel. The bar appears to be protruding from a background plane. Both the facing surface of the block and the background appear to be randomly stippled with small dots. The width of the bar, as indicated in the figure, is the attribute to be discriminated. In this stimulus, all dots appear to be within one of two planes: either on the facing surface of the block, or on the background surface.

In actuality, there is no physical bar. Rather, subjects are hooked up to a virtual reality device with stereoscopic glasses. A stereo image-pair of dots is projected which, given the glasses, engages the subject's binocular disparity-based capacity to recover the depths of the dots. The image-pair is so constructed that, in the no-noise condition, the perceiver will seem to see dots that occur with one of two depths. From these recovered dot depths, representations of protruding and background surfaces are interpolated. The protruding surface is seen as having sharp edges—and hence, a readily discernible width—because of the abrupt depth-discontinuity between dots in the plane nearer to the perceiver and dots in the plane farther from the perceiver.

Despite the fact that there is no physical bar on any trial, I will continue to refer to "the distal bar in a trial." When I say that the bar on a no-noise trial has width 50 mm, this officially is shorthand for: the stereo-image pair produced on the trial contains dot disparities that would be produced, under relevantly similar circumstances, by a stippled bar whose width is 50 mm.

In the no-noise condition, each dot exists at one of two depths; each is coplanar with either the bar's surface or the background. In the *high-noise* condition, by contrast, dots exist at wide variety of depths, as illustrated in Figure 2.2:

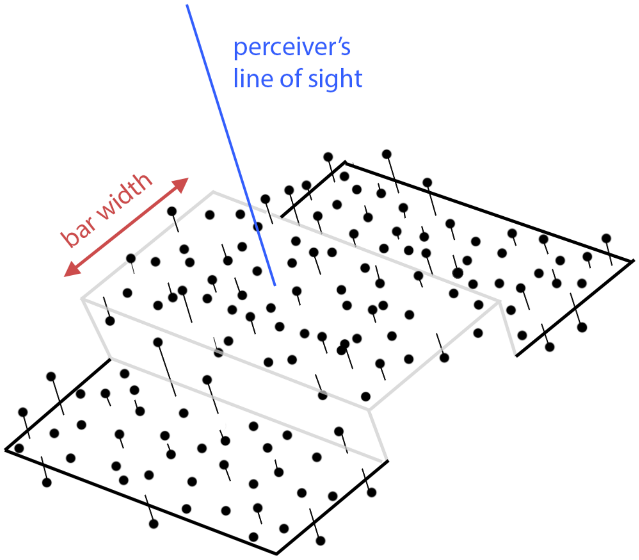


Figure 2.2: Example of a visual stimulus in the high-noise condition of Ernst and Banks (2002)

Even with such a garbled stimulus, the perceiver still interpolates protruding and background surfaces. A line attached to a dot in the above figure depicts the distance in depth of the dot from the surfaces interpolated by the perceiver.

The depths of dots in a high-noise stimulus is determined by a stochastic procedure defined by the experimenters. First, a no-noise stimulus is generated. The width of this no-noise stimulus is chosen uniformly at random from a range of widths. A bar with the given width is generated. The facing surface of the bar and the background plane are stippled with dots, whose locations are also chosen randomly. For the purposes of psychophysical analysis, the width of this no-noise bar is taken to be the "correct" width of the bar, even after it's been garbled. I'll refer to this width as the *distal width* of the bar on a nonzero-noise trial (even though there is no bar and hence no distal width). A high-noise stimulus is constructed from the no-noise bar by randomly perturbing the dots in depth. Specifically, each dot is displaced in depth along the line of sight, either towards or away from the subject, by a randomly chosen amount from some range of possible displacements. What makes the high-noise condition *high*-noise is that each dot is displaced, on average, by a relatively high amount. Visual stimuli in a further condition—the *low*-noise condition—are generated in the same way as high-noise stimuli, except the dots are randomly displaced by an amount that is, on average, relatively low. Which noise condition occurs on a given trial is itself determined randomly, with each noise condition receiving equal probability.

In non-zero noise conditions, due to the displacements, the depth-discontinuities between dots on the "edge" of the block's surface and those in the background plane are less sharp than in no-noise conditions. Intuitively, this makes the block's height less easily discernible. However, in all noise conditions, the depth-discontinuity is sharp enough to (at least typically) support some discrimination of bar widths.

Each visual-only trial is an instance of a type of probabilistic set-up.[1] Each such instance begins with the random selection of a distal bar width from a range of possible widths, each of which is chosen with equal probability. Then, one of four noise conditions is selected with equal probabilities. Once these values are fixed, they feed in as parameters to the stochastic

[1] More precisely, each visual-only trial consists of *two* instances of the set-up type I describe in this paragraph—one for each of the two bars whose widths are to be discriminated. The set-up type I describe types a process that generates a single stimulus that leads to a single sensory response.

procedure of stimulus generation described above. For instance, each trial involves selecting for each dot an amount by which to displace it in depth. The displacement value is chosen uniformly from a range of possible displacements. The range of possible displacements is determined by the noise condition. In a high-noise condition, the range of possible displacements spans minor displacements and drastic displacements. Consequently, the average level of displacement among dots within a single stimulus is higher in a high-noise condition than in a low-noise condition.

In this way, the set-up type specifies the procedure by which stimuli are generated. As needed, we may assume that it also fixes various standing aspects of the experimental context. A fuller specification of the relevant set-up type may include such features as the distance the subject is from the location of the apparent bar, the lighting conditions, the fact that the subject is wearing stereo glasses, the average level of perceiver's attentiveness or boredom, and so on.

On any visual-only trial of a particular noise condition, two bars of (usually) different widths are presented (both garbled via the same noise parameters). The subject must indicate which of the two bars has the greater width. The pattern of subjects' responses, given different distal widths and noise conditions, constitutes part of the main data collected in the experiment.

Haptic-only experiments

Ernst and Banks also ran a series of experiments structurally similar to the vision-only experiments just described, wherein bars are felt but not seen. The same subjects' who completed the vision-only experiments were placed in the same virtual reality chamber as before. In front of them, there is a screen that formerly projected images that (with the stereo glasses) caused a visual representation of a bar. Now, however, no images are projected from the screen. Rather, their hands are placed underneath the screen. Their fingers are placed at a location behind the screen that is identical to the location in 3D space that bars seemed to appear at in the visual-only tasks.

Subjects' thumb and index fingers are fitted into sockets that are attached to rods and a machine. The subject may move their thumb and index fingers freely in a "pinching motion," but once their fingers become close enough, the rods produce resistance that impedes their fingers from moving any farther. The effect is that, when the subjects "pinch" down and meet resistance, it feels as though they are pinching a protruding bar. The machine controlling the rods can vary the point at which the resistance is applied. By varying this distance, the experimenters can vary the width of the "bar" that the subjects are feeling. In one instance, the "bar" might be 50 mm in width—the resistance kicks in when the thumb and forefingers are approximately 50 mm apart. In another instance, the "bar" might be 55 mm. And so on.

Subjects performed width discrimination experiments. On a trial, the subject "pinched" down twice. Each such pinch led to a haptic perception as of pinching a bar of a specific width. Subjects had to judge which of the two "bars" was wider. Such experiments were done over the same range of distal "widths" as in the visual-only experiment. However, unlike the visual-only experiments, no noise was added to the haptic information. (Although it could have been, perhaps by adding jitter to the rods.)

Multimodal experiments

Finally, the same subjects performed a multimodal experiment. Here, in the same virtual reality chamber, subjects were presented with visual and haptic proximal input simultaneously. The visual proximal inputs were the same as in the visual-only trial, and were generated in the same way. Stereo image-pairs were projected, which were generated in the same way as in the visual-only experiment. Similarly, the haptic input was generated in the same way. The visual and haptic input were coordinated, in the sense that, on any given trial, it appeared to the subject that the bar they were feeling was also the bar they were seeing.

However, unbeknownst to the subject, on some trials, the visual and haptic proximal stimulations were in conflict with one another. For example, on a trial, the visual stimulation might derive from a visual distal bar with width 48 mm, and the haptic stimulation might derive from a haptic distal bar with width 52 mm. Another way of putting this: the visual stimulation contained a set of binocular dot disparities that *would*, typically, project from a stippled bar of 48 mm width (potentially under some kind of atmospheric or optical garbling), while the haptic stimulation consisted in mechanical forces that *would*, typically, result from pinching a bar or 52 mm width. Yet a third way of putting this that prefigures later experimental findings: the visual stimulation was an image-type generated by conditions which, in the visual-only experiment, lead on average to a representation as of 52 mm, while the haptic stimulation was generated by conditions which, in the haptic-only experiment, lead on average to a representation as of 48 mm.

The two sources of information (visual and haptic) are called *cues* to width. The conflict in cues did not, apparently, spoil the illusion that the bar being felt was also being seen; subjects reported being unaware of any conflict when asked in debriefing.

Subjects had to perform a multimodal discrimination task. They were asked to judge which of two bars was wider. Both bars were both "seen" and "felt." Only one of the two bars presented on a trial contained a conflict.

Such multi-cue conflict trials were run over a wide range of different visually-specified widths and haptically-specified widths. Moreover, the visual input could be garbled, via the same procedure of garbling from the visual-only trials, and at the same levels. Thus each trial-type in the multi-cue experiment is identified by the visually-specified width, the haptically-specified width, and the level of garbling in the visual input of the conflicted stimulus.

It is assumed that on each trial, the subject produces a multimodal representation of the bar's width. It is not assumed that the processes generating such multimodal representations *will* utilize cues from both modalities. But they may. As before, the experimenters estimate production distributions over the multimodal width representations, given the visually- and haptically-identified widths and a level of visual noise.

2.2 Probabilistic properties of the experiment

2.2.1 Production distributions

Consider the subjects' responses to visual-only trials of a particular noise condition. Ernst and Banks (2002) apply standard psychophysical procedures to this response data. Each presentation of a bar is assumed to produce a psychological magnitude within some internal psychological continuum. Specifically, each visual-only presentation of a bar is assumed to produce a *visual representation* of the bar's width.[2] The subject's judgement of which bar has the greater width on a trial is assumed to derive from a comparison of the visual representations produced; the subject chooses the bar represented as having the greater width. The production of a width representation by and within the surrounding environment is as-

[2] Mathematically, the psychophysical analysis does not require that the psychological magnitudes in the continuum literally represent the distal attribute on the basis of which subjects provide discrimination responses. However, it is standard practice in psychophysical modeling of perception to assume that the magnitudes are representational, or at least registrational (in basically the sense of Burge (2010)). In particular, Ernst and Banks (2002) derive, from the psychophysical procedures, a probability distribution over "internal estimates of height" (i.e. of the bar's width). Since determination of the width operates after stereopsis, and since stereopsis is plausibly a constancy mechanism, I take it as safe that a width representation is produced on every trial. As noted in the main text, in its application to data, the psychophysical analysis presumes that the magnitudes within the continuum produced on a trial drive the subject's response: the subject picks bar 1 as wider if and only if the magnitude deriving from bar 1 is greater than the magnitude deriving from bar 2. Since the continuum is assumed to drive decision behavior in this way, one may wonder how "far" upstream in the processing hierarchy it is from the visual system. If it is much higher up, it may exhibit variability greater than that in the visual representations themselves, owing to the accumulation of internal neural noise and such. Additionally, it could, in principle, access information other than just the visual representations. However, it is standard in psychophysical analysis to assume that the magnitudes are sensory states of a perceptual system. Such an assumption may be bolstered experimentally by employing the methods of signal detection theory. Furthermore, psychophysicists often attempt to estimate the subject's disposition to ignore the stimuli on a trial altogether ("lapse rate") and to guess in response to hard-to-discriminate stimuli ("guess rate"). Such procedures are intended to make the estimates of perceptual representational probabilities more reliable. Ernst and Banks (2002) do not include such a procedure, but they could have. I follow the psychophysics in making this assumption about the perceptual representational location of the continuum.

sumed to be probabilistic: repeated runs of the same trial-type will produce different visual width representations, each with a different relative frequency. For instance, repeated presentation of a bar with a specific width and in the same noise condition will lead to different representations. As far as the psychophysical analysis goes, this probabilistic variation is simply taken for granted. The sources of the variability are not addressed.

The psychophysical analyses enable us to estimate the probability distribution over width representations, given a trial-type. Such a distribution is an example of what I will call a production distribution. A conditional probability distribution is a **production distribution** if it specifies the probabilities with which sensory or perceptual states are produced, given various conditions. The psychophysical analyses are applied to data collected from repeated trials in which the bar has the same width and is garbled via the same noise parameter. The psychophysical analyses provide estimates of production distributions over visual width representations, each conditional on a specific distal width and a specific noise condition.

Two example production distributions are illustrated below in Figure 2.3:[3]

These figures should be interpreted in the same way that Figure 1.1 in Chapter 1 was in the previous section was. Consider Figure 2.3A. On the x-axis, we have the different possible visual representations of a bar's width.[4] The height of the curve above a representation on the x-axis reflects the probability that a representation of that type will be produced on a

[3] The spreads of the distributions do not accurately reflect the numerical variances reported in the original paper. The figure serves qualitative illustrative purposes only.

[4] Note the "as of" locution in Figure 3. The numerical magnitude and the reference to millimeters is used in our "metalanguage" to specify the representatum of the attributive; it is not implied that the perceptual system represents in metric units, or bears representational relations to numbers. Furthermore, the x-axis is assumed to comprise continuum-many perceptual representations, one for each real number within some interval (set by the smallest and largest possible heights that can be perceptually attributed). This is an idealization; presumably there are at most finitely many distinct perceptual representations. This idealization, however, is ubiquitous in the parts of psychophysics that study the representation of continuous magnitudes (such as length), as it permits the use of continuous mathematics in the analysis. Throughout, I follow the psychophysics in making this idealization. Of course, width representations do not occur in a vacuum. Strictly speaking, the event-types on the x-axis should be understood as types like: production of a structured representational state, as of a protruding bar of some shape, size, distance and surface texture, in which a length magnitude of 50 mm is attributed to the vertical span of the bar.

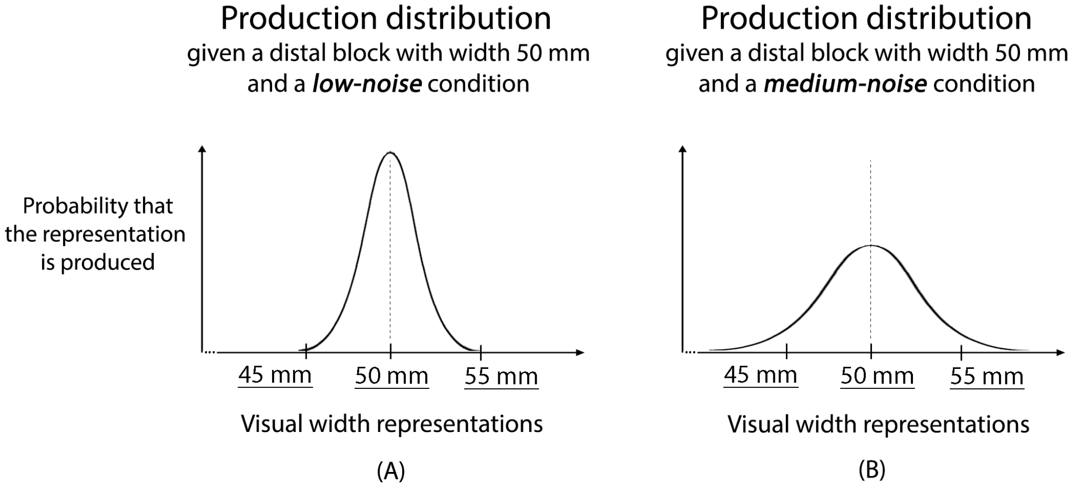


Figure 2.3: Example production distributions from visual-only trials in Ernst and Banks (2002). (A) The production distribution over width representations, given the distal width is 50 mm and a low noise condition. (B) Production distribution for high noise.

low-noise trial in which the distal bar has width 50 mm.[5] Each of the depicted distributions is Gaussian. Thus, given low noise and a distal width of 50 mm, the mean representation produced and the most likely representation to be produced are the same: representation as of 50 mm.

The distributions in both (A) and (B) are conditional on the same distal width: 50 mm. Both distributions have the same mean and mode: a representation as of 50 mm. The distributions are conditional on different noise conditions. (A) is conditional on a low-noise condition, and (B) is conditional on a high-noise condition. The distributions differ in their variances. (A) has a lower variance than (B), as indicated by the greater "width" or "fatness" or "spread" of the distribution in (B). Note that the higher variance in (B) entails that the probability values are more evenly spread out across the possible representation types. Consequently, the probabilities of various representations differ between (A) and (B). For instance, the mean representation as of 50 mm has a lower probability in (B) than it

[5] Since the x-axis comprises continuum many values, the height of the curve above a representation technically reflects the probability density that a representation of that type will be produced. See fn. 4.

does in (A), and similarly for representations near it. Representations fairly far from the mean—for example, representations as of 45 mm—have greater probabilities in (B) than in (A). Thus, for Gaussians with the same means, a higher variance entails that large deviations from the shared mean are more likely.

Notice also that the distribution in (A) has a higher *precision* than the distribution in (B). This illustrates a general trend: the less noise, the more precise the production distribution is.

Ernst and Banks observed several important properties of production distributions on visual-only trials. Let W be the set of distal widths that the bar may have on a visual-only trial. W, N and R are random variables.[6]

First, they observed that the production distribution given any distal width and any noise condition was Gaussian. Furthermore, for all tested distal widths w and all tested noise conditions n, the mean of $\mathbb{D}[R \mid W = w, N = n]$ was a representation as of w. That is, regardless of noise condition, the distal width's being w entails that the mean (hence, given the Gaussian, most likely) representation will be as of w. In general, if a production distribution is conditional on the presence of a stimulus attribute w (potentially along with other conditions), and the mean representation produced is as of w, I'll say that the production distribution is *calibrated*. Ernst and Banks found that production distributions over visual width representations R was were calibrated conditional on any pair of distal width and noise condition. Changing the noise condition does not change the average representation. Intuitively, the distal width is "responsible" for the mean of a production distribution,

[6] Qua set of unadorned experimental conditions, N does not have metrical structure, and so is not a random variable in the technical sense. But again, we will not be asking for averages or variances on N. There is, however, a canonical metrical structure one could apply to the noise conditions. Each noise condition chooses the individual levels of dot displacement from a range of dot displacements. This range is expressed as a percentage of the height that the bar protrudes towards the viewer — 3 cm, in the experiments. In the paper, a 67% noise condition involves drawing displacements uniformly from the range [0 cm, 2.01 cm] (note that $0.67 \cdot 3$ cm = 2.01 cm). The percentages associated with noise conditions induces a metrical structure on the states of N. It also allows us to conceive of N as a continuous variable, of which only four values were used experimentally. In principle, production distributions could differ depending on the specific values of a continuous N.

independently of the noise condition. This finding is partly depicted in Figures 2.3A and 2.3B. Both distributions are conditional on the same distal width, and have the same mean.

The second important observation is that noise condition determines the variance of a production distribution, independently of the distal width. That is, for each noise condition n, there is a unique variance v, such that for any distal width w, $\mathbb{D}[R \mid W = w, N = n]$ has variance v.[7] Increase the noise condition while holding the distal width fixed, and you increase the variance of the resulting distribution. And conversely: if $\mathbb{D}[R \mid W = w, N = n]$ has a greater variance than $\mathbb{D}[R \mid W = w, N = n']$, then n' is a greater noise condition than n. Moreover, if you alter the distal width while holding fixed the noise condition, the variance of the resulting production distribution will be unchanged. This property of production variances is partly illustrated in Figures 2.3A and 2.3B. Holding fixed distal width but moving to a higher noise condition results in greater variance—and so, lower precision.

A specific Gaussian distribution is determined entirely by its mean and variance. In the case of production distributions, two different properties control these parameters separately: distal width controls the mean, and noise condition controls the variance. Thus, together, distal width and noise condition fully determine a production distribution on R.

Similar findings were made for haptic production distributions. Different widths controlled the haptic means, but all production distributions were Gaussian with the same variance. Although different haptic noise conditions were not tested, I will assume that haptic noise conditions would maintain calibration of the haptic production distributions. (See, for example, Serwe et al. 2011.)

[7] The quantification over all distal widths is strictly incorrect. Representation of magnitudes like width mostly obey Weber's Law, which in this case entails that the variance of $R \mid W = w, N = n$ increases as one holds $N = n$ fixed but increases the value of w. It is common in the psychophysics literature to assume that the range of magnitudes (like width) used in an experiment is small enough compared to the full range of magnitudes that Weber's Law may be ignored, and the variances given increasing stimulus values are approximately equal. The quantification over w should be read in this way.

2.2.2 Posterior distributions given representational states

On the model we are working with, objective probabilities are always relative to set-up types; and wherever a set-up type determines a conditional probability $P(E \mid C)$, it also fixes a conditional probability $P(C \mid E)$. Generally speaking, when C temporally or causally precedes E, $P(C \mid E)$ is called a "forward probability," and $P(E \mid C)$ is called an "inverse" or "posterior" probability. (I use E and C mnemonically for 'cause' and 'effect.') These probabilities are related by Bayes' Theorem,

$$P\left(C \mid E\right) = \frac{P\left(E \mid C\right) P(C)}{P(E)}$$

which may be read intuitively here as: the probability of a cause given an effect is the probability of the effect given the cause, times the unconditional probability of the cause occurring at all, divided by the unconditional probability that the effect would occur. The probability P(C) in the formula is referred to as a *prior* probability (since it is the probability that the cause had 'prior' to the unfolding of a process in which E occurred).[8]

Production probabilities are probabilities that a representation will occur, given certain distal stimulus conditions. There are corresponding posterior probabilities over distal stimulus conditions *given* that a particular representation has occurred. For instance, there is a distribution over distal widths, given that the representation produced was R = $\underline{\text{50 mm}}$, and given that the noise condition was n_1. That is, the distribution $\mathbb{D}[W \mid R = \underline{\text{50 mm}}, N = n_1]$ is well-defined in the set-up of Ernst and Banks (2002).

Two remarks about this type of distribution; that is, about posterior distributions like $\mathbb{D}[W \mid R = r, N = n]$ First, it should not be confused with the distribution over $\mathbb{D}[W \mid R = r]$. Both may be called "posterior distributions," since the condition is temporally after the properties receiving the probability. But they are distinct posterior distributions. In

[8] I describe this in terms of cause and effect for the sake of illustration. Bayes' Theorem is a theorem, and holds between pairs of events or random variables regardless of their temporal and/or causal relations.

general, the distributions need not be the same.[9] The latter distribution puts no constraints on what the noise condition is. The probability $P(W = w \mid R = r)$ is an average of the specific probabilities $P(W \mid R = r, N = n_i)$ over every noise condition n_i, weighted by the probability $P(N = n_i \mid R = r)$. This latter distribution characterizes a tendency among types of processes that resulted in $R = r$ to have involved different distal widths, but which processes may fold in any way that is compatible with their resulting in r. It counts r tokens produced in various noise conditions as contributing to the condition. The former posterior distribution characterizes a tendency in processes of a more restricted type: a tendency of processes with a specific noise condition that result in $R = r$ to have involved different distal widths. The two distributions *may* be identical, but generally they will not be. Either because the different noise conditions each make comparable contributions to the relevant process, or because they make no contribution at all, we could imagine scenarios in which the distributions are identical.

The second remark is that the posterior distribution $\mathbb{D}[W \mid R = r, N = n]$ is also governed by a form of Bayes' Theorem, in which the posterior probabilities for each w are mathematically related to the production probabilities we have already discussed: $P(R = r \mid W = w, N = n)$.

Generally speaking, one can know the forward probability without knowing or being able to compute corresponding posterior probabilities. However, the set-up and empirically observed properties of Ernst and Banks have various special properties that *do* permit calculation of the relevant posteriors.

Each distal width w is chosen with equal probability in the experiments. Thus, the unconditional probability $P(W = w)$ is a constant for all distal widths. Additionally, widths and noise conditions are chosen by statistically independent processes, so that $P(W = w \mid N = n) = P(W = w)$. Third, for any visual noise condition n, the production distribution

[9] This is a general point about probability distributions. X | Y =y, Z =z may generally have a different distribution than X | Y = y.

over R is, as remarked, a calibrated Gaussian with a fixed variance.

These three properties entail that the posterior distribution over $\mathbb{D}[W \mid R = r, N = n]$ is itself a Gaussian distribution, whose mean is the distal width indicated by the representation r and whose variance is the variance that $N = n$ specifies for production distributions. So for example, consider the posterior distribution over W, given that a representation $\underline{\text{50 mm}}$ was produced and the noise condition was low-noise. This distribution is illustrated in Figure 2.4.

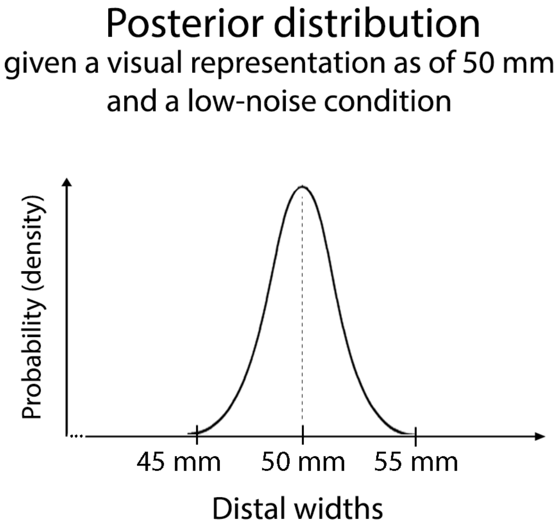


Figure 2.4: Posterior distribution over distal widths, given that a representation of 50mm has been produced and the noise condition is low.

Notice the similarity between this distribution and the distribution of Figure 2.3A. They have the same shape, and identical variances. 2.3A exhibits a kind of 'alignment' between distal width and width representation: the distal width was 50 mm, and the most likely width representation was as of 50 mm. Figure 2.4 illustrates a similar 'alignment': the representation is as of 50 mm, and the most likely distal width is 50 mm.

2.2.3 Probabilities of veridicality and natural analogues

Within the experimental set-up of Ernst and Banks (2002), the probability that any given distal width and noise condition produces a *veridical* representation is 0. Recall that I am using "distal width" to refer to the width of a hypothetical "no-noise" bar from which the presented stimuli are derived. There is no physical bar, hence no width. The representation of a bar suffers a reference failure, and so the attribution of width cannot be veridical. Nevertheless, we can imagine a natural environment that is structurally similar to the Ernst and Banks set-up. In the environment, actual bars appear before the visual system. Bars with different widths are equally likely to appear. The bars have stippled surfaces and protrude out of a stippled background; they appear at similar viewing angles and distances and under similar lighting conditions as in Ernst and Banks. A natural process determines the stippling patterns on the bars, with the same probabilities as in Ernst and Banks. Suppose that the atmosphere in this natural environment can have bizarre properties. Under some atmospheric conditions, the stereo image projected by a bar is distorted so that disparities between corresponding points in the stereo pair are shifted. We may suppose that the atmospheric conditions correspond to the noise conditions of Ernst and Banks. That is, assume that for any laboratory noise condition, there is a type of atmospheric condition that, together with the properties of the physical bar, produces stereo-image I with probability p if and only if the laboratory noise condition and the corresponding distal width produce stereo-image I with probability p. We assume that whatever probabilities govern the sensory transitions from stereo-image to width representation are the same as in the laboratory conditions. We assume that the production distribution over width representations given a distal width and atmospheric condition is identical to the corresponding distribution in the laboratory. I will call any such natural setting that is probabilistically equivalent to a laboratory set-up a **natural analogue** of the laboratory set-up.

By construction, such a natural setting exhibits the same probabilistic structure as the laboratory conditions of Ernst and Banks. So the production distributions for width will be mathematically the same. Given a low-noise atmosphere and a physical bar of 50 mm, the distribution on width representations will be mathematically the same as in Figure 2.3A. In the natural setting corresponding to 2.3A, however, the distribution entails nonzero probabilities for the production of *veridical* representation.[10] Presence of a distal bar with width 50 mm in low-noise atmosphere will most likely lead to a representation as of 50 mm; that is, it most likely produces a perfectly veridical representation. The probability that a 50 mm width will cause a perfectly veridical representation in a high-noise condition (corresponding to Figure 2.3B) is lower. Thus, the different variances in 2.3A and 2.3B correspond to different patterns in veridicality, in the natural setting. Given the calibrated Gaussians, larger variance entails lower probability of a perfectly veridical representation's being produced.

Furthermore, a 50 mm width is more likely to produce an *approximately* veridical representation in 2.3A than in 2.3B. Say that a representation is approximately veridical if the property it attributes to a particular and the property that the particular in fact has differ by less than ε. For instance, suppose the distal width is 50 mm. Then a width representation is approximately veridical only if it is as of a width that is in the interval (50 – ε mm, 50 + ε mm). In other words, the width representation produced is approximately veridical only if it is in the interval of size 2ε centered around the mean representation $\underline{\text{50 mm}}$. The probability that the width representation is in this interval is given by the area under the distribution curve between $\underline{50 - \varepsilon \text{ mm}}$ and $\underline{50 + \varepsilon \text{ mm}}$. So defined, the area under the curve is much less in 2.3B than in 2.3A. So, a 50 mm bar width is less likely to produce an approximately veridical representation in high-noise atmospheres than in low-noise atmospheres. This difference again traces to the fact that 2.3A and 2.3B are both calibrated Gaussians with different variances.

[10] Ultimately, there are complications here that I bracket. For instance, veridical width representation requires successful instances of singular representational content: reference to the bar, and also perhaps a singular representation of the width-instance. There are 'non-deviant causal chain'-type conditions on such singular representational relations' holding. Thus, the probability of veridicality is, in principle, sensitive to probabilities of such causal routes' obtaining.

Notice also that the greater variance of 2.3B entails that the produced representation is more likely to diverge significantly from perfect veridicality in 2.3B than in 2.3A. Numerically, variance may be thought of as a measure of the average deviation of a stochastic quantity from its average. In the natural setting, the variance of the production distribution may be thought of as a measure of the average deviation of produced width representations from perfect veridicality, given a specific distal width.

Thus, in the natural analogue, the production distributions correspond to probabilities that an accurate representation will be produced, given a specification of the environment. The posterior distributions corresponding to production distributions, in the natural analogue, determine probabilities that a representation, *once produced*, is veridical. Suppose a representation 50 mm has been produced in a specific noise condition. Since the posterior distribution is unbiased and Gaussian, the most likely distal width is 50 mm. Each specific distal width corresponds to a specific signed level of representational inaccuracy. Thus, among all possible levels of representational accuracy or inaccuracy, perfect veridicality is most likely. Moreover, the area under the posterior between $50 - \varepsilon$ mm and $50 + \varepsilon$ mm gives the probability that the distal width is in that interval. If again we assume that the representation is approximately veridical if the distal width is within ε of its representatum, then this area gives the probability of the representation's approximate veridicality.

If we compare two representations produced in different noise conditions, we find that their probabilities of approximate veridicality differ. The probability that 50 mm is approximately veridical in a low noise condition is higher than the probability that 50 mm is approximately veridical in a high noise condition. Recall that the variances of the posteriors in noise conditions are identical to the corresponding production variances in those noise conditions. Since the high-noise posterior curve is "compressed" or "squashed" by its higher variance, the probability of approximate veridicality is lower. Corresponding to this difference in probability of approximate veridicality is a difference in probability of *radical inaccuracy*. Distal widths quite far from 50 mm have greater probabilities (given 50 mm) in high noise than in low noise settings. If a representation is radically inaccurate if the property diverges by a large amount from its representatum, then, in a high noise setting, $\underline{\text{50 mm}}$ is more likely to be radically inaccurate.

I emphasize again, however, that there are multiple posteriors over the distal stimulus dimension conditional on a representation. The distribution $\mathbb{D}[W \mid R = \underline{\text{50 mm}}]$ is one thing, and the distribution $\mathbb{D}[W \mid R = \underline{\text{50 mm}}, N = n]$ is another. The latter distribution is the one we discussed in the previous paragraph. The former is a weighted average of the latter types of probabilities.

2.3 The central findings of Ernst and Banks (2002)

2.3.1 The two central findings

Let R_M be the set of multimodal width representations. Ernst and Banks observe that, as in the unimodal cases, the distribution over R_M given visual and haptic widths and a noise condition is Gaussian. The two critical findings involve the mean and variance of this distribution, and their relations to the means and variances of relevant single-modality production distributions.

To illustrate the first finding, suppose the subject is confronted with a conflicted stimulus in which the visual width is 50 mm and the haptic width is 56 mm. Suppose also that the visual stimulus was garbled with a medium amount of noise. We find that the subject's mean multimodal representation in such a circumstance is as of 54 mm. Thus, the average multimodal representation is closer to the haptically-specified width than the visually-specified width. But the exact value of the mean multimodal representation—54 mm—is also important. The visual-only production distribution given width 50 mm and medium noise has mean as of 50 mm and, let's suppose, variance 4. The haptic-only production distribution given 56 mm has mean as of 56 mm and, let's suppose, variance 2. Now take a weighted average of the two means of the single-cue distributions. Let the weights in the weighted average be defined via the variances of the single-cue distributions. In particular, let the weight on the visual mean be 2 / (2+4). That is, the weight on the *visual mean* is the *haptic variance* over the sum of the two variances. And let the weight on the haptic mean be 4/(2+4)—that is, the visual variance over the sum of the two variances. The weighted average, then, is:

$$\frac{2}{2+4}50 + \frac{4}{2+4}56 = 54$$

The mean multimodal representation given visual width 50 mm and haptic width 56 mm can be expressed as a weighted average of the means of the visual-only and haptic-only production distributions given 50 mm and 56 mm respectively. The weights in the average are defined by the variances of the unimodal production distributions, *conditional* on the same noise conditions that characterize the type of multimodal trial. In these multimodal conditions, the visual stimulus was garbled with medium noise. In the vision-only case, that noise level leads to a production variance of 4; the haptic stimulus is generated in the same way as in the haptic-only case, which led to production variance 2. These variances define the weights for that type of multimodal trial.

The first central finding of Ernst and Banks (2002) is that this weighted average relationship holds given any of the tested visual widths, haptic widths, and noise conditions. The mean of the multimodal representation given any triple of visual width, haptic width and noise condition is a weighted average of the means of the corresponding single-modality production distributions. The weights are determined by the variances of the corresponding single-modality production distributions, conditional (in the visual case) on the specific noise condition of the multimodal trial.

The particular weighting scheme entails that the mean multimodal representation is closer to the mean of the modality that has the *lower variance* (equivalently, higher precision) in the corresponding unimodal trials. Return to our example two paragraphs ago. The haptic representation has variance 2. The total variance is 2 + 4 = 6. The haptic variance thus accounts for *less* of the total variance than the visual variance. The fraction 2/6 (haptic variance over total) is applied to the *visual* mean of 50 mm, and thereby pushes the multimodal representation farther from it. Meanwhile, 4/6 (visual variance over total) is applied to the *haptic* mean of 56 mm, and thereby pulls the multimodal representation towards it. As a result, the average multimodal representation is closer to the haptic mean than to the visual mean.

Moreover, the mean multimodal representation is closer to the haptic mean *by a certain amount.* The distance between the mean multimodal and haptic representations is 2 mm; the distance between multimodal and visual is 4 mm. Thus the ratio between the multimodal-visual discrepancy and multimodal-haptic discrepancy is 4 / 2 = 2:1. This is exactly the ratio that the corresponding variances fall into. The ratio between visual variance (given the relevant noise condition) and haptic variance is 4 / 2 = 2:1. So, another way of putting the findings of Ernst and Banks is that the mean multimodal representation always falls between the unimodal means, with the relative distance to each unimodal mean determined by the ratio of the unimodal variances.

The first finding of Ernst and Banks may be illustrated graphically. Consider a multimodal trial type. Suppose the noise condition is n. Suppose the visual input on the multimodal trial is one that leads to a mean visual-only representation $\bar{R}_V$. Suppose that the haptic input is one that leads to a mean haptic-only representation $\bar{R}_H$. Then the relationship between the mean multimodal representation $\bar{R}_M$in these conditions and the single-cue means is depicted in Figure 2.5 below.

The gray line on the left depicts the haptic-only production distribution given the haptic width. We indicate the average haptic-only representation, $\bar{R}_H$. The gray line on the right depicts the visual-only production distribution given the visual width (and the fixed background condition n). We indicate the average visual-only representation $\bar{R}_V$. The unbroken line depicts the multimodal production distribution, given the visual and haptic widths. Its

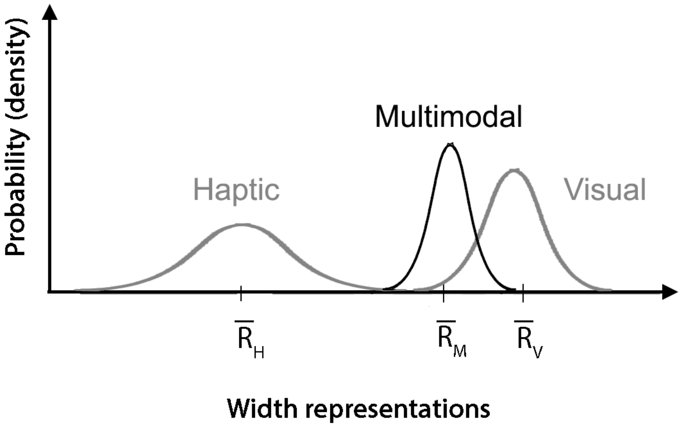


Figure 2.5: Graphical illustration of the main findings of Ernst and Banks (2002). The gray curves are unimodal production distributions. The center curve is the multimodal production distribution. See text for further explanation. Figure is inspired by an explanatory figure from Ernst and Banks (2002).

average is indicated as $\bar{R}_M$. Here, the weighted average relationship can clearly be seen. Since it's a weighted average, $\bar{R}_M$ must fall somewhere in between $\bar{R}_H$ and $\bar{R}_V$. How much closer it is to one or the other depends on the relative variances of the unimodal distributions. In the figure, the haptic variance is much greater than the visual variance. In particular, the haptic variance is about four times as great as the visual variance. Consequently, the multimodal mean is much closer to the visual mean. If the noise condition had been *n'* but we held the other stimulus parameters fixed, and if *n'* induces a visual variance even greater than the haptic's variance, than the multimodal mean would be closer to $\bar{R}_H$ than to $\bar{R}_V$.

Officially, I'll define Finding 1 of the paper is as follows. Let w_V be a visually-specified width, w_H be a haptically-specified width, and let n_i be some noise condition determining the garbling of a visual input. I suppose that $\sigma_V[\cdot]$ is a function from visual width and noise condition to visual production *precision* (the reciprocal of variance) given those parameters, that $\bar{R}_V[\cdot]$ is a function from visual width and noise condition to the mean representation of the visual production distribution given those parameters, and so on. We can then state **Finding 1**: for every visual width w_V, haptic width w_H, and noise condition n_i, the following relationship holds:

$$\bar{R}_M[w_V, w_H, n_i] = \frac{\sigma_V[w_V, n_i]}{\sigma_H[w_H] + \sigma_V[w_V, n_i]}\bar{R}_V[w_V, n_i] + \frac{\sigma_H[w_H]}{\sigma_H[w_H] + \sigma_V[w_V, n_i]}\bar{R}_H[w_H]$$

The second critical finding of the paper centers on the precision of the multimodal production distribution. Visual inspection of Figure 2.5 reveals that multimodal production distribution is narrower than either of the corresponding visual or haptic production distributions. Thus, its variance is less than the variance of either. That is, its precision is greater than either. In fact, its precision can be expressed as a specific function of the precisions of the two production distributions: it is simply their sum. Officially, **Finding 2** of the paper is, for every visual width w_V, haptic width w_H, and noise condition n_i, the following relationship holds:

$$\sigma_R[w_V, w_H, n_i] = \sigma_V[w_V, n_i] + \sigma_H[w_H]$$

This implies that the multimodal precision will always be strictly greater than both the visual and haptic precision (relative to the same noise conditions).

Of course, the actual findings of the paper are not as pristine as my formulations of them in Findings 1 and 2 are. There is a limit to how precisely a production distribution can be estimated, and hence there is some approximation error in estimates of the various means and variances. The findings of the paper suggest that the production distributions bear the cited relations approximately. The empirical results match the findings, as officially stated, quite nicely.

2.3.2 Cue integration in natural analogues

Let us consider again a visual system that operates in a natural analog of the laboratory conditions of Ernst and Banks (2002). Actual bars with actual widths appear, with probabilities in line with the experimental set up, etc. The unimodal production distributions give probabilities that various levels of representational accuracy will be achieved, and the unimodal posterior distributions give probabilities of various levels of representational accuracy for the representation that is in fact produced. Intuitively, these probabilities correspond to different kinds of representational *reliabilities* for the visual and haptic systems, operating independently.

These differences in reliability are, at least given the stimuli of Ernst and Banks (2002), reflected in proximal stimulation in some way. High noise conditions make for high posterior variance; they also tend to produce certain kinds of stereo image-pairs (namely, those with widely varying disparities between corresponding dots). In the single-cue case, the representation produced from such proximal stimulation is quite likely to be significantly inaccurate. Consequently, in a multimodal case, if the system receives that kind of visual proximal stimulation, and if that stimulation conflicts with the haptic stimulation, the system should rely on it less, if at all possible, in the integration process. The closer the multimodal representation is to that kind of stimulation, the closer it is to the mean single-cue representation, and thus the more likely it is to be inaccurate. Finding 1 indicates that the cue integration process has at least some capacity to temper its reliance on visual and haptic inputs along these lines. On average, given such types of visual proximal stimulation, the multimodal representation is weighted away from the width suggested by the stimulation. But of course, it shouldn't downweight its reliance if the haptic stimulation is even worse (that is, it tends even more heavily to produce quite likely quite wrong haptic representations in the single-cue cases). The system appears to effect a reasonable compromise between the visual and haptic inputs along these lines.

In one sense, there are no cue conflicts in a natural setting. No bar can have two different widths simultaneously. Moreover, the bar's width determines optical and tangible aspects of the bar jointly. Thus, any visual cue to width that is directly produced by the bar must agree with the tangible cue it directly produces. But independent sensory information channels may nevertheless often provide conflicting information about an individual property. The result of each information channel is probabilistically determined, relative to various set-up types, by surrounding conditions, both internal and external to the creature. Suppose that, in the multimodal integration process, the system produces two single-modality representations of the width, and combines these in some process. The two single-modality representations may, of course, diverge, even in natural settings. Each process is stochastic. It involves the build up of neural noise, and so on. The transmission of optical or tangible information through a medium like the air or touch may involve corruption. Dynamically changing environments entail that a distal stimulus may fall outside the bounds of a creature's discriminative capacities. Dynamically acting creatures entail that a distal stimulus may only be glanced at or looked over. Relative to any of these conditions, the representations produced may have different variabilities.[11]

Thus, divergence between *token* single-cue representations is to be expected, even if the production and posterior distributions always have overlapping *means*. But the principle remains: if a representation is produced by a process, or in a circumstance, in which it is likely to be substantially inaccurate, it should be relied on relatively less in the cue integration process.

Finding 2 also bears a further significance in the natural analogue. As with the single-cue production distributions, the posterior probability over distal widths is Gaussian given any particular multimodal representation and noise condition. That is, given any multimodal representation and noise condition, there is a Gaussian distribution specifying the probabilities that the distal width is one or another value. Finding 2 implies that the probability of the multimodal representation's veridicality is higher than that of either single-modality

[11] See also Knill (1998; 2003), who points out that slant representations based on texture become biased at higher slants. That is, given a large slant S, the mean of the production distribution for texture-only slant representation will not be as of S, but rather will be as of a slant as much as 10 degrees away. Meanwhile, the mean stereo-only representation given S will be as of S. This constitutes another kind of natural case in which divergences can arise. If the distal slant is S, then most likely the texture cue will diverge significantly from the stereo cue.

representation (relative to the same noise condition). The distribution over widths is narrower given a multimodal representation than given either a visual or haptic representation only. Since all distributions are unbiased, this implies that the multimodal representation is more likely to be veridical. Additionally, the multimodal representation is more likely to be *approximately* veridical. For any fixed standard of ‘approximate veridicality’, the multimodal representation will be approximately veridical with higher probability.

Finding 2 also reveals a further representational benefit that accrues to a multimodal representation. Not only is the specific multimodal representation produced more likely to be veridical; given a particular distal width, Finding 2 implies that a veridical representation of it is more likely to be produced multimodally than within an individual modality.

2.4 A model of the cue integration process

2.4.1 How the findings constrain the cue integration process

Findings 1 and 2 pertain to probabilities *of* representational production. The findings characterize various probabilities of multi-cue representation production in different kinds of circumstances. They relate these probabilities to probabilistic aspects of single-cue perceptual processes operating in relevantly similar circumstances.

What kind of process is operating on each trial of the multimodal experiment? What kind of process determines which multimodal representation will be produced on a given trial? How is that process influenced by the sorts of states produced in vision and touch on a multimodal trial? How do the behavioral findings constrain the answers to these questions?

The first step in answering these questions is to identify the unimodal states that determine the multimodal representation produced. Of course, many unimodal states are causes of the multimodal representation produced. The initial visual registration of the retinal image is *a* cause of the multimodal representation. But we want to identify the unimodal states that most *directly* determine the multimodal representation. We can call these the unimodal *inputs* to the integration process. Once we have identified the unimodal input states, we can then ask which multimodal representations are produced by which combinations of unimodal states. That is, we may ask about the *mapping* from sets of unimodal states to multimodal representations. Having the correct mapping is a prerequisite for any more detailed explanation of the integration process.

In order to specify a mapping, I will make three assumptions about the multimodal integration process. The first is that the mapping is deterministic. Once the unimodal representations have been fixed, the multimodal representation is thereby fixed. This assumption is primarily an idealization to simplify various matters. There is potential stochasticity in any psychological transition, including the unimodal-to-multimodal transition. However, we can plausibly assume that the noise in the transition is relatively negligible.[12]

The second assumption is about the unimodal inputs to the mapping. I assume that, on each multimodal trial, the visual and haptic systems produce their own representations of the bar's width. These unimodal representations are (or are among) the unimodal inputs to the integration process on that trial. Thus, such representational states are arguments to the mapping.

This is a substantive assumption. To see that it is substantive, consider the correlative assumption for *within*-modality cue integration. A visual representation of a surface's slant may be formed from different kinds of sensory information. A slant representation can be produced given only information about stereo disparity. A slant representation can also be produced given only monocular information about texture gradients. A slant representation can be formed when both kinds of information are available. It is, of course, possible

[12] Note also that the Findings rule out various stochastic mappings. For instance, one might propose a mapping whereby one of the two unimodal representations is chosen on each trial to determine the multimodal representation entirely. There might be different probabilities of choosing each unimodal representation, however. Suppose the visual representation is chosen with probability $p_v/p_v + p_h$ and the haptic representation is chosen with $p_h/p_v + p_h$ (where p_i is the precision of modality i's production distribution). Then Finding 1 is predicted. However, such a mapping conflicts with Finding 2. See Rohde et al. (2016).

that, when both kinds of information are available, the visual system produces three slant representations: two based exclusively on a single source of information, and a third multicue slant representation derived from these two. But another possibility is that the system simply forms a single multicue slant representation from the two sensory sources of information. A similar option is available for integration *across* modalities. The two modalities might not produce individual representations, but rather pass forward other kinds of sensory information that serve as the inputs to the multimodal integrator.

Nevertheless, I take the assumption to be empirically warranted for the multimodal cue case. Consider, for example, Hillis et al. (2002). Subjects performed a discrimination task in the same virtual reality set-up as in Ernst and Banks (2002). As before, the stimuli were bars presented both visually and haptically. A bar could be *standard*, indicating that the visual and haptic cues both suggested the same width. Alternatively, a bar could be *conflicted*, in which the visual and haptic stimuli suggest discrepant widths. The task for subjects was an "oddity" task. On a trial, three bars are presented. Either two bars were standard and one conflicted, or two were conflicted and one standard.[13] Subjects had to indicate which of the three bars was the "odd one out"—i.e. the one that seemed different than the others.

Three hypotheses are available about how the oddity task is performed. One hypothesis is that no multimodal representation is produced at all. Subjects perform the oddity experiment only by producing two unimodal width representations per bar, and choosing "odd" if either exceeds some threshold. A second hypothesis is that subjects use *only* a multimodal representation (produced in roughly the way it was in Ernst and Banks (2002)). Subjects perform the oddity experiment only by producing a multimodal representation for each bar, and choosing a bar as "odd" if it exceeds the other two representations by some amount. A third hypothesis is that subjects produce and use all three representations. If any one of them is sufficiently "deviant" from the representations for the other bars, then they

[13] If the standard presented in a trial had width S, then the conflicted stimulus had visually-specified width $S + \Delta S_v$ and the haptically-specified width $S + \Delta S_h$. The values for ΔS_v and ΔS_h varied across the experiment. If two standards were presented on a trial, they both had width S.

will signal that bar as odd. These different hypotheses make different predictions about the results of the experiment. The results of the experiment were, in fact, consistent with the third hypothesis, that subjects perform the oddity task by accessing one of three available width representations.

The third assumption pertains to the production distributions of the visual and haptic representations *under multimodal conditions.* I'll assume that the production distributions for the unimodal representations are identical in multimodal conditions to the corresponding distributions in *unimodal* conditions. For example, I'll assume that the production distribution for visual representations of width given distal width 50 mm and noise condition n is identical in both the unimodal and multimodal case. This assumption is tantamount to assuming that, in the multimodal trials, the processes producing visual representations are not substantially affected by the processing going on in the haptic system, and vice versa. This assumption is questionable; we know that there is cross-modal influence in perception under certain circumstances. Nonetheless, I follow Ernst and Banks in assuming independence of the visual and haptic processes *under* the conditions of the experiment. Ernst and Banks point out that, given the production distributions, Findings 1 and 2 imply that multimodal representations are produced in a statistically optimal way. The sense of "statistically optimal" will be discussed in detail below. However, the optimality assumes that the productions of visual and haptic states—including unimodal representations—are conditionally independent given distal width. I take this to provide support for the assumption of independence. Generally and defeasibly speaking, we can expect sensory systems to be driven towards optimality by evolutionary and developmental factors. The optimality of the process assuming a condition about the sensory processes is thereby reason to believe the condition.

Given these two assumptions, what can we say about the trial-by-trial processes that lead to a multimodal representation?

First, the findings imply that the unimodal representations cannot be the *only* unimodal inputs. Suppose that they were. That is, suppose that the unimodal representations produced on a trial fully determine the multimodal representation produced. The same pair of unimodal representations would produce the same multimodal representation regardless of noise condition. The observed differences in multimodal production distributions between visual noise conditions would thus have to be explained entirely on the basis of the different visual production distributions—and in particular, on the basis their differing precisions.

Consider, for example, one kind of trial in which the haptically-specified width is 45 mm, the visually-specified width is 55 mm, and the visual and haptic production distributions have equal precision. In this case, per Finding 1, the average multimodal width representation will be 50 mm. The precision of the multimodal production distribution is greater than that of either unimodal production distribution. Call this the (A) trial type. The three production distributions of relevance in (A) are depicted in Figure 2.6A below.

Now consider a similar trial type (B). This trial type is identical to (A), except that the visual noise is greater, and so the visual production precision is less than the haptic production precision. Again, per Finding 1, the average multimodal representation is closer to the average haptic representation. The production distributions are depicted in Figure 2.6B.

How can we explain this difference if the unimodal representations fully determine the multimodal representation? The only relevant difference between the two cases is that, in (B), visual representations near 55 mm are less likely than they were in (A), and visual representations sufficiently farther away from 55 mm are more likely than they were in (A). Consider the point on the x-axis corresponding to a representation 52 mm in Figures 2.6A and 2.6B. In (A), that representation has almost no probability of occurring, but in (B), it has a greater probability. To explain the shift in multimodal mean from (A) to (B), then, the representations that become more likely in (B) must, under a fixed mapping, somehow push the multimodal representation towards the haptic representation. For instance, the mapping might assign visual representation 52 mm and various haptic representations r_h to

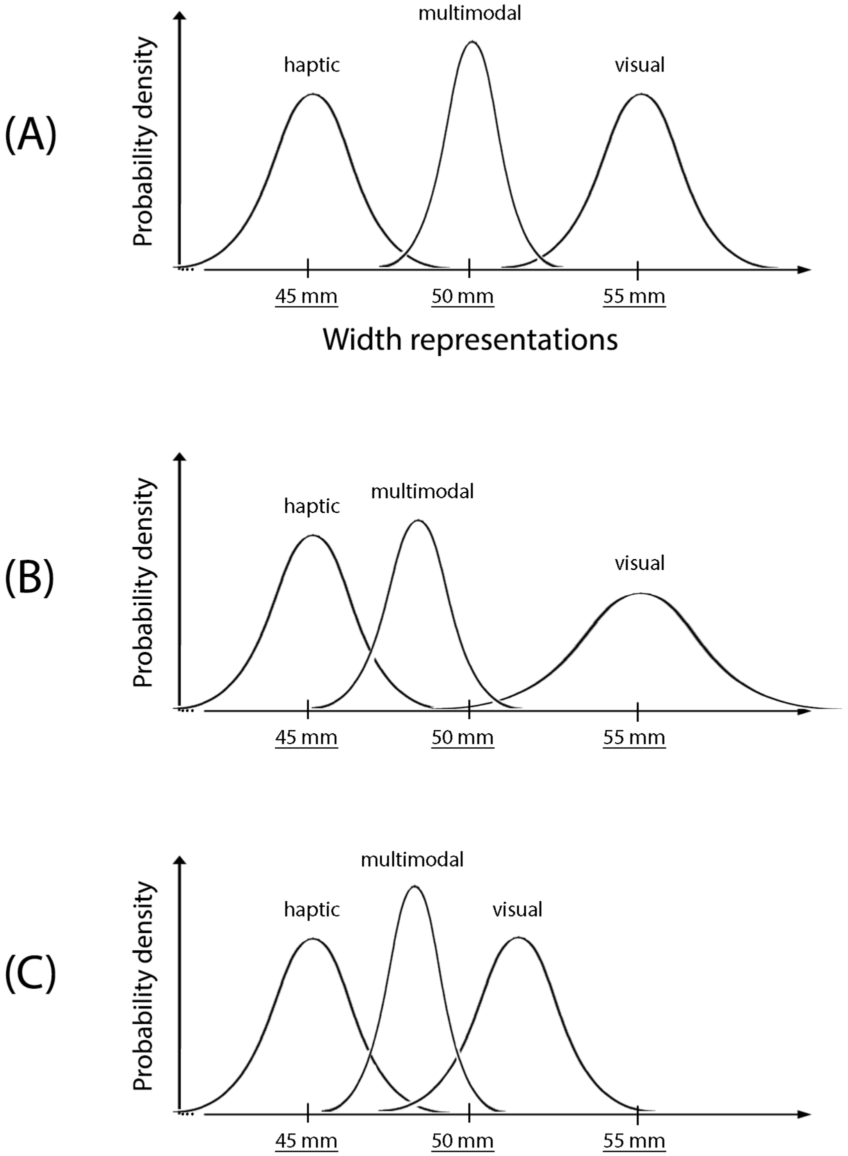


Figure 2.6: Unimodal and multimodal production distributions in three different types of multimodal trial. Each unimodal production distribution is conditional on the distal width represented by its average representation. (A) Visual and haptic distributions are equal variance. (B) The visual variance is greater than in (A). The multimodal variance is correspondingly larger than it is in (A). (C) The visual and haptic distributions have equal variance, but the mean of the visual distribution is shifted leftward compared to (A).

a multimodal representation that is quite close to r_h. The visual attributive $\underline{\text{52 mm}}$ is on the "tail" of the (B) distribution. If sufficiently many visual "tail" attributives lead, under a fixed mapping, to multimodal representations close to r_h, that would push the average multimodal representation towards the mean haptic representation. In principle, such a mapping might explain the Findings as reflected in (A) and (B).

One way to implement such a mapping would be to diminish the influence of a visual representation on the multimodal representation the farther that representation is from the mean of the relevant visual production distribution. On such a model, the visual representation $\underline{\text{52 mm}}$ together with r_h maps to a multimodal representation close to r_h because $\underline{\text{52 mm}}$ is fairly far from the mean of the production distribution from which it was drawn—that is, from $\underline{\text{55 mm}}$. This proposal is implausible, however. The system has no means for estimating the mean of the production distribution on a trial *separately* from the visual representation itself. For another thing, this proposal already conflicts with the supposition that the unimodal representations are the *only* unimodal inputs to the integration process. The proposal requires that some sort of information constituting an estimate of the visual production mean combines with the visual representation to help determine the multimodal representation. The estimate of the mean would then also be a unimodal input.

For a fixed mapping, then, to explain the multimodal shift, the visual representations that become relatively more likely in (B) must exert a constant "push" on the multimodal representation away from the visual representation. The constant may differ for the different "tail" representations of the (B) distribution. Perhaps a visual representation $\underline{\text{51.5 mm}}$ pushes the multimodal representation farther than $\underline{\text{52 mm}}$ does.

However, consider a third trial type (C), illustrated in Figure 2.6C. Here, the haptically-specified height is 45 mm, the visually specified height is 52 mm, and the unimodal production distributions have equal variance. Per the Findings, the average multimodal representation will be half-way between the average haptic and visual representations—that is, it will be $\underline{\text{48 mm}}$. But under these circumstances, the visual representation $\underline{\text{52 mm}}$ is now the most likely visual representation to occur. In order to explain the shift from (A) to (B) in multimodal mean, we stipulated that the visual attributive 52 mm, together with haptic representations r_h that are likely to occur given haptic width 45 mm, lead to multimodal representations close to r_h. In (C), the haptic situation has not changed. Consequently, the constant "push" that 52 mm and nearby representations exert on the multimodal representation towards the haptic entails that the multimodal mean should be fairly close to 45 mm. This conflicts with Finding 2.

I conclude that the unimodal representations cannot be the only unimodal inputs to multimodal integration. More generally, I conclude that a given pair of unimodal representations must, in general, produce different multimodal representations in different noise conditions. Thus, the multimodal integrator requires inputs that effectively signal which noise condition is currently instantiated.

More precisely, suppose that the visual noise conditions are $n_1, \ldots, n_k$. I claim there are sensory states $v_1, \ldots, v_k$, such that state v_i correlates strongly with visual noise condition n_i. That is, I assume that the probability that v_i occurs given that the noise condition is n_i is very high, and conversely, given state v_i, the noise condition is n_i with high probability. I'll call each v_i state a **noise-tracking state**, understood simply to reflect the fact that the state correlates with a specific noise condition. Similarly, I assume that there are haptic noise-tracking states $t_1, \ldots, t_l$ (where we assume a background of l-many haptic noise conditions).

The multimodal integration process thus takes four unimodal inputs: two unimodal representations and two unimodal noise-tracking states. What can we say about the mapping from each possible quadruple of unimodal input states to multimodal representations? The obvious mapping is one that mirrors Finding 1 directly. Each visual noise condition n_i is uniquely associated with a magnitude p_i. The magnitude p_i characterizes the precision of the relevant production and representational posterior distributions that will arise given n_i. A similar magnitude p'_j is associated with haptic noise condition j. Consequently, we may define our mapping so that, if the unimodal inputs are r_v, r_h, v_i and t_j, then the multimodal representation produced will be equal to the weighted average $p_i/(p_i + p'_j)r_v + p'_j/p_i + p'_j)r_h$. Supposing that the noise-tracking states correlate with noise condition very strongly, such a mapping predicts both Findings precisely.

Here we assumed that a certain magnitude — a weight— characterizes the causal impact of noise-tracking states v_i and t_j on the multimodal mapping. We assumed, further, that this magnitude was equal to the magnitude of a probabilistic quantity associated with the relevant noise condition. The causal and probabilistic magnitudes need not match exactly. We may assume, for instance, that the visual noise-tracking state v_i contributes a magnitude equal to $p_i + \varepsilon_1$ and the haptic contributes a magnitude equal to $p'_j + \varepsilon_2$. If ϵ_1 and ϵ_2 are very small, then we predict the Findings approximately, though not exactly. However, the causal magnitudes cannot diverge too far from the probabilistic magnitudes. Suppose the visual standard deviation is 4 mm and the haptic standard deviation is 2 mm. Suppose the causal magnitude associated with v_i is $p_i + 0.2$. That is, the causal magnitude is within 5% of the relevant probabilistic magnitude. In such a case, however, the multimodal production distribution would diverge appreciably from those predicted by the Findings. Where the multimodal average should be 47 mm with standard deviation 1.79, the multimodal average would be 50 mm with standard deviation 2.27.

2.4.2 A basic model of cue integration

Each visual noise-tracking state v_i is canonically associated with noise condition n_i. The noise condition n_i is canonically associated with magnitude p_v^i. That magnitude is the magnitude of the precision of any visual production or posterior distribution that is conditional on n_i. Transitively, each visual noise-tracking state v_i is canonically associated with the magnitude p_v^i. Similarly, a haptic noise-tracking state t_j is canonically associated with a magnitude p_h^j.

Suppose that visual noise-tracking state v_i is produced on a trial. Let p_v be the magnitude canonically associated with v_i. Similarly, suppose haptic noise-tracking state t_j is produced on a trial, and let p_h be the magnitude associated with t_j. I will put double-brackets around a term denoting a representational type to denote the property that the type *indicates* and functions to *attribute* to entities. So, if $[\![r_v]\!] = w$, the visual representational type r_v indicates the width property w. Suppose the unimodal representations produced are r_v and r_h. Then, following the previous section, according to our basic model of cue integration, the multimodal representation r_m that is produced will satisfy:

$$[\![r_m]\!] = \frac{p_v}{p_v + p_h}[\![r_v]\!] + \frac{p_h}{p_v + p_h}[\![r_h]\!]$$

Under this mapping, each visual noise-tracking state v_i has a specific kind of effect on multimodal representation production. When the perceptual system is in state v_i, the resulting multimodal representation will bear a specific quantitative relation to the unimodal representations that are produced at the time. The multimodal representation will be as of a width that is a weighted average of the widths that are the indicants of the unimodal representations. Whenever v_i occurs, the multimodal width indicated will always be expressible as an average in which the weight on the visually indicated width is a quantity that is determined partly by p_v^i—the precision magnitude associated with v_i.

2.4.3 The causal properties of noise-tracking states

Of course, knowing just the visual noise-tracking state produced on an occasion, we can predict nothing specifically about the multimodal representation. The visual noise-tracker is one of four variables that jointly determine the multimodal representation. We can, however, consider commonalities in the effects of a visual noise-tracker by looking for commonalities entailed by the mapping when holding various of the other three variables fixed.

The noise-tracking states in each modality jointly control the level of "compromise" that the multimodal representation strikes between the two unimodal representations. Different uses of a noise-tracker v_i may yield a common effect on the level of compromise. A measure of compromise must compare the proximity of the multimodal representation to the visual representation with its proximity to the haptic representation.

One measure of the level of compromise is

$$\frac{r_m - r_v}{r_m - r_v} = \frac{p_h}{p_v}$$

This measure is defined as the ratio between the multimodal-visual distance and the multimodal -haptic distance. When the ratio is 1, r_m is halfway between the unimodal representations. (For example, if the unimodal representations are as of 40 mm and 50 mm, a multimodal representation as of 45 mm is halfway between the two.) Because 1 is the point of equal compromise, let us call this measure the **unity measure**. When the measure is greater than 1, the visual-multimodal distance exceeds the haptic-multimodal distance, and so the haptic representation is favored. The degree to which it is favored is measured by the magnitude of the measure. When the measure is less than 1, the multimodal representation is closer to the visual representation, and the visual representation is favored. Given the current probabilistic set-up, the unity measure is equal to the ratio between the relevant precision magnitudes. Suppose that the haptic precision p_h is 2, and p_v is 1. Here, the relevant haptic distributions are more precise, hence less variable, than any of the relevant visual distributions. The measure then has the value $2/1 = 2$, and indeed, the visual-multimodal distance will be twice the haptic-multimodal distance. (For example, if $r_v = 55$ and $r_h = 40$, then $r_m = 45$) and our measure is $10/5 = 2/1 = 2$.) The unity measure may be positive or negative depending on whether or not the haptic representation is greater than the visual representation, but the sign of the measure will not be important.

A second measure of the level of compromise is

$$\frac{r_m - r_h}{r_v - r_h} = \frac{p_v}{p_v + p_h}$$

This measure is defined as the ratio between the haptic-multimodal distance and the visual-haptic distance. The measure lies between 0 and 1. Near 0, the multimodal representation is very close to the haptic representation; near 1, it is very close to the visual representation; and at 1/2, it is exactly halfway between them. Call this the **visual half measure**, since the point of equal compromise is 1/2 according to it.

Knowing just the visual noise-tracker that has been produced, we cannot predict anything specific about the level of compromise that will be effected in the multimodal representation. The level of compromise (on either the unity or half measures) does not depend on the unimodal representations, but it does depend on the haptic noise-tracker as well. Idealizing for the moment that unimodal noise-trackers form a continuous suite of states, given a visual noise-tracker v_i, for any level of compromise c, there is a haptic noise-tracker that, if produced, would (together with v_i) lead to a multimodal representation bearing that level of compromise.

Nevertheless, the visual noise-tracker v_i does, by itself, constrain the ways in which specific levels of compromise may be reached. It fixes, for example, a minimum degree that the haptic noise-tracker must exceed, if the level of compromise is to exceed a certain value. Consider the visual half-measure. Suppose that the visual precision is 1/9 (corresponding to a standard deviation of 3 mm), and that this corresponds to visual noise-tracker v_i. For a level of compromise to exceed 0.5 (via the visual half-measure), the haptic precision must be greater than roughly 0.1. See Figure 2.7. Additionally, the visual noise-tracker determines how rapidly the level of compromise will change as the haptic noise-tracker changes. The relationship is illustrated in Figure 2.7.

Suppose the haptic noise-tracker is associated with precision 0.5. If the visual noise-tracker is v_i, then slight changes in the haptic noise-tracker lead to relatively large changes in the degree of multimodal compromise. Notice that the curve in Figure 2.7 above 0.5 accelerates relatively rapidly compared to other haptic precision points.

Now suppose, instead, that the visual noise-tracker is v_j corresponding to a precision of 1. v_j correlates with a situation in which the visual production distribution is less noisy, and so more precise, than v_i. Consider Figure 2.8.

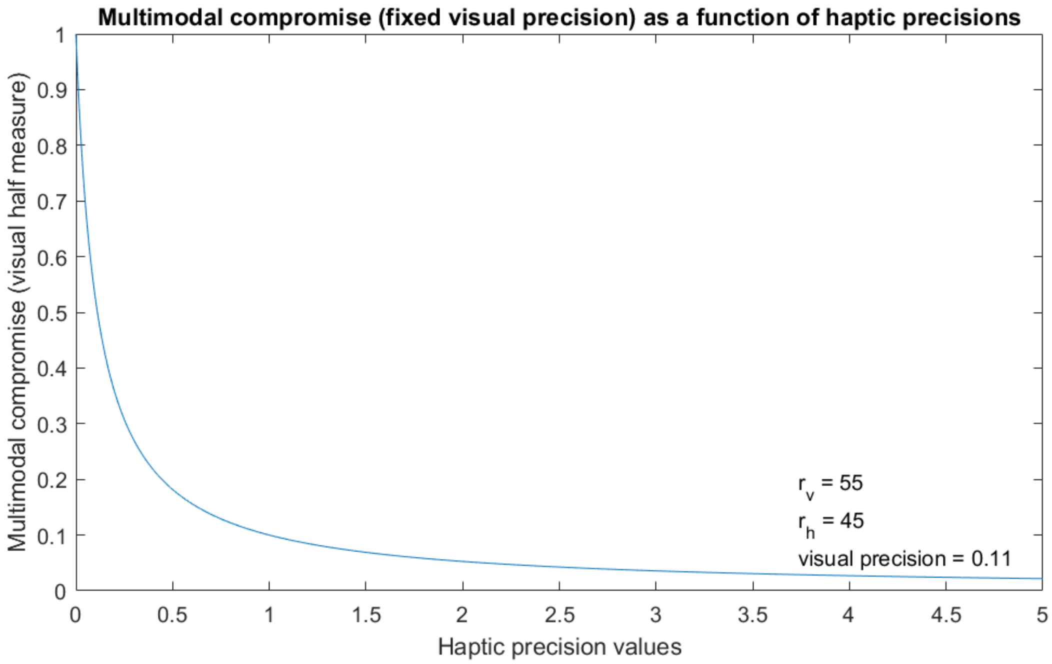


Figure 2.7: Multimodal compromise as a function of haptic precision, given visual precision 0.11.

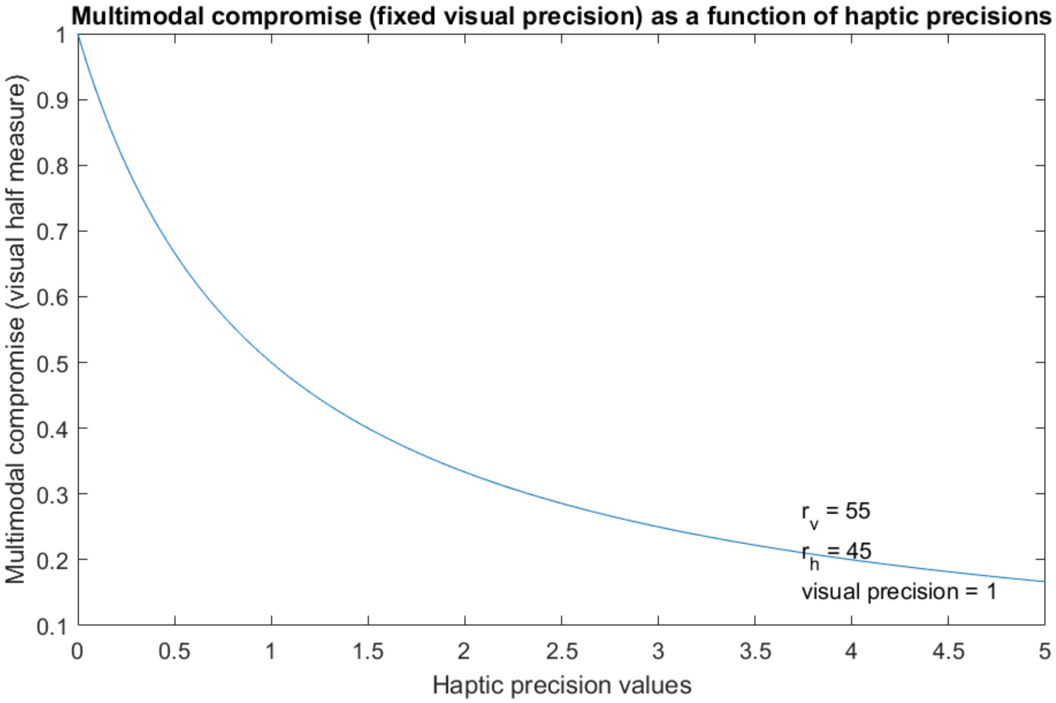


Figure 2.8: Multimodal compromise as a function of haptic precision, given visual precision 1.

Compared to v_i, v_j puts a more demanding constraint on the haptic noise-tracker if a given level of compromise is to be reached. To reach a level of 0.5 multimodal compromise, the haptic noise-tracker must now be associated with a precision of 1—an order of magnitude than was required of it under v_i to reach the same level of compromise. Moreover, slight changes in haptic noise-tracker lead to less drastic shifts in multimodal compromise. If the haptic precision varies around 0.5, the multimodal compromise is less affected in the presence of v_j than in the presence of v_i. Notice the shallower bend in Figure 2.8.

Thus, instances of v_i are alike in the types of effects they entrain on the level of compromise in multimodal representation. The similarities across instances of v_i are systematically related to the dispersion magnitude—precision—that the trackers correlate with, as discussed.

I will summarize the causal properties of a noise-tracking state by saying that the state determines a **level of pull**. The level of pull of noise-tracker v_i may be measured by its associated precision, p_v^i.[14] The level of pull of a noise-tracker characterizes the effect that the tracker has to bring the multimodal representation in line with (or "pull it towards") the visual representation. The level of pull summarizes a pattern of commonalities in the uses of the noise-tracker, and also indicates the dimension along which it differs causally from other noise-trackers in the same modality.

I am conceiving of level of pull as a causal variable somewhat like gravitational force. The gravitational force exerted by a body X on a particle does not, by itself, have implications for how the particle will move. The particle's acceleration will be determined by the joint influence of all bodies exerting gravitational influence on it at any one time. Any degree of acceleration is consistent with the gravitational force exerted by X—just add the right bodies with the right masses and distances. Nevertheless, the gravitational force from X

[14] I do not assume that the level of pull is identical to the precision. Rather, I assume that levels of pull form a ratio scale, with the property that v_i's level of pull is k times as great as v_j's level of pull if and only if p_v^i is k times as great as p_v^j. This identifies levels of pull as a set of psychologically explanatory magnitudes up to a multiplicative factor.

summarizes a common contribution that X makes to acceleration. Such contribution can be seen by studying patterns in acceleration as other attractors are changed while X is held fixed, and also as X's force is altered while holding the other attractors fixed.

Similarly, although the level of pull exerted by the noise-tracker does not itself determine the multimodal representation, it does mark a commonality in the noise-tracker's contributions to the multimodal representation. It marks a pattern in how levels of compromise change across different haptic noise-trackers but given v_i. It marks a pattern in how levels of compromise are fixed across changing representations but holding fixed the haptic and visual noise-trackers. And it marks a difference in the contributions between v_i and other visual noise-trackers. In light of its use in marking these patterns of v_i's causal influence, I will assume that exerting or enacting a certain level of pull is a real effect of v_i.

2.4.4 Noise-tracking states as kinds of modes of presentation

Each variance tracking state v_i is assumed to correlate heavily with noise condition n_i. v_i renders n_i very likely, and vice versa. But the correlation cannot hold magically. The correlation must be induced by sensory means. The correlation must hold because there is some kind of sensorily available information that tends to be caused in and only in n_i, which information in turn tends to cause and only cause v_i. This raises the question of the type of sensorily available information from which the noise-tracking states are produced.

One possibility is that the noise-tracking states are produced from sensory information separate from the kinds of sensory information that produce the unimodal representations. I mention two examples.

The first example is a *sampling model.* On the sampling model, we assume that, on a single trial, *many* unimodal width representations are rapidly produced. Further, we assume that each representation is produced via the same production distribution. Thus, the set of produced representations constitutes a *sample* from that distribution. The system could then compute a magnitude this sample, using standard methods of statistical inference. It might compute, for instance, a *sample variance*, the magnitude of which is used directly in the subsequent weighted average.

The second example is a *learned cue model.* The system might learn that a visual cue naturally unrelated to the computation of width representations signals a level of variability in width signals. We can imagine that each noise condition is associated with a different colored light. A light of that color shines if and only if its associated noise condition obtains. We can imagine that, via a process of perceptual learning, the system comes to recognize a long-term correlation between light color and variability of width representations. At the end of this learning period, it comes to produce a distinct noise-tracking state in response to the different colors. The noise-tracking states then operate as before in guiding the visual representations' influence on the multimodal representation.

Neither model seems likely, at least as applied to Ernst and Banks (2002). The sampling model faces challenges in explaining the psychophysical results. If multiple width representations are produced, how is the single unimodal width percept determined? A plausible rule is that the average width representation in the sample determines the final width percept. But if the number of samples is enough to establish a reliance variance estimate, it will also be enough to establish a reliable estimate of the distal width. That is, the variability in the final percept would be quite small; smaller than what we observe in the psychophysical results.[15] The learned cue model may have applications, but there is no separate cue *to* be learned in the set-up of Ernst and Banks.

I shall assume that the noise-tracking state on a trial is determined by the same kinds of

[15] It follows from a well-known result that, if one uses the sample mean $\bar{R}$ of a set of n produced representations to estimate the mean of the distribution from which the representations were produced, then the variance of this estimate of the mean is σ^2/n, where σ^2 is the variance of the underlying production distribution. For appreciable n, this estimator variance is quite different from the production distribution variance. But if the estimator *is* the representation of width, then its variance should be the sample variance. That said, I do not mean to discount the applicability of sampling-based models in general. For instance, various kinds of popular *diffusion* models in perceptual can be understood as a time-series of samples from a distribution. See, for example, Gold and Shadlen (2007). I leave analysis of such models for future research.

information from which distal width representations are produced. In the context of Ernst and Banks, we may assume that the width representations are produced from representations of the various dots at various distances. We can imagine that surfaces are interpolated between the dots by the system, and that the final surface interpolated automatically involves representation of width.[16] In different noise conditions, the representations of dots will vary in depth around the interpolated surface by different amounts. In a no-noise condition, the dots will all be represented as coplanar with one of two interpolated surfaces (ignoring slight inaccuracies in the representations of the dot-depths themselves). In a high-noise condition, the average displacement that a dot is represented to have from the interpolated surface will be quite large. These average displacements correlate very strongly with noise condition.[17] The noise-tracking states could be differentially produced by different average displacements in dot representations.

In principle, the noise-tracking states could be caused separately from the unimodal representations. However, it will be edifying for us to assume that the noise-tracking states are characteristics or types of width attributives that are individuated by *modes of presentation*. Heretofore, we have been speaking of (for example) the production distribution over width representations given a distal width and noise condition. The production distribution gives the probability that the representation produced is *as of* some distal width, for each distal width. In the terminology of Burge (2010), it gives, for each distal width, the probability that the representation produced *indicates* that distal width and *attributes it* to the surface. But the class of width representations indicating a distal width w is more finely individuated.

[16] See Grimson (1981) for a development of Marr's theory of stereopsis on random dot stereograms which involves a critical phase of surface interpolation.

[17] In a system that perfectly interpolates a surface at the correct depth, and one that is noiseless, the average displacement of dot representations will almost certainly be very nearly approximately equal to the noise condition (where noise conditions are understood as the parameters that govern the distribution from which individual dot displacements are drawn on a trial).There are many dots, each of whose displacement is an sample from the distribution governing displacement. The set of dot displacements is an independent and identically distributed sample. Thus, the sample average of displacement is almost sure to be equal to the average of the displacement distribution—that is, to the parameter that individuates noise conditions.

The class comprises a wide set of *attributives* that all indicate the same distal width w but do so under different modes of presentation. Plausibly, a width attributive that is systematically produced from representations of dots varying widely in depth will have a different mode of presentation from a width attributive produced from representations of dots that vary only slightly in depth.

On the current proposal, a noise-tracking state v_i is a *type* of width attributive. On this proposal, the formation law posited by the basic model can be seen as a mapping from a space of pairs of visual and haptic width attributives—now individuated by modes of presentation—to a multimodal representation. On this interpretation, if the visual width attributive produced on a trial is as of w and is of type v_i, then the multimodal representation will be as of a width that is a weighted average of w and the width attributed by the haptic system, with weights determined partly by the visual magnitude σ_i associated with v_i.

How finely grained is the space of visual width attributives that all indicate the same distal width? Burge (2014) suggests that, in general, the space of modes of presentation is extremely finely grained. Attributives indicating single color-shade differ in mode of presentation depending on the specific luminance and illumination cues that they are produced from. Attributives for motion, produced from broadly similar kinds of proximal retinal motion, can differ in mode of presentation if the proximal stimulation that produces them is foveal or peripheral. I take it that attributives can differ by mode of presentation if they are produced under differing conditions of blur, contrast, lighting level, duration of glance, fog, distance, and so on. I take it they can differ generally on the basis of the kinds of information they use. Two attributives for the same slant differ in mode of presentation if one is produced from texture compression cues and the other from information about disparity. I take it that attributives produced in conjunctions of these conditions make for further unique modes of presentation. For instance, I assume that a shape attributive produced in good light but with blur, a shape attributive produced in bad light with no blur, and a shape attributive produced in bad light with blur all have different modes of presentation, even while indicating the same shape.

One of Burge's motivations for typing perceptual representations with this level of grain is to account for the different explanatory potentials that representations as of the same property have in different kinds of circumstances. A representation as of a color-shade may take longer to produce and/or be less likely accurate if it is produced from one pair of luminant-illuminant cues than another. This difference is an explanatory lawlike difference between the two kinds of color-shade representations. Since attributives are the basic kinds that lawlike psychological generalizations subsume, the attributives differ by type. Considerations analogous to Frege's puzzle motivate ascribing this difference to the representational kinds specifically.

I will assume that for each distal width w the class of visual width representations as of w comprises attributives typed in this extremely fine way. To a first approximation, there are different mode of presentation of a width representation corresponding to differences in such factors as the distance to the surface to which width is attributed, the direction of the surface from the viewer, the type of proximal cues (e.g. dot disparities vs. some other sort of texture) from which the width attributive is produced, the level of lighting, the level of blur, the contrast in the retinal image, and so on. I will also assume that, in the unimodal cases of Ernst and Banks, each different *specific* representation of dots in depth corresponds to a different mode of presentation.[18] This assumption will not be strictly necessary, but I would like to err on the side of pluralism rather than stinginess about modes of presentation.

One can think of the mode of presentation of an attributive as of w as being specifiable

[18] To a first approximation, suppose that, given a distal width w, the set of attributives as of w that might be produced on a trial of Ernst and Banks (2002) is determined in the following way. Suppose that, at the finest level of spatial acuity with which the system can discriminate surface points, the system represents 1,000 locations on (or along the line of sight passing through points on) the purported distal surface. The system represents each location as having either a dot or no dot. Furthermore, each dot can be represented as having one of 20 values of depth (considering only those depths that have any appreciable probability of occurring in the experiment). Then the set of width attributives corresponding to w is approximately $2^{1,000} \cdot 20 \cdot 1,000$— an astronomical number indeed. Not all of these possible configurations is necessarily associated with a width w representation, and presumably this cuts much more finely than the lawlike differences between width attributives do. But I am erring on the side of profligacy here.

by a set of specific values for different parameters corresponding to different *ways* that modes of presentation can differ. One parameter might correspond to blur, and its values might constitute different levels of registration of blur that make for lawlike differences in the psychological role of a w attributive. A second parameter might correspond to lighting condition, with values constituting different levels of registered lighting condition that impact the role of a w attributive. And so on. Thus, we can assume that that the width attributives produced in Ernst and Banks (2002) all have identical values for every parameter other than the one corresponding to dot-depth. The attributives produced do not vary according to lighting condition, distance to surface, or blur, since these conditions are essentially held fixed. They vary only according to the dot-depth representations from which they are formed. I'll call the class of attributives that vary in this way the *producible* width attributives (in the context of the experiment).

The noise-tracking state v_i groups or type-identifies a subset of the producible width attributives. It does so by grouping those attributives which have the same kind of effect on the multimodal integration process. A width attributive a is of the noise-tracking type v_i if it leads to a multimodal representation whose representatum's proximity to the indicant of a is partly determined by σ_i. Grouping width attributives by noise-tracking type cross-cuts grouping them by indicant. Two width attributives differing by indicant may both be of the same noise-tracking type. Two width attributives with the same indicant may be of different noise-tracking types.

2.4.5 Why aren't modes of presentation associated with more fine-grained conditional distributions?

Any width attributive a that is of noise-tracking type v_i will produce a multimodal representation that bears a systematic relation to the indicant of a and the probabilistic precision magnitude p_i. One may wonder, however, why the maximally specific attributive a should lead to a discounting of its indicant on the multimodal representation by the specific value

p_i. Precision p_i is (for example) the precision of the posterior $\mathbb{D}[W \mid R = r, N = n_i]$, where a is an attributive of type r. This distribution specifies the relative probabilities of different distal widths, given that the representation is of type r and the noise condition is n_i. Although an instance of a is an instance of r, it is not the only one possible. Other attributives a', a'', and so on are also instances of r. Attributive a specifies not only an indicant but also a particular specific mode of representing the indicant. In principle, this further specificity might constitute further information that impacts the probability of width representations. That is, in principle, $\mathbb{D}[W \mid a, N = n_i]$ might be very different from $\mathbb{D}[W \mid R = r, N = n_i]$. If it is, then why, according to the basic model, does a lead to a discounting of r on the multimodal representation by the variance/precision of the latter distribution and not the former?

Plausibly, however, the specific attributive a provides no further information about distal width beyond the information it provides about noise condition and the information its having a specific indicant provides. More precisely, assume that $a_1, \ldots, a_m$ are all and only those width attributives that can be produced within the Ernst and Banks set-up that are of noise-tracking type v_i. That is, assume that v_i is coextensive with $a_1 \vee \cdots \vee a_m$. Further, make two assumptions. First, for any attributive a_j that is of variance-tracking type v_i, suppose that $P(n_i \mid a_j) = 0.95$. That is, attributive a_j makes noise condition n_i very likely. P(n$_i$ | a$_j$) = 0.95. It then follows (by definition of conditional probability) that $P(n_i \mid v_i) = 0.95$, as we have assumed.

Second, we assume that the distal width impacts the probabilities of attributives only by impacting the probabilities that the produced representation indicates one width or another. That is, the probability of an attributive a, given distal width w, noise condition n_i, and given that the produced attributive indicates width w_a, is equal to the probability of a given n_i and w_a (for all distal widths w). Intuitively, this condition says that holding fixed the indicant and noise condition but varying the distal width does not change the probability of any specific attributive a that has that fixed indicant. Put another way, given that a representation with indicant w_a has occurred, the probability of each specific attributive for w_a is determined solely by the noise condition. Suppose further that this distribution is uniform. That is, $\mathbb{D}[a \mid W = w, n_i, [\![a]\!] = w_a]$ is identical to $\mathbb{D}[a \mid n_i, [\![a]\!] = w_a]$, and both are uniform distributions over the relevant subset of attributives that are compatible with n_i and indicate w_a.

Given these two assumptions, it follows that for any attributive a, the posterior distribution $\mathbb{D}[W \mid a]$ is approximately Gaussian with precision p_i (where, recall, we have assumed that a is of noise-tracking type v_i). The approximation is better the closer $P(n_i \mid a)$ is to 1. Additionally, the distribution $\mathbb{D}[W \mid a, n_i]$ is exactly Gaussian with precision p_i.[19]

2.4.6 Summary

According to the basic model, there are visual noise-tracking states that correlate with a noise condition and that make a systematic causal contribution to multimodal integration. In correlating with a noise condition, a noise-tracking state also correlates with various instantial probability properties. Furthermore, the systematic quantitative effect of the noise-tracking state can be explained via these probability properties. In the next section, I want argue that, in virtue of these relations to probabilistic magnitudes, noise-tracking states in fact *signal* the presence of those probability magnitudes.

[19] For the second claim, note that $W \mid a, n_i$ is Gaussian with variance σ_i if the likelihood function $\mathcal{L}(w; a, n_i) = P(a \mid w, n_i)$ is proportional to a Gaussian with variance σ_i. (Recall that $P(w|n_i) = P(w) = 1/c$ for a constant c.) By the law of total probability, $P(a \mid w, n_i) = \int P(a \mid w, n_i, r)P(r \mid w, n_i)dr$. Let r_a be the indicant-individuated type of a. $P(a \mid w, n_i, r)$ equals 0 unless r is r_a. So, $P(a \mid w, n_i) = P(a \mid w, n_i, r_a)P(r_a \mid w, n_i)$. The former term in the product is constant in w, by assumption, and the latter term is a Gaussian function of w with variance σ_i. Thus, $W \mid a, n_i$ is Gaussian with variance σ_i. For the first claim, note that $P(n_i \mid a) = 0.95$ entails that $P(w \mid a)$ is approximately equal to $P(w|a, n_i)$, which we just showed to be a Gaussian distribution in w with variance σ_i.

2.5 Sensory signaling states

My main question in this chapter is: do sensory systems have capacities or sub-systems that function to discriminate instantial probability properties? The notion of a discriminative capacity is familiar from sensory and perceptual psychology. As I'll explicate the kind, a **discriminative capacity** is a sensory capacity to systematically and functionally distinguish between situations that differ along some dimension. An exercise of a discriminative capacity functions to "signal" or "indicate" that a specific property within the dimension is instantiated in the current situation. Such "signalings" may be exercises of representational capacities or of non-representational capacities. To give an example of a discriminative capacity, suppose the current retinal image contains a sharp light discontinuity or "edge" at a certain location. The edge will have one of several possible orientations, measurable by an angle between the edge and some retinocentric coordinate axes. Early visual processes involve a capacity for discriminating the orientations of edges at that location. Such a capacity is a capacity for distinguishing between retinal images whose edges at the location have different orientations. An exercise of the capacity functions to "signal" or "indicate" the specific orientation that the edge in the current retinal image has.

A discriminative capacity is individuated fundamentally in terms of a dimension or set of properties. The possible orientations of an edge constitute a dimension that individuates a discriminative capacity. The pair of properties {*is a body, is not a body*} constitutes a dimension, and may individuate a discriminative capacity.[20] An exercise of a discriminative capacity is typed, at one important level of abstraction, in terms of the property whose instantiation it functions to "signal" or "indicate." A particular exercise of the orientation capacity may function to "signal" or "indicate" that the edge currently has 30° orientation. Thus, a discriminative capacity is naturally associated with a set of state- or exercise-types

[20] There are complexities attending to capacities to distinguish between two properties, one of which is a negation or complement of the other. I here include the example of body just to signal that I do not presuppose that discriminative capacities are individuated via dimensions that comprise a class of magnitudes, such as specific orientations or distances.

that correspond to the different properties in the dimension. If a type of exercise or occurrent state of a discriminative capacity functions to "signal" or "indicate" instantiation of property P, I'll say that the type **signals** P. I'll call the state-type a "signaling state" or a "signaler." I will also say that an instance of the type signals P. "Signaling" is a functional relation by definition. The orientation capacity involves states that signal 45° orientation, 46° orientation, and so on.

In addition to the example of orientation discrimination, I take the following to be paradigmatic examples of discriminative capacities:

1. A capacity to distinguish between different intensities of light impingent on a specific retinal location. Such a capacity is naturally associated with the different states of a photoreceptor cell.

2. A capacity to distinguish between different average amounts of luminance across a specific region of the retinal image. Such a capacity is naturally associated with different states of intermediate retinal neurons that receives projections from photoreceptor cells spanning the specific region.

3. An ant's sensory capacity to distinguish between different directions and lengths of the shortest path between the ant's current location and the location of its nest. Such a capacity is naturally associated with different states of the ant's internal "global vector."[21]

4. A capacity to distinguish between different distances to a distal individual along a certain line of sight. Such a capacity is naturally associated with different representational states as of a distance and direction.

[21] I borrow the term "global vector" from Burge (2010, pp. 499ff.). Ant navigation is explained by postulating an internal capacity whose states are modeled by different vector quantities, corresponding to different vectors that characterize the spatial properties of direct paths from ant to nest.

I want to highlight a commonality shared by paradigm cases of signaling states. The commonality is that *instances* of a signaling state *succeed or fail* in some way, at least typically, if they are not roughly contemporaneous with instantiation of the property the state signals by the relevant entity in the local environment. Suppose that an edge is present at a location in the retinal image, and that it has 15° orientation. Suppose that, at roughly the same time, the capacity for discriminating edge-orientations at that location is exercised. The exercise is an instance of a state that signals 30° orientation. Intuitively, the exercise suffers a failure. Such a failure can be made vivid by considering the ways in which the orientation capacity is relied on by downstream processes. Suppose states of orientation capacities feed into subsequent processes of edge-grouping. The edge grouping mechanism functions to identify salient patterns of light discontinuity across the retina. The mechanism effects a mapping from edge orientation states to groupings. The mapping is formed to optimize the groupings *if* the antecedent edge orientation states faithfully signal the orientations of edges in the retinal image. If one of the edge orientation states is off, the resulting grouping might be inappropriate. In that way, the grouping mechanism relies on the orientation capacity to faithfully signal the retinal orientation. Other processes of early image segmentation may rely on the capacity in similar ways. I do not mean to suggest that an instance of the orientation capacity fails only if such a downstream process fails because of the orientation-state's failure.

Such successes and failures are even clearer for exercises of *representational* discriminative capacities. Consider a token state with the representational content

$$\underline{\text{that } \mathrm{x}_1 \text{ blue}(\mathrm{x}_1) \text{ surface}(\mathrm{x}_1)}.$$

The state is a structured exercise of different representational capacities. It involves exercises of capacities for referring to individuals, and for attributing properties of color and surfacehood to individuals. Suppose the state successfully refers to a surface and successfully attributes the property *surface* to it. Suppose the referred-to surface is not blue. Then the token state is not fully veridical, and fails to fulfill a representational function to be fully veridical. Moreover, the failure is specifically attributable to the exercise of the capacity for attributing color that is embedded in the complex representational state. The nature of the failure is in a misattribution of color. The exercise of color attribution functions to represent the surface *currently* referred to as *currently* blue.[22] Such a function is fulfilled only if the referred-to surface is currently blue. The exercise fails to fulfill this function. The failure is partly grounded in the non-instantiation of *blue* by the referred-to surface.

If an exercise of a discriminative capacity fails partly because the relevant entity in the local environment does not, at the time of the exercise, instantiate the property signaled by the exercise, I will call the failure a **signaling failure**. Paradigm examples of signaling states have instances that sometimes suffer failures of co-occurrence. Several questions about failures of co-occurrence arise. Is susceptibility to co-occurrence failure constitutive of all signaling states? With respect to what function or functions is the failure a failure, in general? I bracket these questions. For now, I introduce the notion simply to mark one respect in which discriminative capacities are fundamentally capacities for reporting on the *current* condition of the creature's local environment.

2.5.1 Sensory signaling as a natural kind of functional encoding

In order to further explicate signaling states, I want to consider them in the context of the broadest possible class of functional sensory relations. I will reserve the term **functional encoding relation** for the broadest class of relations between sensory states and other conditions that involve or entail functional differential sensitivity.

I will give an explication of this broad class. A sensory state, process, transition, capacity,

[22] Here too I am being slightly loose. The occurrent referential applications entails that the accuracy conditions of the state are indexed to the present moment. In saying that the exercise functions to represent the surface as currently blue, I simply mean that the exercise has an attributional function that is fulfilled only if the surface is currently blue. I do not mean to imply that the perceptual system has an attributive like currently, much less that it has an attributive like referred to.

or system (type or token) S of an organism. **functionally encodes** a state (type or token) C if there is some environmental set-up type E such that (a) S bears an informational relation to C in E, and (b) the fact that S bears that informational relation to C in E is part of an explanation of at least some activity, capacity, behavior, process, transition, or use of a sub-personal state, either by or within the organism, or another member of the organism's species (past or present), or by an ancestor species of the organism. [23]

Functional encoding is an extremely capacious kind, owing primarily to the breadth of ways in which a condition can "be part of an explanation" of a sensory state. The class includes both representational and sub-representational sensory states. It includes our paradigm examples of signaling states. It also includes relations between sensory states and various conditions that do not seem to be signaling relations. At least, it includes relations that do not seem to fit the kind suggested by our paradigm examples.

To give an example, consider the magnetosome of a bacteria. Suppose that state S of the magnetosome correlates heavily with the local presence of oxygen O. Suppose that the correlation is effected via a series of correlations. Suppose that the presence of local oxygen causes and correlates heavily with the local presence of a certain electrochemical property E. Suppose that E in turn causes and correlates heavily with the presence of a property of the local magnetic field M. And suppose that S correlates heavily with presence of the magnetic property M at its epithelial receptors. It is in virtue of these stepwise causally-grounded correlations that S correlates with O.

Suppose further that the differential sensitivities of S to the presence and absence of O is functional. Suppose, for instance, that possession of the type S in the species of bacteria is evolutionarily explained partly by its sensitivity to O. Suppose O is harmful to the bacteria. Instances of S lead to a scrambling behavior, which typically leads to relocation to more oxygen-poor areas. We can suppose that the scrambling behavior would not be adaptive if

[23] I follow Burge (2010, p. 316-7) in using the term "informational relation" to include causal, probabilistic, nomic and/or structural relations to a condition.

it were triggered too often in the absence of oxygen. (Perhaps it is somewhat metabolically costly.) Thus, in the relevant environment, S confers a fitness advantage on bacteria that possess it, as compared to otherwise similar bacteria that do not possess it. The contribution of S to this fitness advantage centrally depends on its differential sensitivity to O. The fitness advantage is grounded in the fact that the scrambling behavior tends to be triggered when it is advantageous (in helping to avoid deleterious oxygen) and tends not to be triggered when it would be disadvantageous (in wasting energy). This correlation—between the behavior and condition in which the behavior is advantageous—is grounded in the fact that S causes the behavior and that S correlates with oxygen. The relation between S and O is reminiscent of the relation between the ant's sensory vector state and properties of its shortest path home. The relation between S and O appears to be a signaling relation.

The explanation of the state S cites correlation with a condition, which condition directly explains the adaptiveness of the behavior issuing from the state. A fuller explanation of the state, however, cites not just the condition that directly explains the adaptiveness of the behavior, but also those aspects of the creature and its environment that enable the behavior to be adaptive. Were it not for its correlation with magnetism, S would not correlate with O. (We can assume that no other proximal physical magnitude both differentially correlates with oxygen and is minimally discriminable by any easily-evolved epithelial capacity of the bacteria.) Similarly, were it not for the correlation between magnetism and the local electrochemical property E, S would not correlate with O. The fact that S correlates with E is part of an explanation of how S contributes to fulfillment of a biological function. It is thereby part of an explanation for why S persists. Additionally, there is a sequence of internal biochemical events B in the bacteria that lead from activation of S to scrambling behavior. S correlates with B. Had S not correlated with B, S would not correlate with O.

On my explication of functional encoding, S functionally encodes both the intermediary electrochemical condition E and the subsequent biochemical sequence B. But neither relation fits the kind of sensory relation that the orientation states and the ant's vector state typify. The relations do not seem to be signaling relations. S does not seem to discriminate or signal the electrochemical state E in the same way that it signals the presence of oxygen O.

On the explication of functional encoding given, an informational relation counts as a functional encoding relation if it is part of *some* explanation of the state. A natural strategy for identifying kinds within functional encoding, then, is to identify different types of explanation, and different types of roles that informational relations can play in such explanations.

Although a full explanation of the bacterial state S cites informational relations to E, M, O, and B, not all such relations have the same centrality or importance in every explanation of the bacterial state. For instance, suppose our explanation centers on the question: how does S contribute to the fulfillment of biological functions, if at all? Such a question follows naturally from an attempt to understand why bacteria with S persist and proliferate in populations of bacteria. Let us suppose that the biological function in question is the master biological function: staying alive long enough to reproduce. The explanation of S's contribution to fulfilling this function begins by pointing out that an instance of the scrambling behavior contributes to fulfillment of the biological function only if it occurs in the presence of oxygen. In the presence of oxygen, the instance of scrambling is also an instance of *avoiding a harm.*[24] In the absence of oxygen, an instance of scrambling would be an instance of *failing* to contribute to a biological function, since it would involve a metabolic cost for which there was no counter-weighing benefit.

Of course, scrambling in the presence of the electrochemical property E will, most likely, also be a positive contribution to fulfilling the biological function—because, by dint of E's

[24] Of course, the bacteria may scramble *into* an oxygen-rich area by chance. Then the scrambling behavior avoids one harm but not another—and then, presumably, does *not* contribute to fulfillment of the biological function of staying alive long enough to reproduce. A fuller account would incorporate details like this, but for now, I elide them. One may assume that in the presence of oxygen, scrambling always leads to oxygen-poor locations.

correlation with O, the scrambling is most likely also an avoidance of oxygen. But there are plainly differences between E and O in this explanation. For one, given the probabilistic structure of the case, there are possible instances of the scrambling behavior without E that contribute to the function, but there are no possible instances of the scrambling behavior without O that contribute to the function. More fundamentally, the presence of oxygen is part of what *makes* the scrambling behavior a contribution to the biological function. Swimming through oxygen would be deleterious to a bacteria because the *oxygen* would cause the harm to its body, not the correlating electrochemical state E.

So, scrambling near oxygen contributes to fulfilment of the function, and scrambling without oxygen contributes to non-fulfillment of the function. The presence or absence of oxygen here does not only extensionally specify which instances of scrambling are or are not positive contributions to the function; it identifies part of *why* they would or wouldn't be positive contributions to the function.

The explanation of S goes on to point out that $P(O \mid S)$ and $P(S \mid O)$ are very high, and $P(\sim O \mid S)$ and $P(S \mid\sim O)$ are both very low. Since S, we may suppose, entails performance of scrambling, these probabilistic relations between S and the state of oxygen entail that a performance of scrambling is very likely to contribute to fulfillment of the biological function.

The relation of S to O is central to this explanation, in a way that the relation of S to E is not. S makes a contribution to fulfillment of biological function fundamentally because it correlates with O—because it correlates with a condition that directly determines whether a scrambling behavior contributes to the function or not.

Of course, S's contribution to biological function is routed through its entrainment of the scrambling behavior. The basic structure of the explanation of S's contribution cites its correlation to O and its entrainment of scrambling. But we may ask a more specific explanatory question. S is a sensory state. A major dimension of explaining and understanding a sensory state involves its differential sensitivity to conditions external to the sensory system. One may ask of S, not just how it contributes to fulfillments of biological functions, but specifically how its *differential sensitivities* contribute. In asking this question, one takes for granted the fact that S causes or has certain impacts on downstream behavior. One then asks: what sorts of environmental conditions does S differentially respond to, which conditions explain the biological functionality or non-functionality of instances of the behaviors and effects of S?

I suggest that the relation between O and S exemplifies a distinctive sub-kind within the broader class of functional encoding. The sub-kind is marked by its central role in explaining how a sensory state's differential sensitivities help make the uses of the state contributions to fulfillments of functions. I conjecture that this sub-class is the class of signaling states. As a first-pass explication, we may say that sensory state S signals condition C if there is an information relation between S and C, there are uses A of S whose fulfillment or contribution to fulfillment of some function F depends on instantiation of C, and C partly explains why A conduces to fulfillment of F.

This explication of signaling relations deserves further elaboration. One pressing question is: does it correctly identify the informational relations between orientation detectors and edge orientations as signaling relations? The case of orientation detection is, in several important respects, unlike the simpler cases of the magnetosome or the ant's internal vector. The latter states only have one canonical use (scrambling or entraining a specific return course). An orientation-detection state has many different uses. It may be used by processes of texture-segmentation, figure-ground separation, contour identification, and so on. Via its influence on these processes, an orientation-detection state may have various kinds of influence on downstream perceptual processes of shape identification, object tracking, and so on. Via its influence on these perceptual products, it may influence subsequent planning and behavior. There are thus multiple different functions that the state may contribute to the fulfillment of. Unlike the ant's vector or the magnetosome, however, the contribution of the orientation-detection state to fulfillment or non-fulfillment of these functions is much more diffuse. The orientation-detection state does not itself suffice to determine (e.g.) how a texture will be segmented. It influences texture segregation only together with other orientation-detection states at other locations in the retinal image.

A further complication with orientation states is that, presumably, the orientation of an edge in the retinal image does explain why uses of the state constitute fulfillments of functions when they do. Inasmuch as a process of texture segregation fulfills representational or biological functions, it will presumably do so by correctly representing a distal texture. What explains the fulfillment of the function is the texture of the distal surface.

It seems to me that there are two broad options here. One is simply to identify ways in which correctly signaling an edge orientation contributes to the well-functioning of further sensory processes. Different sensory processes in effect rely on the orientation state to reliably signal the presence of a certain edge-orientation. The operations of texture-segmentation, for instance, may be understood as exploiting natural correlations between image-types and distal conditions in a normal or evolutionarily important environment. Although a retinal image underdetermines its distal source, it (roughly) uniquely determines its distal source given various further assumptions. Some such sets of assumptions may be prevalent in the natural environment. The processes of texture segmentation are built around these assumptions. The conformity of the processes with these assumptions depends, in part, on the orientation-state getting the orientation right. If it gets the orientation wrong, it may entrain a *different* segmentation process that would be appropriate if the retinal edge did have the orientation signaled by the current orientation detector. A similar analysis may be given of processes of figure-ground separation that use the orientation state.

A second option would be to identify a different class of explanations in which relations to proximal stimulus properties are central. For instance, one may focus on the particularities of sensory processing. Why are various states used in the ways that they are? Some such uses may be explicable only by adverting to the system's grappling with idiosyncracies of the proximal stimulus.

I put these issues aside, however. The explication of signaling states given above is sufficient for our purposes.

2.6 IPP Signaling States

All of the examples of signaling states canvassed in the previous section signal properties that are not instantial probability probabilities. The orientation of an edge is not an instantial probability property. States of the retinal ganglion cell signal an average luminance in a region of the retinal image. But this average is an *empirical* or *sample* average. It is not the mean of a probability distribution. Hence the main question of this chapter: do sensory systems have capacities or sub-systems for discriminating instantial probability properties?

I propose that the noise-tracking states discussed in §4 signal instantial probability properties. The relation of a noise-tracking state to an instantial probability property seems to be one of functional correlation. I want to argue that the relation is, more precisely, a signaling relation.

2.6.1 The function of noise-tracking states

2.6.1.1 Loss functions and levels of inaccuracy

Any instance of a perceptual representational state achieves a *level of accuracy*. We restrict our attention only to those representational states that indicate properties within a natural dimension (such as individual widths, within the dimension of widths). We restrict our attention only to instances of *attributive* representational states, like <u>red</u> or <u>width of 50 mm</u>. Instances of such states function to attribute their indicant to an environmental particular that is purportedly referred to by a perceptual singular representation, whose scope the attributive falls within. We assume henceforth that the singular representations successfully refer. We focus on levels of accuracy that are grounded only in the divergence between the property attributed to an entity by an attributive and the property within the corresponding dimension that the entity in fact has.[25] When an attributive indicates a property within a natural dimension and attributes that property to entity e, I will call the property within the dimension actually possessed by e the **target** property. If a width representation is attributed to surface e, and e is 50 mm, then 50 mm width is the target width.

In the interest of mathematical convenience and consonance with standard formal techniques, we focus on levels of representational *in*accuracy, and we model it with a **loss function**. The loss function assigns to a representational instance a number, interpreted as reflecting a level of representational inaccuracy appropriate to the kind of representation. More formally, we consider a set of representations and a dimension of properties indicated by representations in the set. The loss function assigns a number to every pair consisting of a representational state from the set and property from the dimension. The number assigned to the pair is interpreted as reflecting the level of inaccuracy that accrues to an instance of a representation of the type if the target property is the other member of the pair. We assume that if the state is perfectly veridical, it achieves a loss of 0.

In principle, any function from pairs to numbers could be (or model) a loss function. Three loss functions are standard however. Suppose r is an attributive representation and w is the target property. On a *squared error* loss function, the loss is $L(r, w) = (\llbracket r \rrbracket - w)^2$. On a *linear absolute error* loss function, the loss is $L(r, w) = c|\llbracket r \rrbracket - w|$ for some constant c. On an all-or-nothing error, the loss is $L(r, w) = 0$ if $\llbracket r \rrbracket = w$ and c otherwise, for some constant c. Henceforth, I will assume that the loss function is either squared or linear error.

[25] Assuming successful reference simplifies various matters. For instance, it enables us to more easily define probabilities that an attributive type bears various levels of accuracy. Under our assumptions, these probabilities are derivable from simple joint probabilities like $P(w, \underline{\mathrm{w}})$ (i.e. the probability that the distal width is w and the representation produced is $\underline{\mathrm{w}}$). We do not have to worry about factors governing probability of referential success, and we do not have to worry about modeling levels of inaccuracy for attributives that are bound to non-referring singular representations. Furthermore, we ignore other kinds of accuracy that may accrue to attributives. Burge (2014) claims that a color attributive as of c attributed to entity e may be inaccurate, even if e is c. The case he considers is one in which the mode of presentation of the attributive is, in some way, mismatched to the current scenario. The attributive may indicate white, but its mode of presentation may be one canonically linked with representing white surfaces under blue light. If, in the scenario, the surface is illuminated by white light, the attributive is inaccurate. We however explicitly restrict attention to forms of inaccuracy grounded only in divergence betweeen indicant and distal property.

2.6.1.2 Average inaccuracy

Relative to a set-up type, given a perceptual representational type, each level of inaccuracy has an objective probability. That is simply the probability that the representational type together with any target property that would yield that loss. For example, there are two circumstances in which a representation 50 mm accrues a loss of 5. The target property must either be 45 mm or 55 mm. The probability that a representation of 50 mm accrues that level of loss, then, is equal to $P(W = 45 \text{ mm or } W = 55 \text{ mm}) = P(W = 45 \text{ mm}) + P(W = 55 \text{ mm})$.

More importantly, we can consider a representational system's *average inaccuracy* (more typically known in statistical contexts as "expected loss"). This can be thought of as a weighted average. The quantity being weighted is the loss-value for a specific attributive-target pair. The weight on this loss-value is the probability that the pair occurs. For instance, one term in the weighted average might be $L(w, r_v) \times P(w, r_v)$. The full weighted average for visual representations is

$$\mathbb{E}[L(W, R_v)] = \int \int L(w, r_v) P(w, r_v) \, dw \, dr_v.^{26}$$

This quantity is the average level of inaccuracy in visual width representations in the set-up type. It gives us the level of inaccuracy that visual width representations achieve on average, taking into account variability in both the distal widths and the visual width representations.

2.6.1.3 Average inaccuracy and optimal multimodal systems

The average inaccuracy for the multimodal width system, relative to a set-up type, is fixed by the mapping from unimodal states to multimodal representations and the probabilities of

[26] Here, $\mathbb{E}$ is the expectation operator. It gives the mean of a random variable, potentially conditional on further values. The double integral may be thought of a continuously infinite extension of a finite weighted average. If there were two representations and two widths, the average inaccuracy would be $L(w_1, r_1)P(w_1, r_1) + L(w_1, r_2)P(w_1, r_2) + L(w_2, r_1)P(w_2, r_1) + L(w_2, r_2)P(w_2, r_2)$

those unimodal states. Suppose that the multimodal representation is fully determined by the two unimodal representations (contrary to our model, which assumes that the multimodal representation also depends on the unimodal dispersion-tracking states). Suppose δ is the mapping from pairs of unimodal width representations to multimodal width representations. Then the average inaccuracy in multimodal width representations, under δ, is

$$\mathbb{E}_\delta[L(W, R_m)] = \int L(w, \delta(r_v, r_h))P(w, r_v, r_v)\; dw\; dr_v\; dr_h$$

Each term in this weighted average concerns a specific case, in which w is the target width and unimodal representations r_v and r_h are produced. The quantity being weighted is the level of inaccuracy that would accrue to the multimodal width representation produced from r_v and r_h—that is, it is $L(w, \delta(r_v, r_h))$. The weight on this quantity is the probability that the three events jointly occur: $P(w, r_v, r_h)$.

Let M be a set of mappings, and suppose $\delta \in M$. We'll say that mapping δ is **optimal within** M if no other mapping in M achieves strictly lower average inaccuracy than δ.

A mapping may be optimal in one class of alternatives but suboptimal in another. Consider the set M of mappings that produce multimodal representations only from the unimodal representations produced, and not any other sensory information. (More precisely, the mapping depends only on unimodal states individuated by the widths they indicate, and not by further detail such as their modes of presentation.) Let δ be the optimal mapping in this set. Consider now the set M' that includes M but also contains all possible mappings that take into account dispersion-tracking states. Thus, M' includes mappings from quadruples consisting of unimodal representations and unimodal dispersion-tracking states. Given our set-up, δ will not be optimal in M'. The optimal mapping in M' is, as noted, the mapping δ' posited by our model.

The reason that δ' outperforms δ in M' is that δ' effectively takes into account different noise conditions. In our set-up, δ will map r_v and r_h to some value r_m, regardless of the variability in the visual representations. In a high noise situation, the visual representation is likely to diverge quite substantially from the target width, and intuitively should be discounted from the multimodal representation of that width. Failure to make such discounting will lead to substantially inaccurate multimodal representations in high noise. δ' effects such discounting, but δ does not. Thus, δ' achieves a lower average inaccuracy—in fact, the lowest available within M'.

2.6.1.4 Optimality and norms

The master representational function of a multimodal system of width representations is to represent widths veridically. Following Burge (2010), I will say that there are *norms* for such a system that flow from its master representational function. A norm, for Burge, is a "standard or level of possible performance that is in some way adequate for fulfillment of a function" (ibid., p. 311). One norm for such a system is "to be reliably veridical and to perceptually represent as well as possible given the perceptual system's natural limitations, its inputs, and its environmental circumstances" (ibid., p. 312). I'll assume that this norm entails several sub-norms norms that are graded, depending on *how* reliably the system represents. I'll assume that such norms are graded by different levels of average inaccuracy. If the multimodal system has average inaccuracy ε, then it satisfies a norm to represent with reliability at least ε.

Within the set-up of Ernst and Banks (and its natural analogue), the multimodal width representation system satisfies a reliability-based norm. In fact, it satisfies the the most demanding reliability-based norm: to represent as reliably *as possible*, given the system's natural limitations, inputs, and environmental circumstances. Perhaps, in principle, there is sensory information available to the creature that *could* improve reliability, if it were detected and utilized in the right way. But detection of such information is, I'll assume, beyond the "natural limitations" or "inputs" of the system, as it currently exists. It is within the creature's power to alter its multimodal representation production on the basis of cues to noise condition. We can thus ask: how reliable is the system, given the *most* reliable it could possibly be by utilizing cues to noise condition? The answer is that the system is maximally reliable.

The noise-tracking states make a systematic positive contribution to the system's fulfillment of this reliability norm. I propose that it is relative to this contribution that noise-tracking states are IPP signalers. The correlation of a noise-tracking state with an IPP explains how the noise-tracking state contributes to the fulfillment of a function. That is, it explains why the effects of the noise-tracking state *do* contribute to fulfillment of the reliability norm. Before explaining this further, we must consider the different candidate IP properties that might be signaled.

2.6.2 Candidate properties for signaling

In setting up the basic model, I only assumed that each noise-tracking state v_i was uniquely associated with some magnitude p_i. The aim, however, is to view the set of noise-tracking states as constituting a discriminatory capacity for a dimension of distributional IPPs. On that view, there is a dimension of distributional IPPs that differ by being associated with difference precision values. A noise-tracking state v_i functions to signal the presence of the distributional IPP associated with p_i. But what are the distributional properties in the dimension specifically?

The first candidate dimension is one that comprises determinable production distributions over indicant-specified width representations, given width and noise condition. Recall that, henceforth, I am taking the random variable R to comprise a class of width attributive *types*, individuated by indicant. Thus, R = $\underline{\text{50 mm}}$ on a trial if and only if the width attributive (regardless of its mode of presentation) is as of 50 mm. As previously discussed, for any distal width w, the distribution over R given distal width w and noise condition n_i has precision p_i. We thus have a class of maximally determinate distributions (conditional on a specific w and specific noise condition n_i) all of which have the same variance and precision. Each member of this class is thus a determinate of a determinable distribution-type: $\mathbb{D}[R \mid W, N]$ *has precision* p_i in x (where x is a free variable).[27] An instance of the Ernst and Banks set-up instantiates this determinable distribution property if and only if the noise condition is n_i in that instance. (n_i is sufficient since the specific w cannot change the variance, and it is necessary since each other noise condition determines a distinct variance.) Thus, we may consider the following dimension of determinable IPPs (recalling that we assumed there to be k many visual noise conditions and hence k many distinct precisions for production distributions):

$$\{ \; \mathbb{D}[R \mid W, N] \text{ has precision } p_i \text{ in } x : 1{\leq}i{\leq}k \; \}$$

I'll use d_i as an abbreviation for the determinable IPP $\mathbb{D}[R \mid W, N]$ has precision p_i in x. We assumed that the noise-tracker v_i is highly probable given n_i and vice versa. To be concrete, let us suppose that $P(v_i \mid n_i) = P(n_i \mid v_i) = 0.95$. Instantiation of n_i is necessary and sufficient for instantiation of d_i. Thus, v_i correlates strongly with d_i. Given v_i, it is 95% likely that the current trial is one that instantiates d_i, and vice versa.

As discussed, the posterior distributions in the set-up are also Gaussians with variances in p_1,. . . , p_k. That is, for any visual width representation r, the posterior distribution $\mathbb{D}[W \mid R = r, N = n_i]$ has precision p_i. The noise condition determines not only the production variance independently of distal width, but also the posterior variance independently of the width representation that was produced. Consequently, each specific posterior distribution $\mathbb{D}[W \mid R = r, N = n_i]$ is a determinate of the determinable distribution-type: $\mathbb{D}[W \mid R, N]$ *has precision* p_i. An instance of the background set-up instantiates this determinable

[27] Recall that this determinable distribution is distinct from the determinable distribution: R | N having precision p_i. The latter determinable distribution is instantiated by a trial only if the trial instantiates noise condition n and R | N = n has precision p_i. In the set-up, R | N = n is not Gaussian, and almost certainly will not have precision p_i. Knowing just the noise condition tells you very little about the representations that will be produced, since the likelihood of the representations depends heavily on the specific width also. We can expect R | N = n to be more or less flat over the set of representations R.

distribution type if and only if the noise condition is n_i in the instance. Thus, we have another dimension of determinable IPPs:

$$\{\ \mathbb{D}[W \mid R, N] \text{ has precision } p_i \text{ in } x : 1 \leq i {\leq} k; \}$$

I'll use d_i once again to name the determinable IPP: $\mathbb{D}[W \mid R, N]$ has precision p_i in x. Once again, we have a robust correlation between noise-tracking state v_i and d_i. Given v_i, d_i is 0.95 likely, and vice versa.

There is one final candidate dimension. Any specific production distribution $\mathbb{D}[R \mid X = x]$ can be understood as a relation between a single value of the condition (x) and *every* possible value of r of R. The distribution encodes the probabilities of every representation in R, given the specific value of x. But we can also ask: given a single representation r, what is the probability that it is produced *given each possible value x of X*? Answering this question provides a relation $\mathcal{L}$ between a single representational value r and *every* possible value x of the condition variable X. So, $\mathcal{L}$ can be understood as a function on the set of X values, with r as a fixed background parameter. Notationally, $\mathcal{L}(x; r) = s$ entails that $P(R = r \mid X = x) = s$. The function $\mathcal{L}(x; r)$ is called the **likelihood function** *given* r *of* x. .

The value of the likelihood function at a value x indicates how probable r is, assuming that the random variable X takes on value x. Critically, the likelihood function $\mathcal{L}(x; r)$ is conceptually and metaphysically distinct from the posterior distribution $\mathbb{D}[X \mid R = r]$. The posterior distribution reflects the tendency of X to have certain values x, given that the process of representational production has led to r. The tendency of X in such circumstances is impacted partly by the tendency of X to take on certain values regardless of how the representational process has unfolded. That is, it is impacted by the *prior probability* $P(X = x)$. If $P(X = x_0)$ is astronomically low, then $P(X = x_0 \mid R = r)$ will be miniscule as well. Nevertheless, if $X = x_0$ *were* to occur (*per improbabile*), then $R = r$ would be, we can imagine, overwhelmingly likely. We can suppose that $P(R = r \mid X = x_0) = 0.999$. It is this kind of probability relation that the likelihood function encodes. The likelihood function $\mathcal{L}(x; r)$ does not encode how likely r makes x; it's how likely each x makes r.

What kinds of instantial probability properties correspond to likelihood functions? I'll assume that a specific likelihood function $\mathcal{L}(X; r)$ corresponds to some instantial probability property that is instantiated by a set-up if and only if $R = r$ in the instance. The instantiation of $\mathcal{L}(x; r)$ indicates the relative rarities of the r-token *if* in fact the event-type x were instantiated. $\mathcal{L}(x_1; r)$ being low indicates that, if the event-type x_1 were instantiated, then the production of r would be a low-probability result of x_1. $\mathcal{L}(x_2; r)$ being high indicates that, if the event-type x_2 were instantiated, production of r would be expected. Thus, $\mathcal{L}(x; r)$ can be seen as a *sort* of measure of the objective probability of x given r, one that ignores the intrinsic variability in X and focuses only on modal relationships between the x-values and r. If $\mathcal{L}(x; r)$ is low, and r is instantiated on a trial, then that renders x unlikely, in the sense that if x *were* the value of X, the current trial would be an instance of a highly improbable type.

Likelihood functions, in general, are not probability distributions over the variable X. If each x leads to a specific r with probability 0.99, then the sum of $\mathcal{L}(x; r)$ over all $x \in X$ is greater than 1. However, in some special cases, a likelihood function is mathematically a probability distribution over X. This is true in the set-up of Ernst and Banks (2002). For instance, given any specific representation r and noise condition n, the likelihood function on distal widths $\mathcal{L}(W; r, n)$ is a Gaussian distribution with the same mean and variance as the posterior distribution over W given $R = r$ and $N = n$. That is, $\mathcal{L}(W; r, n)$ is Gaussian with precision p_i if and only if $n = n_i$. The variance of the likelihood function, in this case, can be understood as a measure of how many distal widths lead to r with appreciable probability and how many distal widths lead to r with negligible probability. A large variance for $\mathcal{L}(W; r, n)$ indicates that many distal widths produce r with similar (low) probabilities. A small variance for $\mathcal{L}(W; r, n)$ indicates that a small interval of widths produce r with high probability, and all others produce r with negligible probability. Thus, here too, the variance of a likelihood function can be understood as characterizing the objective likelihood of r's veridicality—in a way that focuses only on the specificity of the sensory channel between r and W without consideration of the unconditional variability in widths. A small σ_i for $\mathcal{L}(W; r, n)$ corresponds to a kind of objective likelihood of approximate veridicality. Any case of radical inaccuracy would be a rare event. If the distal width's being w would constitute r's radical inaccuracy, this case of radical inaccuracy would be quite rare—even if the unconditional probability of w is extremely high. Regardless of how common w is, if it were to occur, production of r would be overwhelmingly unlikely.

As with probability distributions, likelihood relations come at different levels of determinacy and determinability. For each r, the likelihood of W given $R = r$ and $N = n_i$ has precision p_i. Thus, each such specific likelihood is a determinate of the determinable: likelihood of W given R and N has precision p_i. I will abbreviate this determinable likelihood as: "$\mathcal{L}[W; R, N]$ has precision p_i." Notice, now, that an instance of the set-up instantiates noise condition n_i if and only if it instantiates the determinable $\mathcal{L}[W; R, N]$ *has precision* p_i. Thus, we have one further dimension of immanent probability properties:

$$\{\ \mathcal{L}(W; R, N) \text{ has precision } p_i \text{ in } x : 1{\leq}i{\leq}k\}$$

We want to understand how a noise-tracker contributes to fulfillment of the reliability norm. We want to understand how the uses of the noise-tracker constitutes a contribution to fulfillment of the norm, and *why* this use is a contribution.

In brief, my answer will be as follows. Each use of the noise-tracker exerts a single level of pull on the multimodal representation towards the visual representation. Exerting that level of pull given the relevant visual noise condition is a necessary condition for the multimodal mapping to be optimal—hence, for fulfillment of the reliability norm. The explanation for *why* exerting that level of pull on the visual representation is necessary for optimality adverts in a central way to the dispersion of the visual likelihood function. Because the dispersion of the likelihood function has a central role to play in explaining why uses of the noise-tracker contribute to fulfillment of the function, they are the distributional properties signaled by the noise-tracking states. Thus, a noise-tracking state is a state that signals a determinable dispersion property of likelihood functions.

2.6.3 Likelihood functions and reliability norms

2.6.3.1 Optimality and the duomodal posterior

Consider the posterior probability distribution $P_i(W|r_v, r_h)$. I will call this the **duomodal posterior**. It reflects the probabilities of various different widths, given that the visual representation is r_v and the haptic representation is r_h (and that the visual noise condition is n_i). It is a necessary and sufficient condition for the multimodal mapping to be optimal in n_i that r_v and r_h be mapped to the multimodal representation that is the mean of the duomodal posterior.[28] More precisely, suppose that δ maps unimodal representations to a multimodal representation in n_i. Then δ is optimal in n_i if and only if, for all unimodal representation r_v and r_h:

$$\delta(r_v, r_h) = \mathbb{E}_i[W|r_v, r_h]$$

Furthermore, given the probabilistic set-up, the mean of the posterior in noise condition n_i is

$$\mathbb{E}_i[W|r_v, r_h] = \frac{p_v}{p_v + p_h}[\![r_v]\!] + \frac{p_h}{p_v + p_h}[\![r_h]\!]$$

where p_h is the haptic precision and p_v is the visual precision in n_i. In other words, the mean of the duomodal posterior given r_v and r_h will always be a precision-weighted average of the indicants of r_v and r_h. The indicant of r_v always exerts the same degree of pull on the duomodal mean in n_i. Holding fixed the haptic precision, the level of compromise in

[28] See Theorem 9.1 in Wasserstein (2004). More precisely, the claim holds if (a) the loss function is either squared or absolute loss, and (b) the duomodal posterior is Gaussian.

the duomodal mean between the haptic and visual indicants is fixed, whichever haptic and visual representations happen to be produced. Varying the haptic precision, the proximity of the mean of the duomodal posterior to $\llbracket r_v \rrbracket$ varies in a way that is fixed by the visual precision p_v.

Let us assume that noise-tracker v_i occurs if and only if the noise condition is n_i. Then, whenever it occurs, v_i pulls the multimodal representation towards the visual representation r_v to the same degree that the mean of the duomodal posterior is pulled towards $\llbracket r_v \rrbracket$. If the haptic noise-tracker pulls the multimodal representation towards the haptic representation with degree p_h and the haptic noise condition has precision p_h, then the resulting multimodal representation will be as of the mean of the duomodal posterior. If the haptic precision is p_h but the haptic noise-tracker pulls the multimodal representation to degree p'_h, then the multimodal representation will not be as of the duomodal mean. But the divergence will be due to the haptic noise-tracker, not the visual noise-tracker.

Thus, in pulling the multimodal representation towards the visual representation to the degree that it does, the visual noise-tracker contributes to aligning the multimodal representation with the mean of the duomodal posterior. Such alignment is necessary and sufficient for optimality. So, the visual dispersion-tracker's effect on the multimodal representation contributes to fulfillment of the reliability norm.

The mean of the duomodal posterior exhibits a strong regularity across different r_v and r_h pairs, as discussed. This regularity is exploited by the multimodal system. Why does the mean of the duomodal posterior have this regularity? To answer this, I will argue that features of the duomodal posterior are grounded in and explained by features of certain likelihood functions and the prior probability distribution over widths.

2.6.3.2 Explaining features of the duomodal posterior

First, consider a simpler case. Consider the posterior probability $P(w|r_v)$ (where I assume the noise condition fixed, and suppress reference to it). This posterior reflects the probability that the distal width is w given that the representation is r_v.

I claim that this posterior probability is grounded in and explained by two other kinds of probabilities involving various widths w: the likelihood of w given r_v and the prior probabilities of w. Let $\mathcal{L}(w; r_v) =_{df} P(r_v|w)$ be the likelihood for w, and $P(w)$ be the prior for w.

First, notice that the joint probability $P(w, r_v)$ is grounded in and explained by the likelihood and prior. It is, of course, a trivial consequence of the probability calculus that $P(w, r_v) = \mathcal{L}(w; r_v) \times P(w)$. But in the current case, the right-hand side explains the left-hand side. $P(w, r_v)$ may be understood as the limiting relative frequency of instances in which the width is w and the representation is r_v, among the entire class of instances of the background set-up (supposing that all such instances are n_i). There is a causal order, however, between w and r_v. Widths are themselves first chosen via some process. In our set-up, we have assumed that they are exogeneously determined. The factors determining which width is produced are not specified by the nature of the set-up type. The set-up type only specifies that they are produced with certain probabilities. Once a width has been chosen, it (together with ancillary factors) produces a visual width representation. The probability that w produces r_v is given by $\mathcal{L}(w; r_v)$

The process selecting widths, together with the factors governing representation production given a width, account for the joint probability $P(w, r_v)$. For example, suppose that $P(w, r_v) = 0.0075$. Less than 1% of instances of the set-up are, in the long run, instances in which the width is w and the representation is r_v. Suppose that $P(w) = 0.05$—i.e. 5% of set-up instances are ones in which w is the distal width. Suppose that $\mathcal{L}(w; r_v) = 0.15$—i.e., 15% of cases in which the distal width is w are cases in which representation r_v is produced.

Then of course the joint probability $P(w, r_v) = 0.0075$. 5% of all cases produce w, and 15% of these produce r_v. 15% of 5% is 0.0075. This two-step determination of joint probabilty reflects the two-step causal process linking widths and representations. First a width is produced, and then a representation is produced from it.

Next, consider the probability $P(r_v)$. This probability may again be understood as the limiting relative frequency of r_v occurrences, within the class of possible repetitions of the set-up type. This probability too is determined by likelihoods and priors. By the probability calculus, $P(r_v)$ is a weighted average of likelihood and prior values. If there were only two widths, then $P(r_v) = P(r_v|w_1)P(w_1) + P(r_v|w_2)P(w_2)$.[29] But, again, this identity belies an explanatory asymmetry; the righthandside explains the lefthandside. r_v must be produced by *some* width. Some proportion of all r_v cases will be those produced by w_1. The proportion of all set-up instances that such r_v cases make up is $P(r_v|w_1) \times P(w)$. The remainder of all r_v cases are produced by w_2. The proportion that all r_v cases make, relative to all instances of the set-up, is the sum of these two proportions.

Now consider the posterior probability $P(w|r_v)$. This probability reflects the proportion of r_v cases in which the distal width is w. Such a proportion can be understood as $P(w, r_v)/P(r_v)$. The proportion of w cases among r_v are all and only the instances of w and r_v. Thus, $P(w, r_v)$ and $P(w|r_v)$ assign proportions to the same set of possible set-up instances—those in which w and r_v both occur. The former is the proportion of such cases relative to *all* possible instances of the set-up, and the latter is the proportion of such cases to cases of r_v. But, as argued in the previous sections, both quantities are grounded in the priors $P(w)$ and the likelihoods $\mathcal{L}(w; r_v)$. Those quantities determine both how frequently w and r_v occur, and also how frequently r_v alone occurs. They thereby determine how frequently w occurs within the class of r_v occcurrences.

Now consider the posterior $P(w|r_v, r_h)$. This reflects the proportion of w cases within

[29] The general case is $P(r_v) = \int \mathcal{L}(w; r_v)P(w)dw$.

the class of cases in which the visual representation is r_v *and* the haptic representation is r_h. The same considerations as above entail that this posterior is grounded in likelihoods of the form $\mathcal{L}(w; r_v, r_h)$ and the prior $P(w)$, for various w. However, in the current set-up, the production of unimodal representations are independent of one another. The factors that lead from distal width to production of visual representations operate in the same way, and with the same variabilities, regardless of how the process leading to haptic representations unfolds—and vice versa. A mathematical consequence of this assumption is that, for all w, $\mathcal{L}(w; r_v, r_h) = \mathcal{L}(w; r_v) \times \mathcal{L}(w; r_h)$. Once again, I propose that this identity belies an explanatory asymmetry: the right hand side explains the left hand side. Suppose the proportion of r_v and r_h instances among w is 0.0005. Such instances are the results of independent processes operating. Any such case must be one in which r_v is produced. Suppose $\mathcal{L}(w; r_v) = 0.05$—5% of representations produced by w are r_v. So, there is only a 5% chance of w leading to one necessary condition for r_v and r_h to both hold. By dint of independence, the proportion of r_h cases within this subclass of r_v instances produced by w is equal to the probability that w produces r_h at all. Suppose $\mathcal{L}(w; r_h) = 0.01$. Then 1% of cases in which w produces r_v will also be cases in which w produces r_h. 1% of 5% is 0.0005.

The above considerations apply for any width w and any representation r_v and r_h. So, since each posterior probability $P(w|r_v, r_h$ is grounded in likelihoods $\mathcal{L}(w; r_v)$ and $\mathcal{L}(w; r_h)$ and the prior $P(w)$, I conclude that the posterior *distribution* $P(W|r_v, r_h)$ is grounded in the likelihood *function* $\mathcal{L}(W; r_v, r_h)$ and the prior *distribution* $P(W)$.

Our claim is more specific, however. Our claim is not just that the duomodal posterior distribution is grounded in the likelihood function and prior distribution. It is that the *mean* of the duomodal posterior is grounded in and explained by parameters of the likelihood function. I turn now to this claim.

Recall that the likelihood function $\mathcal{L}(W, r_v)$ is, mathematically, a Gaussian probability distribution. It can be understood as a bell-shaped curve over widths, assigning values to each that jointly satisfy axioms for probabilities. Mathematically, then, we can apply operations like averaging and taking a variance to the likelihood function. However, these operations must be interpreted with care. For instance, we may define

$$\mathbb{E}[\mathcal{L}(W, r_v)] =_{df} \int w \cdot P(r_v|w)dw$$

Such a quantity tells us something about which width is *most likely* to produce representation r_v. In fact, since $\mathcal{L}(W; r_v)$ is a Gaussian function, $w = \mathbb{E}[\mathcal{L}(W, r_v)]$ if and only if w *is* the width most likely to produce r_v. Similarly, the variance of the likelihood function may be understood as measuring how many widths produce r_v with any appreciable probability. Alternatively, suppose w is the mean of the likelihood function, and w' is a width some fixed distance away from w. The variance of the likelihood function can then be understood as measuring the difference in the probabilities with which w and w' produce r_v. We may assume that, given a low variance, w produces r_v with far greater probability than w' does. A higher variance would indicate that w and w' produce r_v with more comparable probabilities.

The mean of the likelihood function $\mathcal{L}(w; r_v, r_h)$ is determined by the means and precisions of the likelihood functions $\mathcal{L}(w; r_v)$ and $\mathcal{L}(w; r_h)$. Since the former is a Gaussian function, the mean of $\mathcal{L}(w; r_v, r_h)$ is the w that makes r_v and r_h most likely. The mean of $\mathcal{L}(w; r_v)$ is $[\![r_v]\!]$ and the mean of $\mathcal{L}(w; r_h)$ is $[\![r_h]\!]$. Suppose p_v and p_h are, respectively, the precisions of these likelihood functions. Suppose $p_v >> p_h$. Then the w that makes r_v and r_h together most likely will be one that is very close to $[\![r_v]\!]$. Since the precision of the visual likelihood is high, such a w will make r_v quite likely (though not quite as likely as $[\![r_v]\!]$ itself would). Such a w will not make r_h particularly likely, but since the haptic precision is so low, *no* w makes r_h particularly more likely than any other. r_h is, roughly, to be as much expected given w as w'. This extreme case exemplifies the way that the unimodal likelihood precisions and means ground the mean of $\mathcal{L}(w; r_v, r_h)$. In less extreme cases, the w that makes r_v and r_h most likely will be one that strikes a certain compromise between $[\![r_v]\!]$ and $[\![r_h]\!]$. But the compromise will, again, be determined by the unimodal likelihood precisions.[30]

So, the unimodal likelihood precisions and means determine the mean of $\mathcal{L}(w; r_v, r_h)$. In the present case, however, the prior $P(W)$ is flat. Thus, the posterior $P(W|r_v, r_h)$ is determined entirely by $\mathcal{L}(w; r_v, r_h)$. In that way, the unimodal likelihood parameters determine the mean of the duomodal posterior.

Thus, the visual likelihood precision makes a systematic contribution to the mean of the duomodal posterior's being what it is. The causal impact of v_i mirrors the contribution of the likelihood precision. Thus, the correlation of v_i with the likelihood precision helps explain why its causal contributon to the multimodal process *is* a contribution to fulfilling the reliability norm. Thus, I conclude, the noise-tracking state v_i signals a likelihood precision.

2.7 Why probability signaling states are perceptual

So far, I have argued for the existence of probability signaling states that critically impact how perceptual processes unfold. What reason is there for thinking that such states are themselves states *of* or *in* perceptual systems?

In principle, the probability signaling states could be produced by post-perceptual processes. It is, of course, a delicate and contentious matter to precisely specify the conditions under which a process is perceptual or post-perceptual. That a process involves *propositional* representational capacities is very strong evidence that the process is not perceptual. It is overwhelmingly plausible, on both conceptual and empirical grounds, that perceptual states never have propositional representational contents. The structure that makes a content propositional is not needed for explanation of perception in perceptual psychology, and

[30] Here it is worth once again emphasizing how much this inference depends on the special features of the case in the set-up we have been working with. In general, likelihood functions need not be identical in shape across r_v and r_h, and the relations between the parameters of $\mathcal{L}(w; r_v, r_h)$ and $\mathcal{L}(w; r_v)$ and $\mathcal{L}(w; r_h)$ need not be as simple as weighted averages. The particular explanatory role of unimodal likelihood means and precisions in grounding the duomodal likelihood flows from the particularities of the set-up we have assumed.

it is not needed to explain the functions that perception has.

The probability signaling states discussed, however, do not in any way seem to require propositional representational states. Propositional representational states typically (perhaps constitutively) signal capacities for literal inference. Such capacities involve a certain level of sophistication. One could imagine a system that signals probabilistic magnitudes by executing propositional statistical inference. Such a system would move from premises about, for example, collections of past observations and apply representations of statistical generalizations to reach conclusions about the probabilities. But there is no reason to posit such sophisticated processing to account for the formation of the probability signaling states. Simpler explanations are available. As the account of those states made clear, all that is required for a system to signal probabilities (in the relevant sense) is to be differentially sensitive to them and to have systematic effects that are centrally explained in terms of the presence of the probabilities. Furthermore, the differential sensitivity to probabilities is effected by exploiting or developing familiar sensory capacities. The probability signaling states need only distinguish sensory cues related to the relevant probabilities—sensory cues like blur or dot garbling. Nothing more than such straightforwardly sensory processing is needed to account for the production of the state.

Additionally, the processes that use the probability signaling states are relatively simple. Each state need only be associated with a causal magnitude—a "level of pull." The causal efficacy of the signaling state is simply a matter of the state interacting with other states in virtue of their psychological magnitudes. Such causal relations are familiar from perceptual psychology. Explanations of the production of sensory and perceptual states advert to antecedent states interacting in virtue of their magnitudes. A representation of slant — itself a state characterized by a certain magnitude — may be produced by several antecedent sensory magnitudes interacting in a certain way (for instance, psychological magnitudes representing the aspect ratios of texture elements on the surface). Ma et al. (2006) provide an influential model that demonstrates how optimal cue integration could be implemented using only simple and uncontroversial sensory neural processes.

Furthermore, it is relatively easy to imagine how evolutionary and developmental processes could gradually lead to states that have the causal powers we attributed to the signaling states. Over generations, or over developmental time, the causal magnitude of the state is refined and shifted. The push towards refinement stems from background biological functions—in this case, ones that prioritize reliable representation. In virtue of the functions and the surrounding ecology, the causal magnitude of the state is refined until it cannot be refined any further for the purposes of the function. The end result is the probability signaling state as we postulated it.

Similar considerations count against the states' being post-perceptual but *not* propositional. For instance, we may imagine central psychological modules (akin to Carey (2009)'s core cognition modules) that produce representations of causation and social relations. Those subject matters may be, relatively speaking, more complex than the subject matter of perception. Causal and social relations do not always manifest themselves in a straightforwardly sensory way. They must be gotten at in more indirect ways. This might lead us to postulate relatively sophisticated processes in such modules. We can assume that such processing is not propositional. For instance, Camp (2009) argues that primate social cognition is sophisticated enough to warrant postulating certain kinds of quasi-recursive operations on representations that have a hierarchical *tree*-like structure. The processes seem more sophisticated than at least paradigm perceptual processes, and yet they are limited enough to disfavor positing full-on propositional inference. Similar considerations may be raised for causal cognition that is theorized in terms of operations on causal *graphs*. But it should be clear from the foregoing that no such added sophistication, sub-propositional though it may be, need be postulated to account for probability signaling states.

perceptual experience could assign levels of confidence to representational contents even if perceptual *systems* did not. In principle, the source of the experiential confidence might be post-perceptual but pre-cognitive. For instance, Carey (2009) introduces a notion of a *core cognition* module. It is possible that some such modules are sufficiently domain-general to count as post-perceptual, yet do not involve *propositional* representations—hence, as I use the term, pre-cognitive. A core social cognition module, for instance, may be post-perceptual, and yet exclusively utilize non-propositional tree-like representations (see Camp 2009). Perhaps a similar module is the source of levels of confidence assigned in perceptual experience.

CHAPTER 3

Levels of confidence

Our main question is: are there levels of confidence in perceptual systems? Answering the question involves two components. First, one must identify the perceptual capacity or state-type that is to constitute an assignment of a level of confidence. Second, one must explain why the capacity or pattern constitutes an assignment of a level of confidence. To fulfill this second component, one must have some sense of the nature of levels of confidence. Ideally, one would obtain constitutively necessary and sufficient conditions for a state-type to count as a level of confidence. Less demandingly, one could identify only constitutively necessary features for a state to be a level of confidence. If such identified features are illuminating—if they appear to help explain what makes something a level of confidence—their satisfaction by a perceptual state provides evidence that such a state is a level of confidence. Even less demandingly, we could identify features of levels of confidence that are, if not fully constitutively necessary, at least deeply or pervasively associated with levels of confidence. Satisfaction of such features by a perceptual state would still provide evidence that such a state is a level of confidence.

In this chapter, I will attempt to articulate such features. In particular, I will be articulating two generalizations about confidence. One focuses on the role of a level of confidence once it has been formed. The other focuses on the processes that lead to the formation of the confidence. I will give some reasons for thinking that these roles are constitutive of a level of confidence. But my primary aim is to convince the reader that these are very deep and pervasive features of confidence, so that finding a states' playing those roles counts as strong evidence for that state's being a confidence.

3.1 Identifying the explicandum

Imagine you wake up to the morning news on the radio. The station's meteorologist is talking. The meteorologist really grates on you—he is always so jokey, unserious, and (most of all) often wrong. He says, in an indecisive way, that it will rain today, before moving on to banter with the anchors. As a result of hearing the meteorologist, you enter into a particular state of mind about the proposition it will rain today, one that could be expressed if you were, at the relevant time, to honestly answer the question "How confident are you that it will rain today?" (In this case, we can suppose you would say "Not very.") We could naturally call the state of mind a level of low confidence about or in the proposition it will rain today. My aim is to inquire into the natures of levels of confidence.

I assume that the example I gave above and others I will give below help to identify a concept of a mental kind: level of confidence. The concept is utilized in our commonsense discourse about and our commonsense understanding of our own minds and of others' minds. I assume that states of confidence are relatively familiar from commonsense understanding of our own minds and of others'—roughly as familiar as states such as believing, desiring, or imagining. Our commonsense discourse is suffused with terminology that express our levels of confidence towards various propositions. We speak of a proposition's being "almost certain," "probable," "more likely than not," or "very unlikely." Such qualifiers are sometimes precisified quantitatively, as in "I'm 60% sure that Trump will be re-elected." Sometimes we hedge in quite coarse ways: "Perhaps Jill Stein will win." Such qualifications and hedgings may sometimes be made out only of politeness or deference, but often they function to express one's mental attitude toward a proposition. They function to express one's state of partial commitment or strength of conviction; to express having a shadow of doubt or merely a suspicion in the truth of some claim. I take it that such states are familiar also from reflection, introspection, and self-understanding. We are familiar, by introspection, with deductive trains of thought that entrain the fixation of a belief and with imaginative trains of thought that entrain the fixation of a desire. So too are we familiar with a train of thought that fixes a level of confidence: just ask yourself how confident you are that Trump will be re-elected.

Like perception, belief and desire, a level of confidence is a mental kind countenanced by commonsense mentalistic explanation and ordinary self-understanding. By reflecting on our commonsense concept of perception, for example, we may uncover deep regularities in the conditions required for the concept to apply correctly and for it to be applied in commonsense explanation. For instance, it appears to be a deep fact about the commonsense concept perception of that apple that no event can fall under the concept (nothing can count as perception) unless that apple is involved in causing the event. Such a principle may be (tacitly) assumed in every explanatory use of perception by commonsense. In principle, empirical psychology could show that there are no perceptual states. All of the explanatory kinds of empirical psychology could, in principle, resist the application conditions deeply associated with the commonsense concepts associated with perception. If there were a serious question about whether a theory in empirical psychology did or did not postulate perceptions, this would require, in part, an investigation of what the commonsense concept of perception requires in order to be correctly applied. Our investigation into levels of confidence is analogous. I do not assume that, ultimately, there is any psychological natural kind that is picked out by commonsense concepts of level of confidence and that fulfills the application conditions associated with that commonsense concept.

Throughout, I will also be discussing, where relevant, the understanding of levels of confidence that is evinced in normative disciplines that make essential use of the concept. Foremost, I will be discussing patterns in the explanatory uses of confidence in decision theory and formal epistemology. In particular, I will sometimes be using these fields evidentially. That is, if a claim about levels of confidence is supported by both commonsense and is supported by most theories in decision theory and formal epistemology, I will take that as additional evidence of the truth of the claim.

Throughout, I will also be discussing, where relevant, the understanding of confidence that is evinced in *normative* disciplines that make essential use of the concept. Foremost, I will be discussing patterns in the explanatory uses of confidence in *decision theory* and *formal epistemology.* In particular, I will sometimes be using these fields evidentially. That is, if a claim about levels of confidence is supported by both commonsense *and* is supported by *most theories in decision theory and formal epistemology*, I will take that as additional evidence of the truth of the claim.

I will assume that levels of confidence are type-identified by a representational content, as suggested by a claim like "He is somewhat sure that *p*." Consequently, I assume that ascriptions like "He is a confident golfer" target some other kind of mental or behavioral condition than what I have called a level of confidence. I am interested in what unifies levels of confidence *qua* levels of *confidence*, and also in what distinguishes them, qua *levels* of confidence. Being doubtful that *p* and being almost certain that *p* are different kinds of mental states; I want to understand what constitutively distinguishes them. On the other hand, any level of confidence (whether high or low) is, intuitively, distinct from a level of desire or a level of salience. I want to understand what the different levels of confidence fundamentally have in common, and what differentiates them from other gradeable mental state-types. Finally, I will be tentatively assuming that the commonsense discourse and self-understanding I've alluded to (and will develop below) pick out a unitary mental kind. As we proceed, we may find that, surface grammar and introspection notwithstanding, there are really two sorts of mental phenomena of radically different natures countenanced by commonsense. Perhaps we will find upon reflection that, as it turns out, there are practical levels of confidence and epistemic levels of confidence, and that these are fundamentally different sorts of states. But at the outset, I want to see how far we can get by assuming that there is just one class of phenomena—the levels of confidence—and trying to articulate its nature.

I will begin by trying to find constitutively necessary conditions on being a particular level of confidence. I begin by aiming for constitutively necessary conditions that may be defended whilst prescinding from various theoretical questions about levels of confidence. For instance, paradigm commonsense instances of levels of confidence are propositional attitudes: they are type-identified jointly by an attitude and a proposition. The surface grammar of some commonsense discourse suggests that level of confidence is, itself, an attitude. A claim like "I am confident that it will rain" decomposes into a verb-phrase and a propositional complement, as does "I believe that it will rain." This suggests that a level of confidence is a unique type of attitude, one that can be born towards different propositions. An alternative theoretical option, however, is that a level of confidence is actually a species of belief, so that a high level of confidence in p is best analyzed as being a *belief that probably p*. It is not easy to determine which of these views is correct. On the one hand, it seems plausible that children and non-human animals can be ascribed levels of confidence by commonsense just as literally as adults can. But there is a serious question about whether such creatures can have beliefs embedding the concept probably. For the concept to be part of a belief's representational content, the concept must make a contribution to truth-conditions. In formal semantics, at least, the orthodox view is that "probably p" is true iff p is highly supported by contextually-supplied body of evidence E; the main disagreement among orthodox views is over what semantic constraints there are, if any, on where the body of evidence E can come from. At first glance, this view suggests that a certain amount of sophistication or sensitivity must be in place to possess the concept probably. The possession-conditions for such a concept may entail a sensitivity to evidence (if not possession of the concept evidence) that is more sophisticated than the mere fact that, typically, one's beliefs are or are not formed as a matter of causal fact in a way that tracks evidential considerations. On the other hand, in typical adults, probability operators clearly sometimes type-identify mental states without type-identifying the attitude. One can, of course, embed probability operators under conditionals, as when one thinks that if the economy doesn't rebound, probably Trump will not be re-elected. In any case, my claims in this chapter will not depend on settling this issue. The generalizations I articulate about confidence will be, I claim, equally plausible on either view about the structure of levels of confidence.

Additionally, I will be following commonsense in assuming that the possible levels of confidence towards a representational content *p* form a linear order. I will further assume that locutions such as "I am *very* confident of *p*" can be understood as locating the level of confidence in *p* towards the maximal element of the linear order. However, I will not be assuming that levels of confidence have any richer structure than this. For instance, the orthodox decision theory (both normative and descriptive) used in economics is *expected utility theory.* On such a theory, confidences are modeled as *subjective probabilities.* Thus confidences have a highly fine-grained structure. On such a view, the set of possible levels of confidence one can have towards a content are, at least, isomorphic to the unit interval of real numbers. I do not assume this.

When it is convenient, I will use "LOC(p)" as shorthand for "level of confidence in proposition *p*." I will sometimes write things like "LOC(p) > LOC(∼p)," to be understood as "the level of confidence in *p* is greater than the level of confidence in ∼*p*."

3.2 Reliance

3.2.1 Confidence and reliance in the philosophical literature

In the next section, I will investigate the different processes by which different levels of confidence are formed. But in this section, I will first be asking: from a commonsense perspective, what differentiates levels of confidence in terms of their typical uses and effects in a psychology once those levels of confidence have already been formed? Note that this question is descriptive, not normative. Our main interest will be in differentiating levels of confidence, all of which have the same representational content. What sorts of psychological uses, patterns and effects differentiate (for example) low, medium, and strong levels of confidence in a proposition p?

The presence of a level of confidence in a proposition in an individual's psychology will have many specific consequences. The level of confidence will help fix various dispositions and will have particular effects on the psychological activities that engage or use the confidence. The particular consequences of the confidence will depend on the contingencies of the individual psychology and on the particular activities that use the confidence. My aim in this section is to identify regularities in the psychological consequences of holding a level of confidence. I aim to identify regularities that obtain regardless of the confidence's representational content, and that obtain across individual psychologies and across different uses of the confidence. Furthermore, we want the regularities to be specific to the particular level of confidence. Is there a *type* of disposition that a state of medium confidence (but not low or high confidence) helps to fix? Is there a particular *type* of effect that medium confidence (but not low or high confidence) has on activities that use it? Since we are searching for such regularities among commonsense explanations that advert to confidence, any regularities we discover will almost certainly be ceteris paribus. That is to say, such regularities should support counterfactuals, but they need not (almost certainly will not) do so without exceptions.

Different possible levels of confidence in a proposition vary by degree. It is natural, then, to suppose that levels of confidence may be differentiated in terms of some other degreed quantity or condition. Once we have identified the further quantity, we can differentiate the levels of confidence in terms of their causing or explaining different levels of the quantity. Structurally, this approach to differentiating levels of confidence is exemplified in much of the philosophical literature explication levels of confidence. Take Ramsey's famous explication of "degree of belief":

> We are driven therefore to the second supposition that the degree of a belief is a causal property of it, which we can express vaguely as the extent to which we

> are prepared to act on it. (Ramsey 1926, p. 170)

Ramsey introduces a degreed quantity: degree of preparedness to act on a level of confidence.[1] On this picture, the difference between low and high confidence in p is that one has only a low level of preparedness to act on the low level of confidence and a high level of preparedness to act on the high level of confidence. As the quotation suggests, Ramsey's account is behavioristic. The only causal differences between levels of belief that he countenances are differences in observable behaviors and actions. To illustrate this behavioristic cast, and to see what Ramsey might mean by "acting on a level of confidence," consider one of his examples:

> I am at a cross-roads and do not know the way; but rather I think one of the two ways is right. I propose therefore to go that way but keep my eyes open for someone to ask; if now I see someone half a mile away over the fields, whether I turn aside to ask him will depend on the relative inconvenience of going out of my way to cross the fields or continuing on the wrong road if it is the wrong road. But it will also depend on how confident I am that I am right; and clearly the more confident I am of this the less distance I should be willing to go from the road to check my opinion. I propose therefore to use the distance I would be prepared to go to ask, as a measure of the confidence of my opinion... (Ramsey 1926, pp. 174-5)

Ramsey takes the maximum distance you'd be willing to go off the chosen path as measuring one's "preparedness to act" on the level of confidence one has that one has chosen the correct path. The greater one's confidence, the smaller the maximum distance one would deviate. Equivalently, the greater one's confidence, the greater the minimum distance one would decide *not* to deviate. Although Ramsey speaks of acting *on* a level of confidence, I

[1] The surrounding context of Ramsey's discussion makes it clear that by "degree of belief" he has in mind level of confidence, rather than just a degree that attaches to a state that is already a flat-out belief.

find this terminology slightly obscure. By "act on a level of confidence," Ramsey presumably does not mean just that the confidence plays a causal role in one's deliberation process. The level of confidence plays (or could play) a causal role in both deciding to deviate and deciding not to deviate from the road. It seems rather that acting on a level of confidence is somehow acting on the *proposition* that is the confidence's content. Declining to deviate a particular distance is acting on the proposition <u>I am on the correct path</u>. But what does acting on a proposition mean? In this example, there is apparently some kind of connection between the *evaluation* of adopting the action and the proposition's being *true*. Declining to deviate is, in some intuitive sense, a *good idea* only if one *is* on the correct path. At a minimum, declining to deviate is only successful as an action if one is on the correct path. If one is on the correct path, then declining to deviate saves an extraneous journey. If one is *not* on the correct path, then declining to deviate misses an opportunity at correcting one's path. If we assume that one will only encounter at most one person at a distance at one's path, then declining to deviate if one is not on the correct path entails the further failure of not getting to one's desired destination. On Ramsey's picture, increased confidence that one is on the correct path entails a larger set of distances that one would not deviate over, if a person had been at that distance. Since each such failure to deviate is successful only if one is on the correct path, increased confidence leads to a greater number of actions one is willing to undertake whose success entails the truth of the confidence's content.

As it stands, Ramsey's account is overly behavioristic, in that it focuses too narrowly on the observable behavioral consequences of confidence under idealized conditions. One could be too proud to ever deviate from the road, regardless of one's confidence that it is the correct path. Ramsey's account also focuses entirely on the practical uses of confidence, and it therefore does not straightforwardly identify any commonality between practical uses of confidence and any uses of confidence in less practically oriented activities (such as theoretical inference or estimation). On the other hand, Ramsey's account highlights a link between the truth of a confidence's content and the evaluation of activities that use the confidence.

This link will inspire us as we proceed.

The strategy of differentiating levels of confidence by linking them to differences in some other quantity is also exemplified in this quotation by Joyce:

> A person's credence in a proposition X is her level of confidence in its truth. This corresponds, roughly, to degree to which she is disposed to presuppose X in her theoretical and practical reasoning. (Joyce 2009, p. 263)

Joyce provides this claim to orient his readers; he is not attempting to provide a detailed account of confidence. Still, I find Joyce's claim worth considering. As with Ramsey, levels of confidence are differentiated by their associations with different levels of some other degreed quantity. As with Ramsey, they are differentiated by their corresponding to different degrees of *being disposed* towards something. Unlike Ramsey's, Joyce's explication is mentalistic: the degree attaches to a disposition to use the content of the confidence in a certain way in mental processes (viz. 'theoretical and practical reasoning'), rather than in just a disposition to overtly act in a way that is dependent on the content. Furthermore, Joyce's account explicitly covers both practical and theoretical cases.

Furthermore, like Ramsey's explication of confidence, I take Joyce's to evince a connection between the truth of a confidence's content and the evaluation of some process. I assume that "presuppose p" (where p is the content of the confidence) is to be understood more broadly than just literally assuming or accepting p as a premise. Of course, there are cases in which one's confidence does influence whether to assume p. Consider for example:

> **Traffic**. You have to figure out how to get to the East Side for a dinner. You could take the train, but you have never taken the train. Planning a train route would involve exhausting and boring research of the train schedules, the stops and lines, etc. You know the bus lines quite well, however. The bus will get you there on time only if there's light traffic. You are fairly confident there will be light traffic. So you decide to make a plan involving only bus routes.

In this example, you literally and consciously assume, for the sake of further planning and decision-making, that there will be light traffic. Furthermore, we can imagine that, if your confidence in light traffic were lower, you would not literally and consciously presuppose light traffic. Rather, you bite the bullet and put together a *contingency* plan. You read the train schedule and now have a plan for what to do if you should find heavy traffic. However, many (perhaps most) familiar uses of confidence in *p* do not involve anything like a conscious assumption or acceptance of *p*. If I purchase a bet that Secretariat will win, I do not assume or accept the truth of <u>Secretariat wins</u>.

I take it, then, that the notion of a process's presupposing *p* should be understood, as we did with Ramsey's "acting on a proposition," as an entailment relationship between the process's *success conditions* and the *truth* of *p*. Such a relation holds even in the Traffic case. In assuming light traffic, you make the truth of the confidence's content a necessary condition on the successfulness of your plan. If you had been less confident in light traffic, you would have made the contingency plan. The whole point of the contingency plan is that its success does *not* presuppose light traffic.

Finally, here is a recent quotation from Burge:

> "Individuals rely, in actions and responses, on perceptual states' and perceptual beliefs' being veridical. Their psychologies have grades of reliance on or anticipation of the veridicality of these states and beliefs. The numerical subjective probabilities for veridicality postulated by Bayesian models are just numbers put on these grades of reliance or anticipation." (Burge 2020, p. 125)

Burge is not discussing confidence directly. He is attempting to explain how to understand ascriptions of subjective probabilities in Bayesian perceptual psychology. But since Bayesian models take themselves to be giving a regimented notion of confidence, I will read the claim as a proposal about confidences. I take it that, once again, "reliance" here means "relies on for some sort of success." The account liberalizes Joyce's, in simply specifying the entire *psychology's* having a grade of reliance. And it liberalizes the relation: not one of presupposing but one of reliance. I will follow Burge and use the term "degree of reliance" as the term for a relation between the successfulness of some process, event or action and the confidence's propositional content being true. Our task is then one of attempting to explicate the reliance relation.

Consider consciously deciding whether to perform one of two actions (you must choose one). You believe that each action has one of two particular outcomes depending solely on whether p is true or false. Which action you choose will depend on your level of confidence in p. If your level of confidence is below some threshold, you will choose one action, and if it is above that threshold, you will choose the other action. The aim is to articulate a sense in which each action relies on p's being true to a greater or lesser degree. The degree to which an action relies on p's being true should reflect dependency relations between evaluative properties of adopting the action and the truth or falsity of p.

3.2.2 Decision-theoretic reliance

Orthodox decision theory makes a set of idealizations about human decision-making. These idealizations are severe. However, given the idealizations, one can define a plausible measure of reliance and associate it causally with levels of confidence. I will first set up some idealizations inspired by decision theory and state the measure of reliance and its relation to levels of confidence. I will argue that the measure does capture an intuitive notion of reliance. I will also argue that the measure inspires a less precise notion of reliance, one that applies to agents who fail to conform to the decision-theoretic idealizations. I will also argue that the causal link between reliance and confidence uncovered in the decision-theoretic setting also applies, defeasibly, to actual agents. I will then consider several kinds of cases not covered, even in an idealized way, by our decision theory. I will conclude by discussing how the notion of reliance, and its causal relations to levels of confidence, may be extended to apply to these cases.

I begin by describing the idealizations made in our decision-theoretic perspective. An agent believes he must adopt one of two actions. The agent believes that the outcome of each action depends solely on whether p is true or false.[2] In particular, the agent believes that each action has one outcome if p is true and one (often different) outcome if $\sim p$ is true. The agent represents each outcome as having some quantity of utility. A representation of an outcome's utility is, roughly, a representation of the desirability or value that outcome would have, were it to occur.[3] We assume that utilities may be quantified by real numbers, positive and negative. We further assume that the agent's representations of utility are fine-grained, so that they may be modeled with real numbers. Until specified otherwise, I will be assuming that the agent's representation of the situation is veridical: he really must choose between the two actions, their outcomes do depend solely on whether p, and the outcomes do in fact have the utilities he represents them as having. An agent's representation of the possible actions and the dependencies of their possible outcomes on the truth or falsity of p is called a decision problem. A decision problem can be compactly summarized in a decision matrix. To illustrate, consider an intuitive example:

> **Detective.** You are a detective and you are chasing a fugitive down a city street. You momentarily lose sight of him and come to a fork in the road. You believe that he must have gone down one of the two paths, and you believe that if you successfully follow him down the correct path, you will catch him. To the left is the criminal underbelly of the city. You believe that if you go left, you will be

[2] It is perhaps not exactly correct to say that the agent believes that the action of each outcome depends solely on p. A quantum theorist may have to decide which of two cups to look under to retrieve a ball. We may assume that he accepts, for the purpose of this decision, that picking the right cup is successful (leads to retrieval of the ball) if and only if the ball is under the right cup. But he does not strictly believe this. He believes, rather, that picking the right cup is successful only if the ball is under the right cup and the ball doesn't quantum tunnel through the table.

[3] I am using the term "utility" for the desirability or value of an outcome's obtaining. I do not build anything else into the term "utility." I do not assume, for instance, that the utility of an outcome is measured in an aggregate way, so that the utilities of outcomes in monetary gambling are represented as the utility of one's resultant net wealth.

> beaten up by mobsters (regardless of which path the fugitive took). If you catch the fugitive, you will be showered with glory. If you fail to catch the fugitive, you will be given a slight reprimand and told to keep looking for him.

Here there are two possible actions: *go left* and *go right*. The outcome of each action depends on the truth or falsity of the proposition the fugitive went left. We can assume that getting beaten up by mobsters has -90 utility, catching the fugitive has 150 utility, and getting reprimanded has -10 utility.[4] Then the decision problem in Detective may be summarized in the following decision matrix Figure 3.1:

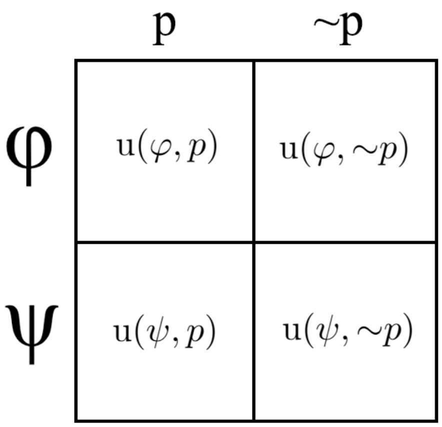
		Contingencies	
		Fugitive went left	Fugitive did not go left
Actions	Go left	60	-100
	Go right	-10	150

Figure 3.1: Payoff matrix for the Detective Case.

We further assume that the possible levels of confidence are isomorphic to the unit interval of real numbers, so that they may be measured by a real number between 0 and 1. Furthermore, we assume that LOC(∼p) = 1 – LOC(p).

[4] I am assuming here that the utility of an outcome is additive in the utilities of the events that it comprises, so that the utility of going left when the fugitive went left is the sum of the utility of catching the fugitive and the disutility of getting beaten up.

We assume that the agent always chooses the action that has the greater *subjective expected utility* (henceforth just expected utility). To explain, first notice that any decision matrix will be of the form



	p	~p
φ	$\mathrm{u}(\varphi, p)$	$\mathrm{u}(\varphi, {\sim}p)$
ψ	$\mathrm{u}(\psi, p)$	$\mathrm{u}(\psi, {\sim}p)$

Figure 3.2: A general payoff matrix.

Where u(φ,p) is the utility of φ's outcome if p is true, u(φ,∼p) is the utility of φ's outcome if $\sim p$ is true, and so on. The expected utility of an action is a weighted average of the utilities of its possible outcomes, where the weight on an outcome's utility is the agent's level of confidence in the proposition on which the outcome depends. Mathematically, the expected utility of φ is

$$\mathrm{EU}[\varphi] = \mathrm{LOC}(p)\mathrm{u}(\varphi, p) + \mathrm{LOC}({\sim}p)\mathrm{u}(\varphi, {\sim}p)$$

We assume that the agent chooses φ if and only if $\mathrm{EU}[\varphi] > \mathrm{EU}[\psi]$. (We'll ignore the case where they are equal.) We call this the agent's *decision rule.*

There are some decision problems for which one's level of confidence is irrelevant, in the sense that one's choice is determined entirely by the utility-structure of the problem, regardless of one's level of confidence in p. If φ's outcome is better than ψ's outcome if p and also if $\sim p$, one will choose φ regardless of one's level of confidence in p. If the utility of

φ's outcome and ψ's outcome are the same if p, then an agent's choice is dictated entirely by the utilities of the two actions if $\sim p$. Henceforth, we will only consider decision problems for which one's level of confidence is relevant, in the sense that there are two levels of confidence the agent could have in p such that having one confidence leads to a choice of φ and having the other confidence leads to a choice of ψ.

I will say that a proposition *p favors* φ over ψ if u(φ,p) > u(ψ,p). That is, if p favors φ over ψ, then φ's outcome if p is true is better than ψ's outcome if p is true. In Detective, the fugitive went left favors going left over going right, since if the proposition is true, the former action would lead to a caught victim (despite the beating), whereas the latter action would lead to a reprimand.

Suppose that D is a decision problem in which p favors φ over ψ. It then follows from the agent's decision rule, the assumption that confidences are relevant, and the formula for expected utility that the agent chooses φ if and only if

$$\mathrm{LOC}(p) > \frac{u(\psi, \sim p) - u(\varphi, \sim p)}{\Big[\ u(\varphi, p) - u(\psi,)\ \Big] + \Big[\ u(\psi, \sim p) - u(\varphi, \sim p)\ \Big]}$$

Call the inequality the *threshold inequality* and call the fraction on the right-hand side the *threshold formula.*[5] Note that, given our assumptions, the threshold formula takes a value in the unit interval of reals. The inequality reflects the fact that, in order for an agent to choose the action that p favors, his level of confidence in p must exceed a certain threshold. That threshold, moreover, is determined entirely by the utility structure of the decision problem. Consider again the case of Detective from above. Intuitively, if you are only slightly more confident that the fugitive went left than that the fugitive went right, you will still go right. Going left is extremely risky and could potentially lead to a needless beating. Going right is relatively risk free, and has a great reward. According to the threshold inequality in this

[5] If confidences are relevant, it follows that p favors φ only if u(ψ,∼p) > u(φ,∼p). (That is, only if $\sim p$ favorsψ). This, in turn, ensures that the threshold is one that LOC(p) must exceed.

case, an agent's level of confidence that the fugitive went left must exceed 0.78 in order for the agent to go left.

Suppose that p favors φ in a decision problem. I propose that the threshold formula provides a reasonable measure of the extent to which φ *relies* on the truth of p (in that decision problem). It suggests a conception of reliance on which reliance is a *relative* measure. It is relative in two senses. First, the extent to which φ relies on the truth of p depends not only on the utility of φ's outcome if p is true, but also on the utility of φ's outcome if p is not true. Second, the extent to which φ relies on the truth of p depends also on the utilities of ψ's outcomes if p and also if $\sim p$.

To motivate the idea that the threshold formula provides a reasonable measure of reliance, it will be helpful to investigate some entailments of its possible values. The threshold formula measures a relationship between how much better φ is than ψ if p is true, on the one hand, and how much better ψ is than φ if $\sim p$ is true. To see this relationship more clearly, define the *advantage* of an action φ if p is true as the difference between φ's outcome if p and ψ's outcome if p. In notation:

$$A(\varphi, p) = u(\varphi, p) - u(\psi, p)$$

The larger A(φ,p), the better φ's p-outcome is compared to ψ's p-outcome. In Detective, if the fugitive went left, then going left has a substantial advantage over going right. On the other hand, if the fugitive went right, then going right has a tremendous advantage over going left. If the fugitive went right, going left would lead to a pointless beating, whereas going right would lead to glory with no beating at all. Hence, A(*going left*, the fugitive went left) is much less than A(*going right*, the fugitive did not go left).

It follows that if, in a decision problem, the threshold formula equals T, then

$$\frac{A(\psi, {\sim}p)}{A(\phi, p)} = \frac{T}{1-T}$$

Call the left-hand fraction the *advantage ratio* for φ (where, again, we are assuming that p favors φ in the decision problem).For example, if T is 3/4, then A(ψ,~p)/ A(φ,p) = 3/1. That is, the advantage of ψ over φ if ${\sim}p$ is three times greater than the advantage of φ over ψ if p. The greater the threshold formula, the greater the advantage ratio.

The advantage ratio helps us to see the sense in which the threshold formula measures a degree of an action's reliance on a proposition p. Consider the difference between the following two decision problems: There is an intuitive sense in which φ relies more heavily on p in D_2 than in D_1. In both decision problems, if one adopts φ, and if p is true, then one gets the same outcome. Indeed, in both decision problems, if p is true, choosing φ incurs the same relative gain over having chosen ψ. In both decision problems, if p is true, the outcome of having chosen φ is better by the same amount than the outcome one would have gained had one chosen ψ. Choosing φ, seen as the act of choosing of φ *over* ψ, relies on p partly in that it relies on p's truth in order to reap the relative benefit of φ's outcome over ψ's. On the other hand, consider the contingency that p is *not* true. In D_1, if one adopts φ and p is not true, then although one reaps 5 utiles, there is an element of regret or missed opportunity in

	p	~p
φ	10	5
ψ	5	15

D_1

	p	~p
φ	10	5
ψ	5	20

D_2

the choice one has made. In having chosen φ, and in its having turned out that p is not true, one incurs a kind of counterfactual regret; one could have done 10 utiles better if one had chosen ψ. Choosing φ, then, relies on p partly in that it relies on p's truth in order to *avoid* incurring such counterfactual regret. In D$_2$, the counterfactual regret of having chosen φ if p is not true is greater than it is in D$_1$. Choosing φ relies on p more heavily in D$_2$ than in D$_1$ because in the latter, φ relies on p's truth to avoid a greater quantity of such counterfactual regret*, relative* to the same fixed advantage of φ over ψ if p is true. This relation is reflected in the advantage ratio. The advantage ratio in D$_1$ A(ψ,∼p)/ A(φ,p) = 2:1. In D$_2$, it is A(ψ,∼p)/ A(φ,p) = 3:1. The advantage of φ if p is the same in both decision problems. Relative to this fixed advantage, the advantage of ψ over φ if $\sim p$ is two times greater in D$_1$ and three times greater in D$_2$.

Consider now D$_3$: The outcomes if $\sim p$ are the same as in D$_2$. Now, however, the utility of ψ if p has gone down from D$_2$. The advantage ratio of φ in D$_3$ equals the advantage ratio of φ in D$_1$. Hence, according to the threshold formula, φ relies on p equally in D$_1$ and in D$_{3.}$ The reason is that, although φ incurs a greater counterfactual regret if p is false in D$_3$ than in D$_1$, it also promises a greater counterfactual *gain* than D$_1$ if p is true. The different counterfactual gains and losses of the two decision problems are equally balanced. Nevertheless, φ still relies to a relatively substantial extent on p in both decision problems. In both, if φ is chosen, the truth of p is still required to avoid a substantial counterfactual

	p	~p
φ	10	5
ψ	2.5	20

D$_3$

regret whilst reaping a relatively modest counterfactual gain.

If used as a measure of reliance, the threshold formula entails that the lower the utility of φ's outcome is if p is true, then, holding fixed the other outcomes, the more φ relies on p's being true. This initially does not seem intuitive. One might expect the relationship to go the other way, so that that φ relies on p more the greater its p-outcome is. After all, if p is true, φ would reap a greater reward. The advantage ratio helps to explain this. The lower the utility of φ's p-outcome, the closer it is to ψ's p-outcome. Suppose that φ's p-outcome is quite close to ψ's, and suppose that one chooses φ. If one had chosen ψ, one would have received a comparable (though lesser) outcome if p is true. But ψ also provides the advantage of incurring a, let's suppose, much better outcome if $\sim p$ is true. Then choosing ψ over φ would have avoided the counterfactual regret of choosing φ if $\sim p$ is true, whilst itself incurring minimal counterfactual regret over φ if p is true. Thus, ψ is, intuitively, the safer option. This is reflected in the advantage ratio. Choosing φ foregoes this safety for pursuit of a minor relative gain that is obtained only if p. Put this way, it is intuitive to say that choosing φ relies heavily on p's truth.

The threshold inequality relates levels of confidence to degrees of reliance, as measured by the threshold formula. It thus enables us to articulate the kind of generalization about levels of confidence we are seeking. It enables us to specify, for any agent who conforms to the decision-theoretic framework described above, a general consequence of an agent's having a particular level of confidence in a proposition.

> **Decision-Theoretic Reliance.** Suppose an agent's level of confidence in p is s. Suppose that p favors φ in decision problem D. Then the agent chooses φ in D if and only if φ relies on p to a degree less than s. Reliance of φ on p is measured by the threshold formula.

3.2.3 Generalizing beyond the idealizations of decision theory

From the perspective of Decision-Theoretic Reliance, the main role of a level of confidence in p in decision making is in determining whether or not to choose the action that is favored by p. One's level of confidence in p fixes the maximum amount an action can rely on p in order to be chosen. If one's level of confidence in p increases, one's dispositions for decision-making change. After the increase, one is now disposed to choose actions favored by p, actions which one was disposed *not* to choose before the increase. Furthermore, those actions that one becomes disposed to choose after the increase rely on p to a greater degree (as measured by the threshold formula) than any action one was disposed to choose before the increase. In this sense, increases in confidence increase an agent's reliance on p.

The threshold formula arises in an extremely idealized setting. It is not that the decision-theoretic framework we've set up is overly behavioristic. The threshold formula applies to decision problems as they arise and as they are conceived by an agent. It does not predict anything by itself about overt behavior. If one sees a person decline a bet that pays only if p, one cannot infer anything about that person's confidence in p. It may simply be that the person observes a religious injunction against gambling, so that the decision problem of whether to bet does not even arise, or, if it does, is conceived of as one whose outcomes depend not solely on p but also on e.g. God's willingness to foregive. However, the threshold formula is limited in that it applies only to agents who have confidences and utilities with extremely fine-grained structure, and who combine them exceptionlessly in accord with the laws of expected utility maximization.

However, I want to make three claims about the usefulness of considering the decision-theoretic perspective. The first claim is that the threshold inequality (or something quite similar) may very well turn out to be an approximately true description of actual human decision making. The threshold inequality is derived from orthodox decision theories, such as that of Savage (1954). Savage's decision theory is widely taken to be descriptively inadequate, and grossly so. Shafir and Tversky (1995) reviews several such descriptive shortcomings. But most of these shortcomings pertain to assumptions in orthodox decision theory about utility and its representation. Orthodox decision theory assumes that the utilities of outcomes are represented in terms of an agent's gross utility after each outcome. It assumes a certain kind of symmetry between gains and losses, so that avoiding a loss of \$5 is equivalent to gaining \$5. Kahneman and Tversky (1979) curry experimental evidence that such assumptions are incorrect. They propose that the value of an outcome is rather represented by the incremental gain or loss relative to some reference point, and they propose a representation of utility on which avoiding losses are preferred to an equivalent gain. These tenets are spelled out in their famous Prospect Theory, a theory that is regarded as descriptively far better than orthodox decision theory. But the basic understanding of decision-making in orthodox decision theory is retained in Prospect Theory. Prospect Theory assumes that agents assign aggregate measures of value to possible actions and adopt the action with the greater aggregate value. The aggregate measure of an action's value is *nearly* a weighted average of the value of each action's outcome (where the value of an action's outcome is measured in accordance with their updated utility representation). I say nearly because Kahneman and Tversky do not require the weights to sum to 1, as we did above. The weights however often will sum to 1, and when they do not, they will often sum to close to 1. In the latter case, the Prospect Theoretic analogue of the threshold formula will differ from the one described above by a small factor.

The second claim is that the threshold inequality entails gross relations between confidence and decision behavior that are intuitively familiar to us, even if our confidences and utilities do not have a fine-grained structure. To illustrate, consider:

> **Friendly Greeting**. It's 1:00 p.m., and you are on your way to a meeting. The road forks; down one road you see a pizzeria and down the other you see a coffee shop. Both are equally far away from you and there are no other eateries in sight. You know that your friend always takes her lunch at 1:00 p.m. You also know

> that her new job is nearby, so that she must be taking her lunch in either the pizzeria or the coffee shop. It would be nice to catch her and say hello. Visiting either of them won't delay you from getting to your meeting, but you only have time to visit one of the two. If you go to an eatery and she's not there, it's no big deal.

In Friendly Greeting, it is intuitive that you will visit the pizzeria if you are only barely more confident that your friend is in the pizzeria than that she is in the coffee shop. (Perhaps you vaguely remember someone at a party praising the pizzeria, but you aren't sure it was your friend.) Now consider a set of different variations on Friendly Greeting.

> **Nostalgia**. The set-up is the same as Friendly Greeting, with the following alteration. Long ago, you and your friend had a great time in the coffee shop sketching each other. If you go to the coffee shop and she's there, not only will you get to say hi, but you will also get to briefly reminisce on those nostalgic times.

Unlike in Friendly Greeting, it's intuitive that being barely more confident that she's in the pizzeria no longer leads to visiting the pizzeria. But, if you were appreciably more confident that she was in the pizzeria, you would visit the pizzeria. (Perhaps you vividly remember your friend recently praising a pizza place she just recently tried for the first time.) The increased confidence required to visit the pizzeria is predicted by the threshold inequality. Her being in the pizzeria favors visiting the pizzeria. Nostalgia differs from Friendly Greeting only in that visiting the coffee shop if she is there has a better outcome in the former than in the latter. Such an increase leads to a higher value of the threshold formula, and hence in a higher confidence that she's in the pizzeria for you to go there. Our next variation:

> **Bad Memories**. The set-up is the same as Friendly Greeting, with the following alteration. Long ago, you and your friend once had a terrible argument in the

> pizzeria. If you visit the pizzeria and she's there, you will get to say hi, which will still be nice, but invariably your conversation will be slightly awkward as you tip toe around bringing up the argument.

Again unlike Friendly Greeting, we can imagine that being slightly more confident that she's in the pizzeria won't lead you to visiting the pizzeria. But if you are fairly confident that she's in the pizzeria, you will visit the pizzeria. This increase in required confidence relative to Friendly Greeting results from diminishing the value of visiting the pizzeria if she's there. Again, this is a pattern predicted by the threshold inequality. Now consider

> **Clear Windows**. The set-up is the same as Friendly Greeting, with the following alteration. The coffee shop has huge and clean windows. The pizzeria has deeply tinted windows. Consequently, anyone in the coffee shop can see who enters the pizzeria, but no one in the pizzeria can see who enters the coffee shop. Your friend loves to people-watch bystanders on the street when she can. Consequently, if you visit the pizza shop and your friend is in the coffee shop, she will see you and think that you are trying to avoid her.

Again, it's easy to imagine that you will visit the pizzeria only if you are fairly confident that she's in the pizzeria. Intuitively, going to the pizzeria incurs risk that it didn't in Friendly Greeting; if she's not in the pizzeria, you risk annoying your friend. The increase in required confidence relative to Friendly Greeting results from decreasing the value of visiting the pizzeria if your friend is not there. This, too, is a pattern predicted by the threshold inequality. Finally, consider

> **Coworker**. The set-up is the same as Friendly Greeting, with the following alteration. You know that your friend's coworker always lunches at 1:00 p.m., and that he never eats pizza. So he must be in the coffee shop. If you visit the coffee shop and your friend isn't there, you'll at least be able to ask the

> coworker to let your friend know that you were looking for her and to say hi for you. Although this isn't as nice as saying hi to your friend in person, it's still a friendly gesture.

Again, it's easy to imagine that you will visit the pizzeria only if you are fairly confident that your friend is there. Intuitively, going to the coffee shop is the safer option, since you're guaranteed to get a greeting to your friend one way or the other. The increase in required confidence relative to Friendly Greeting results from increasing the value of visiting the coffee shop if your friend is in the pizzeria; again, a pattern predicted by the threshold inequality.

We could multiply examples illustrating that the level of confidence in p required to choose an action favored by p changes with changing the utilities of various possible outcomes in roughly the ways predicted by the threshold inequality. For example, one could vary the intensity of the various alterations from Friendly Greeting (e.g. in Bad Memories, the risk is not of awkward conversation but violent argument). I predict that one would intuitively find that such changes in intensity would lead to even higher thresholds on confidence, as predicted qualitatively by the threshold formula. The main point of the foregoing intuitive examples is just to provide evidence for the idea that the threshold inequality correctly identifies some deep general patterns about the relationship between levels of confidence and the agent's representation of the values of various outcomes. Those general patterns hold, even if utilities and confidences aren't fine-grained, and even if agents do not exceptionlessly conform to expected utility maximization.

The third claim is that the threshold formula inspires a reasonable notion of degree of reliance that can apply to agents who do not conform to the decision-theoretic framework laid out above. I argued above that the threshold formula constitutes a reasonable measure of an action's reliance on a proposition's truth. It is a reasonable measure because it is sensitive to the relation of the action's performance to what it actually reaps if the proposition is true and also to what outcomes are *avoided* if the proposition is true. If the action is performed and the proposition is true, one therein avoids the outcome that performing the action would have obtained were the proposition not true. Since performance of one action entails non-performance of the other, if the action is performed, one therein also avoids each outcome that might have transpired if the other action were performed. Each relation between an action's performance and the various possible outcomes identifies one intuitive kind of reliance. An action's performance relies on p's truth in order for it to reap its preferred outcome. An action's performance also relies on *p's* truth to reap a relative gain if p is true, a gain over what outcome would have transpired if the other action had been performed and p were true. An action's performance relies on p's truth in order to avoid the outcome it would have had if p were not true. If p is not true and the action is performed, then the action incurs a relative loss; if one had performed the other action, one would have obtained a better outcome. An action's performance relies on p's truth in order to avoid this kind of relative loss. The threshold formula suggests one way to aggregate these various forms of reliance into a single measure of an action's reliance on a proposition's truth. It suggests a conception of reliance on which relying heavily on p is exposing oneself to great risk if $\sim p$ for relatively little reward if p. On this conception, risk is understood as encompassing both the relative badness of the action's outcome if p is false and the loss relative to the other action if p were false. Reward is understood as encompassing both actual reward gained if p is true and its relative advantage over the other possible reward if p were true. As illustrated by the advantage ratio, the threshold formula suggests that the aggregate measure of reliance on p should depend on some ratio between risk and reward, understood in these relative and counterfactual senses. Such a notion of reliance can apply intuitively to agents who need not or do not conform to the decision-theoretic structures above. For instance, this notion of degree of reliance applies intuitively to visiting the pizzeria in Friendly Greeting and its various alterations.

I take the foregoing discussion as evidence that there is a notion of reliance, broadly similar to the threshold formula in the ways just described, that applies to actual human decision-making. Potential actions may be assigned a degree of reliance on the truth of the proposition that favors it. I'll assume that possible decision problems (as represented by actual agents) are linearly ordered by the degree of reliance of their p-favored action on the truth of p. I propose that levels of confidence in p differ in that they lead to adoptions of actions that differ in their degree of reliance on p. Specifically, a level of confidence in p may be associated with a degree of reliance, such that having that level of confidence in p defeasibly entrains a disposition to choose a p-favored action only if the action relies on p's truth to no more than that degree. Moving from one level of confidence in p to the next higher level of confidence defeasibly entrains that one is now disposed to choose further p-favored actions that one was not willing to choose before. These new p-favored actions are alike in that they rely on the truth of p to a degree greater than any action one was previously disposed to choose.

The purview of the decision-theoretic perspective we described was limited in several ways. The notion of reliance it inspires, and the relation of reliance to confidence, so far only applies to a particular kind of decision situation. Confidence, however, plays a role in other psychological processes also. I would like to consider some other psychological uses of confidence. I will sketch how the notion of reliance we've developed so far can be extended to apply to these uses. And I will sketch how confidence and reliance interrelate in these uses.

The measure of reliance so far only applies when there are two possible actions, each of whose outcomes is materially equivalent to the truth or falsity of a single proposition. But confidence impacts more complex decision-making. Consider

> **Keys**. You are taking out the trash, and when reach into your pocket you find that your car keys are missing. You need them within the hour. You know that they are either in the car, the kitchen, or the bedroom. Each place is equally far and equally easy to get to from your current location at the trash cans. Each place is equally far and equally easy to get to from each other. You believe that the keys will be equally easy to find in each location, if they are there. You form

a plan to search the three locations in a particular order.

This is a more complex decision problem than any we have encountered. It is a simple but paradigmatic example of contingency planning. Although you may not represent them all as you formulate your plan, there are six possible plans you could adopt—one for each possible ordering of the three locations. The outcome of any one plan now depends on which of three mutually exclusive and exhaustive propositions is true. For instance, consider the plan of visiting the car, then (if necessary) the bedroom, and then (if necessary) the kitchen. Intuitively, this plan has the best outcome if the keys are in the car. Its second best outcome entails that the keys are in the bedroom and its worst outcome entails that the keys are in the kitchen. If they are in the bedroom, you'll have made one extraneous journey (to the car); if they are in the kitchen, you'll have made two extraneous journeys.

Intuitively, in Keys, you will plan to visit the locations in an order that matches the ordering of your levels of confidence. If you are most confident that the keys are in the car, less confident that they are in the bedroom, and even less confident that they are in the kitchen, then intuitively you will plan to visit the car first, then the bedroom, and then the kitchen. Furthermore, intuitively, each plan relies on the different propositions to different degrees. Adopting a plan to visit the car, then bedroom, then kitchen relies most on the keys being in the car, since it relies on that to avoid the unfavorable outcomes of the keys being in the bedroom and kitchen. It relies less on the keys being in the bedroom, but (if the keys should fail to be in the car), it relies on it to avoid the worst outcome of the keys being in the kitchen.

In principle, we could extend our decision-theoretic treatment to Keys. We could then search for analogues of the threshold inequality in this more complex case. Rather than engage in such a task, I will simply claim that, in this more complex setting, we should expect reliance and required confidence to change in response to changing utility structures in ways broadly like the threshold formula. Consider

> **Department**. As in Keys, except now, you know that the keys are either in the kitchen, or in the bedroom, or in the department. You were distracted by an emergency phone call while packing up, and consequently, you are quite confident that the keys are in the department. It is, however, a huge hassle to get to the department. So, you form a plan to check the kitchen and then the bedroom first, on the off chance the keys are there. Then you will visit the department.

Unlike in Keys, the location you are most confident in is visited last, not first. Compare two possible plans: < kitchen, bedroom, department > and <department, kitchen, bedroom >. The latter relies very heavily on the truth of the keys are in the department. The relative advantage of the latter plan over the former if the keys *are* in the department is miniscule. If the keys are in the department, visiting the department first saves you having to visit the kitchen and the bedroom, but the cost of visiting those locations is negligible compared to the cost of schlepping to the department. Meanwhile, if the keys are in the bedroom, then visiting the department first incurs a huge counterfactual regret. If the keys are in the bedroom, the former plan terminates with only a minor hitch (having visited the kitchen first), while the latter plan terminates after having incurred a huge cost (pointlessly visiting the department). The latter plan incurs a comparable counterfactual regret if the keys are in the kitchen. The latter plan relies very heavily on the truth of the keys are in the department, and intuitively, you would have to be extremely confident in that proposition before you would adopt the plan. (Perhaps you summon a vivid auditory memory of hearing a clanging sound as you left, unidentified at the time, but now unmistakably recognizable as the sound of keys falling to the floor.)

Confidences can play a role in complex planning that is not easily assimilated to the simple decision-theoretic perspective described above. Consider

> **Traffic**. You have to figure out how to get to the East Side for a dinner. You could take the train, but you have never taken the train. Planning a train route

> would involve exhausting and boring research of the train schedules, the stops and lines, etc. You know the bus lines quite well, however. The bus will get you there on time if and only if there's light traffic. You are fairly confident there will be light traffic. So you assume that there will be light traffic, and decide to make a plan involving only bus routes.

In this example, you literally and consciously assume, for the sake of further planning and decision-making, that there will be light traffic. Because you assume this, you do not bother making a contingency plan involving train routes. Unlike the previous cases, here you are not deciding between physical actions or particular plans to adopt. You are deciding whether or not to *assume a proposition's truth* for the sake of subsequent plan-formation. Since Traffic is a case of conscious deliberation, it is not implausible to say that your assuming a proposition's truth is a mental act. The decision is whether to perform this mental act or not. Of course, deciding whether or not to assume light traffic is sensitive to represented costs and benefits of the eventual plan you might end up with. If it is critically important that you arrive at the dinner on time, then assuming light traffic if there is heavy traffic will leave you with no contingency plan, and will cause you to be late to the dinner. In this case, you might not assume light traffic. In this higher-stakes case, making the assumption that p relies far more heavily on the truth of p, in that the truth of p is required for you to avoid the catastrophic possibility of being stuck in heavy traffic with no idea how to switch to the train. Correspondingly, in this higher-stakes case, one must be extremely confident of light traffic to assume p.

A further limitation of the decision-theoretic perspective we've considered so far is that, even when one must make a choice between two actions, it assumes that the outcomes of each action are materially equivalent to the truth or falsity of a proposition. But often the outcomes of actions can depend, in an intuitive sense, on several propositions. Deciding to act will be influenced by one's levels of confidence in those various propositions. This is already somewhat present in Traffic, but to give two clearer examples:

> **Secretariat Sufficient**. Secretariat is about to race, and I am about to flip a coin that both of us know to be fair. I offer you a bet that pays \$10 if Secretariat wins the race *or* if the coin comes up heads, and \$0 otherwise. The bet costs \$1 to purchase. Do you purchase the bet?
>
> **Secretariat Necessary**. Secretariat is about to race, and I am about to flip a coin that both of us know to be fair. I offer you a bet that pays \$10 if Secretariat wins the race *and* if the coin comes up heads, and \$0 otherwise. The bet costs \$1 to purchase. Do you purchase the bet?

Intuitively, your choice to purchase the bet in either case ought to be influenced by your confidence that Secretariat will win. But Secretariat's winning is only a sufficient condition for winning the first bet, and only a necessary condition for winning the second bet. The orthodox decision theoretic perspective in this kind of circumstance is to assume that confidences are subjective probabilities. To assume this is to assume that the interrelations between different levels of confidence obey the laws of probability theory. Consider Secretariat Necessary. A decision-theoretic agent will purchase the bet only if their confidence in the conjunction <u>Secretariat wins the race and the coin comes up heads</u> is greater than 0.1. Assume the agent believes, plausibly, that Secretariat's status in the race is causally isolated from (hence statistically independent from) the outcome of the coin flip. Then the agent's confidence in the conjunction is determined by his confidence in the individual conjuncts; the former is simply the latter two multiplied. Hence individual manipulations in one confidence (holding the other fixed) will change the agent's choice behavior. Assuming 50% confidence that the coin comes up heads, one must have 20% confidence or greater that Secretariat will win in order to purchase the bet for \$1. Here, it is difficult to define a notion of degree of reliance on an individual conjunct. The threshold inequality applies straightforwardly to the conjunction. And intuitively, if the bet cost \$3, then purchasing the bet would rely more heavily on the truth of the conjunction. Still, alterations in confidence of the conjuncts still broadly lead to increased reliance. Increased confidence that Secretariat will win (while holding fixed confidence that the coin comes up heads) increases the confidence in the conjunction. Consequently, after the increase, there will be bets that the agent was formerly disposed not to accept that he now is disposed to accept. If the agent becomes 60% confident that Secretariat will win, he will now purchase the $3 bet that he wouldn't before. In both the $1 bet and the $3 bet, the conjunctive proposition favors purchasing the bet, but purchasing the bet relies more heavily on the conjunction in the latter. An increased confidence that Secretariat will win leads to adoption of the latter bet, and hence, an action that relies more heavily on its favoring proposition.

3.3 Support

Often, as one acquires stronger evidence for the truth of *p*, one's level of confidence in *p* increases. I learn that Mr. Boddy was murdered by Miss Scarlett or Professor Plum. I learn nothing else about the circumstances of Mr. Boddy's murder, and I have no information about Miss Scarlett or Professor Plum. Consequently, I am as confident that Miss Scarlett committed the murder as I am that she didn't. I then learn that Mr. Boddy was shot, and that Miss Scarlett's pistol was found at the crime scene. My confidence that Miss Scarlett committed the murder increases. I then learn that Professor Plum was out of town when the murder was committed. My confidence that Miss Scarlett committed the murder increases further still. Each time I learn new information, my body of evidence about Boddy's murderer expands. After each expansion, my body of evidence more strongly supports Miss Scarlett's having murdered Body. Correlatively, my confidence that Miss Scarlett is the murderer increases.

Such correspondences between strengths of support and levels of confidence are commonplace. I suggest that levels of confidence typically *function* to correspond to levels of strength of support. Roughly, if a level of confidence in *p* is formed from a body of information, and that body of information supports the truth of *p* to a certain degree, the confidence suffers a kind of failure if the level of the confidence does not roughly match that degree of support. On the view I suggest, a 60% level of confidence that Miss Scarlett killed Mr. Boddy, produced solely from the belief that Mr. Boddy was killed and Miss Scarlett is one of two suspects, would suffer a failure. The failure arises from the fact that the body of information equally supports the truth of Miss Scarlett is the murderer and Professor Plum is the murderer. To reflect this equal support, we can say that the body of information supports the former proposition to degree 50%. The level of confidence fails because it is not in line with this strength of support. If the level of confidence were roughly 50%, however, then it would not fail with respect to this function.

To elaborate and defend the view I suggest, I will first explicate the notion of *support* I have in mind. I will then elaborate the relation between levels of confidence and levels of support.

3.3.1 Conditional support

I begin by introducing a kind of relation between a belief (or set of beliefs) and a proposition. I introduce the relation by considering two examples.

First, consider two belief-types: a belief in p and a belief in *if p then q*.[6] Consider the proposition q. The contents of the two beliefs together form a set of premises of a deductively valid argument to the proposition q. With logical necessity, any situation in which the two beliefs are true is one in which the proposition q is true also.

Second, consider the belief-type that die a is an evenly weighted die. Consider the proposition the next toss of die a will come up facing less than six. Of course, the content of the belief does not entail the proposition. By itself, the truth of the belief-content does not even entail an objective probability that the proposition is true. But relative to a given reference class, the truth of the belief may confer an objective probability on the truth of the

[6] Suppose that the content of the second belief is a material conditional.

proposition. Relative to a reference class in which die rolls are produced by a mechanism whose initial conditions are chosen uniformly at random from some range, the truth of the belief-content fixes an objective probability of 5/6 on the truth of the proposition.

These two example relations between beliefs and propositions exemplify what I will call *conditional support relations.* A conditional support relation marks a systematic connection between the supposition that some beliefs are true and the truth of a proposition. They mark the manners and extents with which the truth of the beliefs would conduce to or indicate the truth of the proposition. Belief-types in p and *if p then q* conditionally support the proposition q. The belief-type that a is evenly weighted conditionally supports the proposition that the next toss of a will come up less than six. Tokens of those types conditionally support the proposition, even if the tokens are false.

Each of the two examples illustrates a different *kind* of conditional support relation. The relation between the beliefs in p and *if p then q* and the proposition q is a *deductive* support relation. In deductive support relations, the proposition is guaranteed by the truth of the beliefs, in a way that is grounded entirely in the syntactic structures and propositional connectives of the belief-contents. The relation between a belief that die a is evenly weighted and the proposition the next toss of a will be less than six is what I will call an *aleatoric* support relation.[7] The proposition's truth depends on the outcome of a stochastic process. The truth of the belief, relative to the relevant reference class, fixes an objective probability for this outcome, and so, derivatively, for the truth of the proposition.

Conditional support relations have a polarity. Both of our examples are examples of *positive* conditional support relations. In both cases, the truth of the antecedent beliefs makes the proposition more likely to be true than false. The truth of p and *if p then q* makes the truth of q more likely than not in the strongest possible way: it guarantees the truth of q, and guarantees the falsity of $\sim q$. The truth of the belief about the die's weight

[7] Aleatoric is defined in the OED as "depending on the throw of a dice or on chance."

makes the proposition about the roll more likely true than false, but short of a guarantee. There are also *negative* conditional support relations. Beliefs in p and *if* p *then* q negatively support $\sim q$ in making its truth less likely than not.[8]

Conditional support relations help explain and ground various psychological norms and functions. Most prominently, they help establish baseline criteria for satisfaction of various *epistemic* norms.[9] Suppose that a belief in r is formed from beliefs in p and q. The belief in r is *epistemically warranted* only if the two antecedent beliefs conditionally support r. A belief's being epistemically warranted marks it as being produced or sustained by a process that is, in some sense, a good route to truth. If the antecedent beliefs do not positively conditionally support r to some degree, the belief in r cannot be warranted. Without positive conditional support, r is more likely to be *false* if the antecedent beliefs are true. Even if a true belief r is produced from the beliefs, and even if the antecedent beliefs are both true and themselves warranted, the truth of the belief in r is not secured by a good route to truth.

Of course, positive support is only a necessary condition on epistemic warrant. The positive conditional support must be sufficiently strong to meet conditions for *epistemic* warrant. Epistemic warrant is constitutively associated with producing knowledge, and knowledge may require a relatively demanding level of conditional support. Furthermore, for a belief to be epistemically warranted, the antecedent beliefs must themselves be warranted. If the antecedent beliefs are not warranted, the transition to the belief that r would not count as a good route to truth—even if the antecedent beliefs conditionally support r to a sufficient degree.

[8] One may also consider relatively positive levels of support. Suppose a process has one of 100 outcomes. The truth of some beliefs, suppose, fixes an objective probability of 0.1 to one outcome, and $\sim$0.009 to the remaining outcomes. The proposition corresponding to the one outcome is not positively supported—its truth is still substantially less likely than its falsity. But the proposition is better supported (by two orders of magnitude) than any proposition corresponding to a different outcome.

[9] As a background on epistemic norms, I follow the account of Burge (2003, 2020).

Conditional support relations may be tied to norms and functions that are not epistemic. Burge (2020) argues that it is not a constitutive function of beliefs to *be* knowledge. A belief does not fail, *as* a belief, if it is not knowledge. This raises the possibility of psychological systems in which beliefs do not have a function to be knowledge at all.[10] Nevertheless, in such systems, conditional support relations may still be associated with norms of various kinds. For instance, being positively conditionally supported might be a kind of contribution to a belief's fulfilling its constitutive *representational* function of being true. Being more strongly positively supported is a stronger contribution to fulfilling that function. So there may be representational norms or standards associated with levels of conditional support. Such norms may apply even in creatures whose beliefs do have epistemic functions. Arguably, when produced from the belief that a is evenly weighted, the belief that the next toss of a will come up less than six cannot be knowledge. The relation of 5/6 probability is simply too insecure. If so, the aleatoric support relation does not satisfy an epistemic norm on the produced belief. But it may still satisfy a representational norm on that belief. The belief's 5/6 chance of being true is representationally better than if it had a 2/3 chance of being true.

There are support relations that are not obviously either deductive or aleatoric. For example, it is not obvious that all intuitively *inductive* support relations are aleatoric. Consider some of the beliefs of a knowledgeable experimental physicist. The beliefs include beliefs about the results of long sequences of observations and experiments. They may include beliefs about theoretical virtues, such as that ontologically simpler theories are better explanations, all else equal. The set of beliefs, we can suppose, strongly conditionally supports the proposition that the Einstein field equations hold at every point in spacetime. Intuitively, after all, a belief in the proposition held on that evidence could be epistemi-

[10] This possibility does not follow from Burge's claim, however, since constitutive relations are hyperintensional. It may be that, necessarily, if a psychology has capacities for belief, then beliefs function to be knowledge in such a system. But the relation may fail to be constitutive if, for example, the explanation for the source of the epistemic function is not the presence of the beliefs, but some other (necessarily concomitant) feature of the psychology.

cally warranted. The beliefs do not entail the proposition. There is also no straightforward aleatoric relation between the beliefs and the proposition. It would be hard to make sense of the claim that the field equations are true with a certain limiting relative frequency within a class of repeated situations in which the beliefs are all true. Even if sense could be made of the claim, it would misconstrue the nature of the conditional support. The truth of the field equations *need not* be the consequence of a stochastic process in order for the physicist's beliefs to conditionally support the proposition. A full explanation of the support relation in this example may, ultimately, advert only to objective probability relations combined in some complex way. Alternatively, it may involve support relations that are not aleatoric. My point here is simply to indicate the intended breadth of support relations.[11]

Suppose that a warranted belief B is produced from a warranted set of beliefs S, and that the conditional support relation underwriting the warrant of the belief in B is an aleatoric one. As previously mentioned, aleatoric support relations hold only relative to a reference class. What determines the reference class with respect to which aleatoric support is evaluated, when aleatoric support contributes or fails to contribute to fulfillment of some function? This is a complex matter, and I will only gesture at a partial answer. But I assume that such reference classes can be fixed by background psychological *states* or background *relations* between the psychology and the environment.

Of course, the reference class may be fixed by the beliefs in S—that is, by the beliefs from which the warranted belief B is formed. One may conclude that the die will come up with a face less than six on the basis of beliefs about its even weightedness *and* further facts about the tossing mechanism. But often the beliefs one infers from will not themselves include beliefs about a reference class. Nor must the reference class be fixed by the beliefs in the set. One may believe that that the tossing mechanism is fair, but not use or access this belief

[11] Potential examples of further distinct kinds of support relation may be relations of necessitation that, unlike deductive relations, are not grounded solely in the syntactic form and propositional connectives of the belief-contents. A possible example of such a relation may be that believing that a cup contains water conditionally supports the proposition that the cup contains H2O.

in the inference. That background belief would, I conjecture, help fix the relevant reference class. More generally, one may have a *set* of background beliefs that help fix the relevant reference class.

Suppose one lacks background beliefs that fix the reference class. However, suppose that one has a certain disposition to make inferences from considerations about the physical symmetries and asymmetries of objects. One is disposed to change one's judgement about die-toss outcomes in systematic ways given beliefs about the distribution of the die's weight. One is also disposed to change one's judgement about coin-toss outcomes given beliefs about the coin's lopsidedness. Suppose such a disposition was formed after interacting with many different stochastic systems, most or all of which involved mechanisms whose initial conditions were drawn uniformly at random from the relevant range.[12] One never, however, stops to consider the matter about the mechanisms underlying the outcomes of the stochastic processes. Consequently, one never forms beliefs about any bias (or lack thereof) in the tossing and flipping mechanisms. I conjecture, however, that the etiology of the inferential disposition fixes the reference class for the relevant aleatoric support relations.

I will speak of *levels* of support. However, my use of "level" here comes with the same caveats mentioned above about my use of "level of confidence." The polarity of support relations is familiar. A set of beliefs may count for or against a proposition's truth, or it may not count for or against it at all. It is also intuitive that support relations may be ordinally compared. At first, I have no evidence about who killed Boddy. Then, I learn that he was shot with Scarlett's gun. The second body of evidence (positively, conditionally) supports the proposition <u>Scarlett killed Boddy</u> more than the first body of evidence does. I learn that the other suspects were out of town at the time, and my new evidence base supports the proposition more than either of the two previous ones. I assume that support relations can be ordered in this way, and that a level of support marks a position in the relative order.

[12] Of course, the assumption about initial conditions is likely to be more abstract and more general. See Chapter 1.

Perhaps some support relations support richer structure, such as *ratios* between levels of support, but I do not assume this.[13] Nevertheless, for ease of exposition, I will often idealize and use numerical quantities to mark levels of support. In particular, I will often idealize a level of aleatoric support as being equal to the underlying magnitude of objective probability.

Throughout, I have focused on support relations that hold between beliefs and propositions. However, as I intend the term 'support relation,' support relations can generally hold between *committal psychological states* and *representational contents*. On my usage, a committal psychological state is one that is partly type-identified by a signaling function (representational or not) and that undergoes some sort of constitutive failure if its fails its signaling function. Paradigm committal states are those partly type-identified by representational contents. Beliefs are committal states. Perceptions are committal states also. A perception constitutively suffers a failure if its representational content is inaccurate. Perhaps various kinds of memory states are committal states, without being beliefs. If an episodic memory-trace is encoded in an iconic or perception-like way, it may not have a propositional content, and so cannot be a belief. Nevertheless, it suffers a failure if the memory is fabricated, or if it is inaccurate of the remembered situation. On my usage, however, non-representational signaling states are committal. A state that functions to signal a retinal edge orientation undergoes a failure if the retinal edge has a different orientation.

I let support relations be broad in these ways in order to account for the full variety of epistemic norms and analogues of epistemic norms. Suppose a perceptual belief is formed solely from a perception, and without the influence of any ancillary beliefs. The belief is warranted only if the perception's veridicality would conduce to the truth of the belief's

[13] I do not assume that support relations are, in general, commensurable. Perhaps one cannot ordinally rank the support that astronomical evidence confers on a stellar hypothesis and the support that criminological evidence confers on a criminal hypothesis. I do not even assume that, given a proposition, any two support relations it bears to different possible evidence bases must be comparable. All I assume is that the support relations may be partially ordered. A "level of support" may then be identified with echelons of the partial order. So, a level of support is a type of support relation, and the different levels of support are linearly ordered.

content. This condition is analogous to positive conditional support at the fully doxastic level, but the perception is not a belief. Similarly, a belief may be warranted if formed on the basis of a non-propositional memory-trace. Here too, I assume, the memory-trace must conditionally support the belief-content positively.

I assume that no states have epistemic functions in a creature incapable of belief. There may nevertheless be *analogues* to epistemic norms in such creatures. For instance, a perceptual state may meet a higher representational standard if it is formed from antecedent perceptual states that positively support the state more strongly. Similarly, a perceptual state may meet a higher representational standard if it is formed from antecedent *sensory registrations* that support the state more strongly. Suppose the proximal stimulation is as the sensory registration signals it to be. Suppose such proximal stimulation induces a certain objective probability on the veridicality of the perceptual state's content. A perception formed from registrations that (if properly signaling) induce a higher objective probability on veridicality thereby meets a higher standard associated with being veridical.

3.3.2 Levels of Support and Levels of Confidence

As I have already remarked, one's levels of confidence in a proposition quite often changes when one learns new evidence that bears on the proposition. I take it that cursory reflection on one's mental life will evince the commonality of this pattern. The pattern is most noticeable when one learns especially strong and obvious evidence for or against a salient proposition's truth. I hear a rumor that a local company has been attempting to quell attempts to unionize, but since it's a rumor and I have nothing else to go on, I am as confident as not that the company is quelling unionization. I then learn that one hundred employees have just signed and released a document detailing the ways in which they have been harassed by the company for labor actions, and my confidence increases. The pattern between confidence and evidence is also familiar in more humdrum cases. I wake up, look out the window, and see an overcast sky. I am only slightly confident that it will rain today. My partner reminds me that it's July (typically a rainy season around here), and I hear the weatherman predict rain. My confidence in rain increases.

I suggest that this pattern reflects a function of levels of confidence. An increasing level of confidence in p in response to an evidence base that has come to more strongly support p fulfills the function. Not increasing the level of confidence in such a situation would be a failure to fulfill the function.

In fact, I suggest that the function of levels of confidence is somewhat more demanding. A level of confidence must not only increase with increasing support; it must match the *level* of support. Even if it increases in response to stronger evidence, a level of confidence in p may fail the function if it over- or under-shoots the level of support afforded p by the states the confidence is formed from.

To explicate the function more precisely, I will first consider levels of confidence that are formed from beliefs. Specifically, I will consider levels of confidence formed partly on the basis of beliefs about statistical and probabilistic quantities. I begin with an idealized example. I will suppose that levels of confidence and levels of support may both be measured by precise numerical values. Using the idealized example, I will state and clarify the function.

> **Overcast.** Suppose it's morning, and I have just woken up. I walk over to my window, somewhat curious about what today's weather will be like, and I see an overcast sky. I form the conscious belief that today's morning is overcast. I then consciously recall my belief that the chance of rain on days with overcast mornings is 65%. I go on to form a level of 65% confidence that it will rain today.

I assume that my two antecedent beliefs conditionally support the proposition <u>it will rain today</u> to degree 65%. For instance, we may assume that each day in my city is an instance or "run" of reference class R. The reference class specifies that certain exogeneous variables (e.g. barometric pressure) are supplied to the city with certain probabilities (e.g. by exogeneously variable storm fronts). R specifies how the values of those variables interact with local variables (such as elevation and local terrain) to stochastically determine meteorological outcomes. Relative to this reference class, the probability of rain on overcast days is 65%. I assume that background factors about my psychology privilege this reference class. Those background factors conspire to constrain the truth-conditions for my belief, so that the belief's truth-value is determined by a conditional probability relative to the reference class R specifically.[14] More generally, those background factors conspire to make R relevant for the determination of support relations.[15] Given that R is privileged, the two beliefs conditionally support the proposition to degree 65%.

In the example, the level of confidence and the level of support match in magnitude. Such matching constitutes fulfillment of a function of the confidence. If the level of confidence had been 95%, then the function would not be fulfilled. I finish the example:

> **Overcast (continued).** After turning from the window, suppose I hear the weather report on the radio. The weatherman says that it will rain today. I form the conscious belief that today's morning is overcast, and the weatherman said it will rain. I also consciously recall another belief of mine, that the chance of rain, on days with overcast mornings *and* in which the weatherman says it will rain, is 75%. On the basis of these two beliefs, I form a confidence of 75% that it will rain.

[14] Perhaps, for instance, I read the statistic in an almanac, and I have various dispositions to defer to the authors of the almanac on matters related to the statistic. The authors are in a position to specify the reference class of the statistic more precisely. Suppose there is an indexical element in the content of my belief that functions to refer to a reference class. Suppose my deferential dispositions to the authors of the almanac help guide the indexical to refer to reference class R. I provide this as an example of one way the reference class might be fixed; there are many other possible ways.

[15] The example involves two propositions about single-case probabilities: it is overcast today and it will rain today. As analyzed in Chapter 1, single-case probabilities are relativized to a reference class. Thus, if R is singled out as the reference class for the 65% chance belief, it must be singled out as the reference class for the two single-case probabilities. Note that I do not say that the reference class is part of the truth-conditions for the belief it is overcast today. The reference class is privileged only in being the one relative to which the truth of the singular belief makes a contribution to support relations, and thereby to any functions or norms that cite the support relations.

As before, I assume that the two propositions conditionally support the proposition that it will rain today to degree 75%. So, after receiving new evidence in the form of the weatherman's report, my confidence changes to maintain match with the level of support.

Given the idealizations about the two kinds of level, matching is a matter of identical magnitude. Each level is canonically associated with a real number, indicating the level's position in and relation to the other levels in the relevant set. Given this background, two levels perfectly match if they are canonically associated with the same real number. However, matching relations can obtain even if the levels do not have such rich metric structure. Recall that the levels of confidence and the levels of support are both sets of linearly-ordered types. Matching relations may be defined so long as there is a natural order-preserving mapping between the two levels. In the simplest case, there are as many levels of confidence as there are levels of support, and both sets of levels are finite. Then each level in each set may be assigned an integer index corresponding to their position in the order. Two levels match if they have the same index.[16]

I will thus assume that there is a *matching mapping* that assigns to each level of confidence one or more levels of support. I assume that the matching mapping is fixed by the content of the level of confidence to be produced and the set of beliefs from which the confidence will be produced.[17] The level of confidence in p that is actually produced from the beliefs then **matches** the level of support for p given the beliefs if the level of support is among the

[16] More generally, either set of levels may be more fine-grained than the other. Suppose that levels of support are more fine-grained than levels of confidence. Then matching levels may be defined via a many-one mapping f from levels of support to levels of confidence. The only structural restriction on the mapping is that, if support levels a and b map to confidences f(a) and f(b), then a is less than or equal to b under the support ordering if f(a) is less than or equal to f(b) under the confidence ordering. Then the levels match if the support level is x and the confidence level is f(x). Similar considerations apply when confidence is more fine-grained than support.

[17] I allow that the matching mapping may differ in different cases. For instance, different kinds of evidence base may simply permit levels of support with finer or coarser grains. Conscious statistical belief may offer real-valued levels of support, whereas murkier evidence bases about (e.g.) who will win the election may only offer a small finite number. If such differences exist, the matching mapping of course must be different. It may change for other reasons also.

values associated with the level of confidence by the matching mapping.

Matching can be illustrated in cases where neither confidence nor support have the rich structure of real numbers in the unit interval. People rarely have precise numerical beliefs about probabilistic magnitudes such as the chance that it will rain on an overcast day. But people commonly do have *approximate* beliefs about such magnitudes. It is plausible, for instance, that people maintain approximate magnitude representations of quantities like the chance of rain on overcast days. Such magnitude representations may be updated on a more-or-less case-by-case basis. The magnitude underlying the representation may be decreased on overcast days with no rain, and it may be increased on overcast days with rain. The system is highly imperfect. The magnitude may be updated only sporadically. The increments or decrements to the representation may be variable even given the same prompting. The representation may stochastically decay in memory.[18] Nevertheless, the different magnitude representations may be naturally linearly ordered. So, they may serve to measure levels of support, and ground matching mappings to levels of confidence.

The function on levels of confidence that I propose can now be stated:

> **Support Matching.** Suppose that a level of confidence in p is formed on the basis of set B of beliefs. Then the level of confidence functions to match the level of conditional support conferred on p by the beliefs in B.

I submit that cases like Overcast, sans idealizations, are quite common. Suppose the chance of rain on overcast days is actually 65%. You have observed many overcast days in your city over many years, and on each day you were generally aware of whether it rained or not. Consequently, you have a general sense for how frequently it rains on overcast days in your city, one that is approximately reflective of 65%. If, on the basis of this sense

[18] My narration of the process of estimating chances from observed frequencies is inspired by the model of Gallistel et al. (2014). For background on my narration of the resulting representations as analogue magnitude representations, see Chapter 5.

alone, you became nearly certain that it will rain, that intuitively would be *overshooting.* You represent the chance of rain as middling (within the range of chances you are able to represent); the level of confidence should not be extremal. Similarly, if, on the basis of the sense of frequency alone, you became extremely doubtful that it will rain, that intuitively would be *undershooting.* Now consider the degree of confidence you would typically form if, in addition to seeing the overcast sky, you heard the radio report that it will rain today. There is an intuitive sense in which, if you formed that level of confidence after having observed the overcast sky but *before* hearing the radio report, that level of confidence would have been overshooting—but not by as much as being nearly certain would be.

I intend for these considerations to give intuitive purchase to levels of confidence and levels of support. The three different levels of confidence (almost sure, highly doubtful, same as given report) each appears to be faulty or criticizeable because it is out of line with a level of support. On the other hand, the three levels of confidence suggest how often things go *right.* Often enough, upon seeing the overcast sky, one forms a level of confidence that *isn't* overshot or undershot. Often enough, you won't become almost certain that it will or that it won't rain. Often enough, you don't form the level of confidence in rain that you typically *would* form if you were to go on and hear a radio forecast of rain.

I take the commonality with which levels of confidence and levels of support *do* match to itself count in favor of Support Matching. Observing a psychological state systematically engaged in a certain type of use is prima facie evidence that the use is a function of the state. To further bolster Support Matching, I would like to consider more carefully cases in which the levels fail to match.

When might a level of confidence fail to match the relevant level of support? The most intuitive examples of such failures involve various kinds of "pathological" influence on the formation of confidence. Consider, for example, the following case:

Rain Dream. Suppose it's morning, and I have just woken up. Before waking,

> I had an extremely vivid dream of a torrential downpour. I walk over to my window, somewhat curious about what today's weather will be like, and I see an overcast sky. I form the belief that today's morning is overcast. That belief, together with my belief that it rains on 55%-ish overcast days, triggers the formation of a level of confidence that it will rain today. However, a memory-trace of the vivid dream interferes with the process.[19] As a result, I become quite confident that it will rain today.

Here, the level of confidence is intuitively faulty. The reason it is faulty is, again, I think intuitively, that the memory-trace contributes no positive conditional support to the proposition. Suppose the memory-trace is one of the states that the confidence is formed "from" or "on the basis of." Even so, the conditional support for it will rain is the same relative to the pair of meteorological beliefs with the memory-trace and without it. The faultiness of the level of confidence stems from its not matching that level of conditional support. The faultiness does not appear to be due to any other factor. Suppose that the memory-trace represents a dream of a certain kind as having occurred. Then we may suppose that the memory-trace is both veridical and warranted. The dream transpired as represented in memory, and the memory is formed by a reliable route from the dream. In that case, the faultiness in the confidence engendered by the memory-trace does not stem from its being false or unwarranted.

It is presumably rare for dream-memories to exert this kind of substantial influence on inferential processes. There are less outre examples. Consider the Gambler's Fallacy. The Gambler's Fallacy, roughly, refers to a human tendency to think that certain outcomes in the near future are more likely, in order to "compensate" for opposing outcomes' having recently occurred more frequently than they do on average—even when the outcomes are statistically independent, and so no such "compensation" will occur. Suppose one has been at a gambling table for a long time observing repeated flips of a coin. Recently, there was a long string of

[19] Suppose, for example, that the memory-trace partly activates a channel within the formation process that typically records having observed a torrential downpour.

tails. To idealize, suppose the person believes that the sequence of coin tosses so far was: HTTHTHTHTHHTHTTHTHTTTHHTHTHTHTHHTTHTTTTTTT. We can suppose that the initial segment of data provides compelling evidence that the coin tosses are independent and fair. We may even suppose that the person has a minimal background competence for recognizing this fact. The person remembers enough of their statistical inference class to, if they worked at it, realize that the initial segment is strongly suggestive of independence. The belief about the sequence, then, supports the proposition <u>the next toss will come up heads</u> to degree ½. However, the person succumbs to the Gambler's Fallacy, and becomes 80% confident that the next toss will come up heads. They are struck by a gut feeling that heads is "due" after so many tails. We may even suppose that the person in question knows of the Gambler's Fallacy, and is sometimes able to leverage this knowledge to counteract its effects.

Here too, the level of confidence seems faulty. It will seem faulty to the person if they consider it. The confidence is based on a belief that conditionally supports the proposition to a middling degree. The level of confidence is disproportionately high. The level of confidence would not seem faulty if it was approximately 50%. The faultiness in the confidence is not explained by deficiencies in the truth value or warrant of the belief its based on. The belief about the sequence is true and warranted.

Other examples supporting Support Matching involve other kinds of deficiencies in the transition from beliefs to a level of confidence. Suppose my friend has tested positive for a disease. I believe that 95% of those with the disease test positive, that no one tests positive if they don't have the disease, and that 12% of the population has the disease. In such a situation, I am disposed to form my level of confidence that my friend has the disease by multiplying the true positive rate by the disease's base rate. Such an operation would bring my level of confidence in line with the objective rate of disease among positive tests. The result of the calculation should be 0.11. However, I am slightly dyscalculic, and in performing the multiplication I conflate 0.12 for 1/2. I consequently wind up with a confidence of 0.475 that my friend has the disease. As a result of forming this level of confidence, I back away from my friend cautiously. He chides: "What's wrong? There's only a 12% chance I actually have the disease." I recognize my mistake and revise my confidence.

I take the above examples to support Support Matching, at least as a generic claim about levels of confidence.[20]

3.4 Perceptual levels of confidence

On the model proposed in Chapter 2, each multimodal trial produces a visual state of two types. The state is of some representational type r, such that it attributes $[\![r]\!]$ to a distal particular. The state is also of IPP-signaling type v_i, which signals instantiation of a visual-width likelihood with precision p_i. The two types are closely coupled. At a minimum, the signaling type determines the causal influence of r on the multimodal representation. More robustly, as we suggested, the signaling type may be a sub-type of r. It marks instances of r as having a certain type of mode of presentation, one that, additionally, functions to signal an IPP.

I propose that, if visual width representation has content r and IPP-signaler v_i are produced on a trial, the IPP-signaling state constitutes a level of confidence in the representational content r.

The IPP-signaling states are naturally linearly ordered. They may be linearly ordered by the precision of the likelihood property they signal. On the ordering, an IPP-signaling

[20] Support Matching generalizes various norms that have been proposed in formal epistemology. The example of Overcast is an instance of an "inference" pattern known as direct inference: from the belief that x is P and the belief that n% of P's are Q's, form confidence n% that x is Q. In one form or another, this pattern is widely seen as a norm on the formation of confidences. (Of a similar rule, Reichenbach 1949 adds the proviso that P must be the narrowest reference class to which x belongs for which you have reliable statistics about Q. Salmon (1971) and Kyburg (1974) adds several modifications to this proviso. Levi (1977) adds the proviso that one must believe x to have been chosen randomly from the class of P's. All, however, take some form of the pattern to be normative on the formation of confidence.) Direct inference forms the inspiration for Lewis' (1980) enormously influential Principal Principle. The principle encodes a norm demanding that one's confidence in p be equal to the objective chance that one believes p to have. Various descendants of the Principal Principle are endorsed today in formal epistemology.

state $v_i \leq v_j$ if $p_i \leq p_j$. Such an ordering also corresponds to degreed differences in the causal powers of the two signaling states. In Chapter 2 we introduced an abstract notion of the "level of pull" that an IPP-signaling state exerts on the multimodal representation. That notion was an abstraction meant to compactly summarize the various more specific causal powers of the state. So, equivalently, we can linearly order IPP-signaling states by their levels of pull.

If the IPP-signaling states constitute different levels of confidence in representation r, given the discussion of Chapter 3, we would expect that greater IPP-signaling states lead to greater reliance on the content r. Consider two IPP-signalers, v_i and v_j, such that $v_i \leq v_j$. Suppose the haptic IPP-signalers and unimodal representations are fixed. If the IPP-signaling state is v_i, the multimodal representation will be r_1, and if the IPP-signaling state is v_j, the multimodal representation will be r_2. Under the model we proposed, r_2 will be proportionately closer to the visual representation r_v than r_1 is.

Reliance is a metric that specifies dependence between the successfulness of the outcome of a psychological process and the accuracy of a representation. Here, the relevant psychological process is that of producing multimodal width representation. The outcomes are the possible multimodal width representations. The function of the outcomes, and the process, is to produce a width representation that is as accurate as possible. Thus, the successfulness of the multimodal outcome may be measured by a loss function. Production of multimodal representation r_m is perfectly successful on an occasion if the loss function given r_m and the actual distal width is 0. The representation is correspondingly less successful as the loss function increases. The visual width representation may be evaluated in a similar way. Its level of accuracy is also judged by a loss function.

In what way does the representational success of r_m depend on the representational success of r_v? Let us say that a multimodal width representation r_m is *approximately accurate* if $|[\![r_m]\!] - w| \leq \varepsilon$, supposing that w is the actual distal width. Suppose that r_m is approximately accurate. This supposition has entailments for how accurate the visual representation

r_v could be. Suppose that $r_m < r_v$. Suppose $w' = \max_w\{w \mid |[\![r_m]\!] - w| \leq \varepsilon\}$. Then r_v's level of inaccuracy is at least $|[\![r_v]\!] - w|$. If r_m is quite far from r_v, then r_m does not heavily rely on the accuracy of r_v. Fulfillment of r_m's function does not entail that r_v is especially close to accurate. Nor does it entail that r_v fulfills *its* representational function. By contrast, the closer r_m is to r_v, the more accurate r_v must be if r_m is to be approximately accurate. In that case, r_m relies more heavily on the accuracy of r_v.

Thus, from the fact that $v_i \leq v_j$, it follows that the multimodal representation produced under v_j relies more heavily on the accuracy of r_v than it does when produced under v_i. Similar consideration apply to any pairs of IPP-signaling states. Thus, the IPP-signaling states satisfy Reliance.

What about Support? Suppose that a visual representation r_v is "substantially inaccurate" if $|[\![r_v]\!] - w| \geq \epsilon$. It follows that a visual representation produced in a high-noise condition has a higher objective probability of being substantially inaccurate than a visual representation produced in a low-noise condition. On any occasion, the noise condition and distal width help determine the type of proximal stimulation. Features of the proximal stimulation are reliable indicators of noise condition. Certain profiles of dot disparity will make noise condition n_1 overwhelmingly likely, while other profiles will make n_2 most likely, and so on. Insofar as the proximal stimulation is registered correctly, the proximal registration is also a reliable indicator of noise condition. Transitively, the proximal stimulation is a reliable indicator of the visual representation's probability of substantial inaccuracy.

We can thus define a metric of conditional support between proximal stimulus registration and visual width representations. Suppose the proximal stimulus registration faithfully registers the proximal stimulus. If so, the objective conditional probability of substantial inaccuracy in r_v is fixed. Faithfulness of the proximal registration implies that the current instantiation of $\mathbb{D}(W \mid R_v, N)$ has a certain precision, which precision in turn fixes the probability of inaccuracy. Probability of substantial inaccuracy will provide our metric of support for a visual representation r_v.

We can thus linearly order proximal registration types by the extent to which they conditionally support a visual representation r_v. Under the support ordering $\leq_s$ (relative to r_v), for proximal registration types x_1 and x_2, $x_1 \leq_s x_2$ if and only if the probability of r_v's substantial inaccuracy is lower given x_2's faithfulness than it is given x_1's faithfulness.

Support requires that, if states v_i and v_j are levels of confidence in r, and if $v_i \leq v_j$, then typically, v_i is produced from proximal registrations that support p less than the typical proximal registrations leading to v_j. This condition is satisfied. If a high-support registration occurs, then most likely the noise condition is low-noise—say, n_j. And if the noise condition is low-noise n_j, then, most likely, v_j is the IPP-signaling state that will occur. Thus, when r_v has a high degree of support, v_j is likely to occur. Conversely, if v_j occurs, then, most likely, the noise condition is n_j, which, in turn, will most likely lead to a high-support registration. So, if v_j occurs, r_v is likely to have a high degree of support. The same considerations show that there is a lower level of support for r_v that tends to occur with IPP-signaler v_i, where $v_i < v_j$.

The probability signaling states help us to see how perceptual systems could have levels of confidence. The notion of reliance is, as discussed above, intuitively associated with confidence. It is also not hard to see how patterns of reliance on representational states might apply within perception. However, the notion of support is also, as I have argued, intuitively associated with confidence. The notion of confidence is paradigmatically associated with notions of evidence. It is less obvious how a perceptual system could ever be oriented towards a body of information (i.e. sensory states) in a way that is relevantly like the orientation one takes to a body of evidence at the level of belief. The link to objective probability—and specifically, to objectie probabilities of accuracy and inaccuracy—shows how this is possible.

By showing how perceptual systems can be oriented towards bodies of information in roughly similar ways to our orientation towards evidence, we explain how a perceptual system can have credal states. The notion of credal state, presupposed by realism about Bayesian models in perceptual psychology, is conceptually tied to such orientation. Thus, the account

I offer removes one potential hurdle for a moderate realism about Bayesian models. The account shows that, and how, perceptual states are credal states. A moderate realism about Bayesian models may require more. A moderate realism may require different sorts of credal states. It might require that there be credal states corresponding to prior probabilities, and it might require that those states interact in certain ways with credal states corresponding to likelihood probabilities—that is, with the perceptual states defended here. This is a fruitful avenue for further research. But the existence of credal states in perception already pushes against instrumentalism. Instrumentalists do not object to credal states corresponding to prior probabilities in particular. They object to the existence of credal states that operate in a manner roughly consistent with Bayesian transitions. An argument for the existence of perceptual credal states undercuts part of that objection.

CHAPTER 4

Analogue magnitude representations and objective probability

Humans and many animals can rapidly discriminate collections on the basis of how many elements those collections contain. They can do this for visually-presented collections, auditorily-presented collections, and even collections that consist in a series of actions that they themselves have performed. The range of this ability is, in general, surprisingly large: many animals can reliably distinguish 8 from 10, 15 from 18, 22 from 28, and beyond. They accomplish this without consciously counting. The ability is approximate and seems specific to number (as opposed to closely correlated properties). It is employed in behavior when number is important, as when a monkey must choose the more plentiful tree to climb. The ability is deemed "numerosity discrimination," and the size of a collection as tracked by this ability is deemed its "numerosity."

To explain this discriminative capacity, psychologists posit an internal system that tracks numerosities. This system is called the "Approximate Number System" (ANS). The properties of this system are increasingly well-characterized, and its neural underpinnings are being dutifully explored. The role this system plays in the development of mature human arithmetical abilities is intensely studied. The system comprises states that are differentially responsive to the numerosity of a collection of objects. Moreover, these states are widely taken to be representational. There is, however, much unclarity in the literature about what properties the states of the ANS represent. Many take the states to represent numerosities—that is, to represent the exact cardinalities of object-collections. Some take the states to represent ranges of numerosities. Others hedge, using terms like “7-ish” or “approximately-7” to denote the properties represented by the states. Although there is widespread agreement *that* these states have semantics, it is unclear *what* semantics they have.

My aim in this paper is to systematically lay out the different semantics one could assign to the ANS and evaluate them in light of the empirical and modeling literatures. I will argue that no state of the ANS represents numerosities, or ranges of numerosities, or partial numerosities. I also argue that, prima facie, no state represents the ratio between the numerosities of two collections. Instead, I claim, states of the ANS represent collections as being more numerous, less numerous, or equally numerous. Here is the plan. In §1, I survey the behavioral data on numerosity discrimination. In §2, I introduce the most widespread model of the ANS, what I call the log-Gaussian model. In §3, I discuss the intra-psychological relations of the ANS. I also systematically lay out the different semantics one might assign to states of the ANS, and I clarify the methodology by which I will evaluate them. In §4, I describe in detail the most obvious semantic proposal: that the ANS represents numerosities. I argue that this proposal is empirically unmotivated. In §5, I argue against partial numerosity semantics. In §6, I consider the proposal that the ANS represents ratios. In §7, I defend the view that the ANS represents comparative numerosity.

4.1 The Behavioral Signatures of Numerosity Discrimination

Consider the behavioral task of Nieder and Miller (2003). On an individual trial, a monkey is visually presented with a *sample* display containing some number of dots. The sample display is shown for 800ms, and then after a one second delay, the monkey is shown a *test* display, containing either the same number of dots as the sample display, or a different number of dots. If the number of dots in the two displays match, the monkey must release a lever to receive a juice reward; otherwise it must wait until a second test display, guaranteed to match the sample, to release the lever. Releasing the lever when the sample and test displays do not match constitutes an incorrect response. Figure 1 illustrates a pair of stimuli that could be shown on a trial.

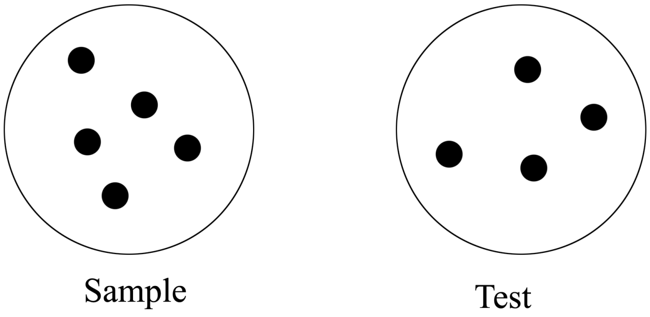


Figure 4.1: Illustration of a stimulus in Nieder and Miller (2003). Here, the correct answer is "different."

Following T. Burge (1977), I will say that this trial presents two different **aggregates** of dots. I call them aggregates rather than sets to honor the fact that the collections of dots in the displays have spatiotemporal locations, whereas sets, being abstracta, do not. And I will say that the two aggregates differ in their **numerosity**: the sample display has a numerosity of 5, while the test display has a numerosity of 4. Intuitively, numerosity is for aggregates what cardinality is for sets.

Nieder and Miller had monkeys perform this discrimination task, with sample displays ranging in numerosity from 2 to 6 dots, and with test displays differing from sample displays by as many as 5 dots. Figure 2 demonstrates monkeys' performance on this task.

The darkness of a curve denotes the number of dots in the sample display, and the x-axis indicates the number of dots in a test display. For example, when the sample display contained 6 dots and the test display also contained 6 dots (peak point on the darkest curve), monkeys correctly responded with "same" roughly 85% of the time. But when the sample contained 6 dots and the test contained 7 dots, monkeys *incorrectly* responded with "same" roughly 70% of the time. When the sample contained 6 dots and the test contained 8 dots,

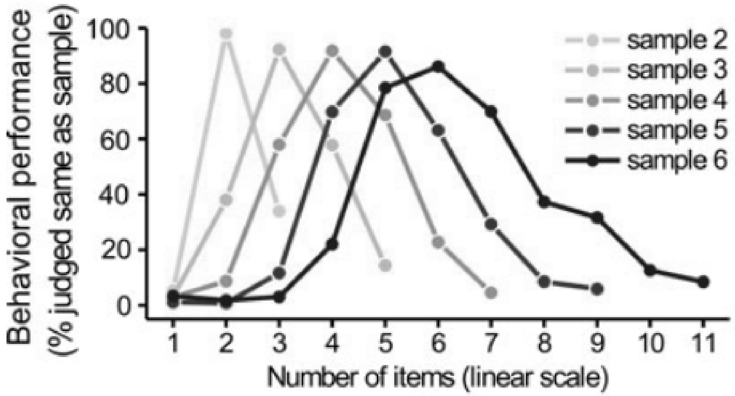


Figure 4.2: Behavioral results of Nieder and Miller (2003), as redrawn in Nieder (2013).

monkeys' error rate dropped to 40%.

Figure 2 illustrates two facts about monkey performance on this same-different task. First, performance shows the **distance effect**: given a particular numerosity N, the reliability with which the monkey can distinguish N from a distinct numerosity M increases as the numerical distance between N and M increases. Notice from Figure 2 that monkeys are fairly unreliable at distinguishing 6 from 7 dots—they can only do it about 30% of the time. But they are more reliable at distinguishing 6 from 8 dots—they can do this about 60% of the time. And they can distinguish 6 from 10 dots 85% of the time. Second, performance shows the **magnitude effect**: given a fixed numerical distance $|N - M|$ between two numerosities, the reliability with which the monkey can distinguish N from M decreases as N increases in magnitude. Let's take our numerical distance to be 1. Figure 2 shows that monkeys distinguish 2 dots and 3 dots fairly reliably (they can do it about 65% of the time). They distinguish 3 dots from 4 dots less reliably (they succeed about 40% of the time), and they distinguish 4 from 5 dots less reliably still (only about 30% of the time). The distance and magnitude effects are nearly universally observed in numerosity discrimination tasks. Every numerosity discrimination study I discuss contains an observation of these effects, unless otherwise noted. Although I shall not focus on it in this paper, response times in many numerosity discrimination tasks also exhibit analogues of the distance and magnitude effects. The rapidity with which a creature can distinguish N from M increases as their numerical distance increases, and, for a fixed distance $|N - M|$, diminishes as the N increases in magnitude.

The sample and test displays illustrated in Figure 1 differ in their numerosities, but they also differ in other respects. The sample display is, overall, less bright than the test display. The sample display is also more densely packed with dots than the test display. One could object, then, that the monkeys in the task were responding not to numerosity but to brightness or visual density. To test this objection, Nieder, Freedman, et al. (2002) had monkeys perform the same task with sets of stimuli that systematically varied such properties as visual density, shape and size of aggregate elements. Figure 3 illustrates such stimuli.

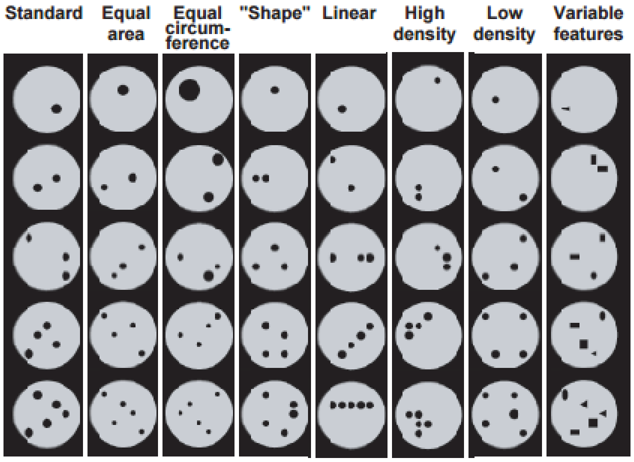


Figure 4.3: Alternative stimuli from Nieder et al. (2002)

Consider the aggregates in the "equal area" column of Figure 3. If monkeys respond on the basis of total inked surface area, then all pairs of stimuli drawn from the "equal area" column should be judged "same" with roughly identical frequencies. However, if monkeys respond on the basis of numerosity, then each pair from the "equal area" column should be judged "same" with different frequencies. Moreover, if the pair comprises aggregates with numerosities N and M, it should be judged "same" with the same frequency reported for

N and M in Figure 2. This is what Nieder, Freedman, et al. (2002) found for the "equal area" column, and indeed, all other columns in Figure 3. For example, monkeys conflated 6-membered and 7-membered aggregates 70% of the time, regardless of their sizes, their shapes, and the arrangements of their component members.

As I shall say, monkey performance displays **response invariance** to non-numerical properties of stimuli. An animal exhibits response invariance if, in any numerosity discrimination task, the statistics of the animal's responses to aggregates with numerosities N and M remain the same, regardless of what other properties those aggregates have. So stated, response invariance is too extreme to ever be literally true, for both trivial reasons (what if one of the aggregates is invisible?) and empirical reasons (e.g. humans are known to systematically underestimate numerosity of large aggregates when they are more densely packed—see Dehaene (2011, p. 60)). Nevertheless, studies in numerosity discrimination routinely involve controls like those of Nieder, Freedman, et al. (2002), and routinely find that subjects attain response invariance to a high degree.[1] Certainly, discrimination performance is largely invariant across changes in the most obvious continuous properties of aggregates, such as the cumulative visible surface area of their members.

Visual numerosity discrimination experiments like the one just reviewed have been conducted on other organisms, including pigeons, rats, crows, infants, and adult humans. Auditory numerosity discrimination experiments have also been run, in which subjects must discriminate aggregates of tones on the basis of how many tones were presented. Response invariance is attained, as revealed by controlling for total duration of the aggregate, total

[1] Response invariance is the primary reason psychologists postulate numerical representations. The thought is that, since animals can differentiate numerical properties as such, they have capacities that represent numerical properties in particular. However, the adequacy of the control experiments supporting response invariance, and consequently the foregoing motivation for postulating numerical representations, has been questioned. See, for example, the essays in Henik (2016), especially Chapter 16. Some authors, for instance, think that "numerosity discrimination" is actually just a form of perception of texture density (see Dakin et al. (2011)). Such skeptics are in the minority, but the issues are subtle and complicated. I shall henceforth assume that the degree of response invariance observed in the empirical literature motivates postulation of some sort of numerical representation, albeit defeasibly.

duration of each tone, total duration of inter-tone interval, and other properties. Similarly, experiments have been conducted in which a rat or a pigeon must enter a food-tray only after having executed some definite number of actions (such as pecks or lever-presses). Such experiments control for total amount of time spent pecking, amount of energy expended in pecking, etc. Such controls again reveal response invariance. Moreover, indices of discrimination performance from visual and auditory tasks tend to agree, suggesting that these behaviors are subserved by an amodal system that receives input from multiple sensory and motor systems. Across species and tasks, performance in these numerosity experiments manifests the distance and magnitude effects. In the experiment of Figure 2, the largest numerosity presented was 11, but other studies have assayed discrimination performance on numerosities as large as 24, 30, or 50.

Response invariance, where observed, suggests that animals harbor psychological systems differentially responsive to the numbers of members in aggregates as such. The magnitude and distance effects suggest, however, that this responsiveness is only approximate. Hence, psychologists postulate an internal psychological system called the "Approximate Number System" (ANS).[2] Such a system is taken to drive animals' performance in numerosity discrimination tasks like the ones just discussed. Moreover, there is widespread agreement in the numerosity literature about how to model the ANS. We turn to this model now.

4.2 The Log-Gaussian Model of the ANS

There is widespread agreement about a mathematical model of the ANS, which I will call the log-Gaussian model (LGM). According to the LGM, the ANS comprises a continuous infinity of distinct state-types. I will call this set of state-types of the ANS the *state space.* The LGM models the state space of the ANS as the set of positive real numbers, $\mathbb{R}^+$. Hence,

[2] Also called the "large number system" or, more poetically, "the number sense."

according to the LGM, the state space has the structure of $\mathbb{R}^+$.[3] In particular, the state-types of the ANS are linearly ordered, satisfy a Euclidean distance metric, and ratios may be defined between them. I will use 'R' to denote the state space of the ANS, and lower-case r, sometimes with a subscript, as a variable ranging over individual state-types in the state space. Because the state space has the structure of $\mathbb{R}^+$, I will sometimes refer to a state-type by the real number it maps to. But the state-type is not itself a real number. Similarly, I will sometimes say that two ANS states r_1 and r_2 are "close" to one another. Again, this refers to a relational property of the ANS itself, albeit one mirrored by the reals.

Suppose a creature attends to an aggregate with a numerosity of N. Presentation of this numerosity causes an event: the creature's ANS now enters into a state of type r. But this causal route from N to r is stochastic. Suppose the animal looks away, and then back at the same aggregate. Now, the presentation of N causes the ANS to enter into a *different* state, $r' \neq r$. In general, different state-types have different probabilities of occurring in response to an N-membered aggregate. The LGM models this fact with a conditional probability distribution over the state space. This distribution—which I will call $P(R|N)$—specifies, for each state-type r, the probability that the ANS enters into state r, given that the numerosity of the perceived aggregate is N.[4] Furthermore, for a specific r, the probability of r's instantiation will, in general, vary with the presented numerosity: r may be very likely given three dots, but extremely unlikely given thirteen. In this case, $P(r|3) >> P(r|13)$. Note that these distributions specify *objective* probabilities, and not subjective

[3] There is no greatest element in $\mathbb{R}^+$, but there is a "greatest" state the ANS can enter. So in fact, the LGM models the ANS as a proper initial interval of $\mathbb{R}^+$, the upper bound of which may be regarded as a free parameter of the model. Additionally, it is not plausible that the ANS—or any psychological system—has a genuinely uncountable number of states. In reality, the ANS has finitely many states—it's just that there are *lots* of them, and they are tightly packed. So it would be better to say that the ANS comprises a quasi-continuous set of states.

[4] Note two abuses of notation that I employ here and throughout. First, I abuse the distinction between a probability and probability density. Strictly speaking, a distribution assigns to each state r a probability *density*, not a probability. Second, on my usage, '$P(R|N)'$ denotes a probability distribution, while '$P(r|N)$' denotes a particular probability. $P(R|N)$ does not denote the probability that *some* state in R occurs in response to N, a probability the LGM models as always equal to one.

ones. They specify the objective probability that an event of a certain psychological type will occur, given that some other event has occurred.

According to the LGM, each distribution $P(R|N)$ is Gaussian. A Gaussian distribution is uniquely defined by specifying its mean and standard deviation. According to the LGM, the mean μ of $P(R|N)$ is a logarithmic function of N. In particular, the mean μ of $P(R|N)$ is $\log(N)$, where 'log' denotes the natural logarithm.[5] So, the mean of $P(R|3)$ is $\log(3)$. This means that, on average, a three-membered aggregate will tend to trigger states "close" to the ANS state corresponding to the real number $\log(3)$. I will use 'μ_N' to denote the mean of the distribution $P(R|N)$. While the mean of $P(R|N)$ depends on N, the standard deviation σ of $P(R|N)$ is the same for all distributions, and may be regarded as a parameter of the model. Of course, the actual value of σ depends on the characteristics of the ANS and the psychology in which it is embedded.

Figure 4 illustrates the LGM. Notice how, as N increases, the conditional distributions "bunch up" and overlap more and more. As we shall see, this is the basic source of the model's ability to predict the distance and magnitude effects.

The LGM models the ANS at a very high level of abstraction. This has two consequences. First, there are alternative models that are mathematically distinct but make identical (or near-identical) predictions. For example, one often finds discussion of the "scalar variability" model of the ANS. Like the LGM, this model postulates Gaussian conditional distributions over a state space modelled as $\mathbb{R}^+$. Unlike the LGM, however, it holds that the mean of $P(R|N)$ is simply N, and its standard deviation is $\sigma = N * w$, for some small fraction w.

[5] More generally, the mean of $P(R|N)$ is $\mu = a\log(N) + b$. I follow Dehaene (2007) in letting $a = 1$ and $b = 0$. These may be regarded as free parameters of the model. Note a technical point, however. The LGM models the state space as $\mathbb{R}^+$, but technically, a Gaussian distribution must be defined over $\mathbb{R}$. This does not impact us too much, because, for most of the Gaussians we consider, the probability that R takes a negative value is vanishingly small. However, given what I have said in the text, the mean of the distribution $P(R|1)$ is $\log(1) = 0$. In this case, $P(r < 0|1) = 0.5$. Hence, we should really set b to a positive value, to shift the distributions over, ensuring that r's being negative given numerosity 1 is vanishingly rare. To simplify matters, I do not do this, pausing only to point out that my arguments generalize entirely to this more realistic case.

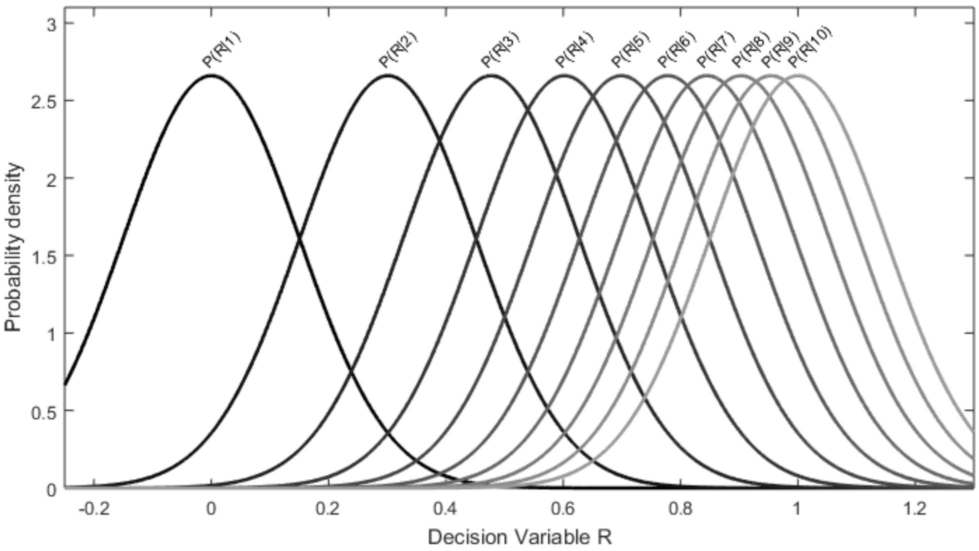


Figure 4.4: The LGM illustrated, with $\sigma = 0.15$.

Like the LGM, the conditional distributions overlap more as N increases. Consequently, the two models make near-identical predictions.[6] Second, the LGM is silent about the processes that lead from proximal sensory input to activations of ANS states. There are, however, several proposed models of these processes in the literature.[7]

The LGM assumes that, when confronted with two numerosities N and M simultaneously, two distinct state-occurrences are elicited "in" the ANS concurrently. With vanishingly small probability, these states will be of the same type; more likely, they will be of different types. When the LGM is deployed to predict the results of a particular type of numerosity discrimination task, it is supplemented with a formal *decision rule* that, in the simplest case,

[6] As Roitman and Brannon (2003, p. 171) note, these models, although mathematically distinct, make extremely similar predictions, and are often treated as "functionally equivalent." Dehaene (2007) calls them "essentially identical" in their empirical predictions. There has been dispute about the extent to which this is so. Brannon, Wusthoff, et al. (2001) claim that only the scalar variability model can accommodate their experimental results; Dehaene (2001) replies by providing a simulation of a neural network implementing LGM that predicts their data.

[7] Two classic models are Dehaene and Changeux (1993) and Meck and Church (1983). The former is most naturally seen as implementing the LGM, while the latter is most naturally seen as implementing the scalar variability model. There are, however, mechanistic models that make distinct predictions from both LGM and SVM. See, for instance, Zorzi and Butterworth (1999).

maps pairs of state-types to behavioral responses. For example, suppose aggregates A and B are presented simultaneously on a trial, and they must be judged "same" or "different." Suppose the numerosities of these aggregates caused occurrences in the ANS of types r_A and r_B respectively. Following the treatment of same-different tasks elsewhere in psychophysics, we may supplement the LGM with the following decision rule: respond "same" iff $|r_A - r_B| < \delta$, where δ is a *decision criterion.* The exact value of the decision criterion could reflect architectural facts about the creature's psychology, or it could be under the creature's intentional control (as it may be if the creature is exceptionally wary of getting the wrong answer). This decision rule reflects a sensible policy: respond "same" only if the elicited state-types are close together. The farther apart they are, the less likely they are to have derived from the same numerosity. If this decision rule is found empirically successful, we can say that it corresponds to a causal law of the creature's psychology, roughly: sufficiently close ANS states cause "same" responses.

Later, I will explain in more detail how the LGM predicts the behavioral facts about numerosity discrimination. For now, simply note the following. First, response invariance is simply assumed by the model. The probability of a state r is conditional *only* on distal numerosity. Second, the distance effect is predicted because, as M gets closer to N, it becomes likelier to cause an internal response r_M that is very close to the internal response to N—and hence, discrimination becomes harder. Third, the magnitude effect is predicted, because for a fixed numerical distance $(N - M)$, the distributions $P(R|N)$ and $P(R|M)$ get closer as N increases, once again making it likely that N and M will elicit very close internal responses.

4.3 Psychological, semantic, and methodological preliminaries

Aggregates and aggregation

Numerosities are akin to cardinalities. Like a cardinality, a numerosity is a property that can only be had by a collection of individuals. Whereas cardinalities apply to sets, I have said that numerosities apply to aggregates. The intuitive notion of an aggregate is simple. An aggregate is simply a collection of physical particulars (including physical events). An aggregate may be regarded as a particular in its own right. Aggregates have spatial locations, undergo change, and exert causal influence. An aggregate may have properties that none of its components have, as when a flock of birds takes on a spherical shape. A collection of components may make up an aggregate and, separately, constitute an individual. All of the particles in a planet constitute the planet, and they are also the components of an aggregate, but the planet and the aggregate are distinct. They have, for instance, different persistence conditions: annihilation of any one component annihilates the aggregate, but leaves the planet intact.

I will assume that any creature with an ANS has some perceptual representational capacity to refer to aggregates and to characterize them *as* aggregates. Equivalently, using the notation of T. Burge (2010b, p. 25ff.), we can say that such creatures have perceptual states with the representational content $\underline{\text{(that } x_1\text{) aggregate(}x_1\text{)}}$.[8] I will call such a state an aggregate representation.[9]

I make two further assumptions about aggregate representations. First, I will assume that

[8] A word about the notation. Underlined expressions denote representational contents. Specifically, they denote a *type* of representational content. The 'that(x_1)' denotes a general ability to refer demonstratively to an environmental entity. For a psychological *event* with this content, the 'x_1' marks an occurrent exercise of this general ability that (if successful) secures an environmental referent. 'aggregate(x_1)' denotes the fact that this content attributes aggregatehood to the environmental entity referred to by x_1, if any.

[9] In future work, I hope to more carefully address what distinguishes singular reference to an aggregate from plural reference to its components. I also hope to address the empirical conditions and limits on aggregate representation.

perceiving an aggregate entails perceiving each of its components. That is, if an aggregate is successfully referred to by an aggregate representation, then each component of the aggregate is successfully referred to by some perceptual representation. If a creature perceptually refers to an aggregate of three dots, then for each dot, there is a representation referring to it. For instance, we would expect to find the representations $\underline{\text{(that } x_1\text{)dot(}x_1\text{)}}$, $\underline{\text{(that } x_2\text{)dot(}x_2\text{)}}$, and $\underline{\text{(that } x_3\text{)dot(}x_3\text{)}}$. My second assumption is that the representations referring to components of the aggregate are in some way *linked* to the representation referring to the aggregate. A representational linkage would suffice: linkage to $\underline{\text{(that } y\text{) aggregate(}y\text{)}}$ would be secured if each component dot i of the aggregate were represented by

$$\underline{\text{(that } x_i\text{) dot(}x_i\text{) component-of(}x_i, y\text{)}}$$

. But, for the purposes of this paper, non-representational linkage would suffice as well. To take an analogy from computer programming, it would suffice for linkage if pointers to the variables x_1, x_2, x_3 were contiguous with the memory location of the variable y.

These two assumptions enable us to amend the LGM's account of the ANS slightly. Component representations, and the aggregate representation to which they are linked, causally drive activity in the ANS. If something is not perceived *as* an aggregate, its numerosity will not be estimated by the ANS. In the previous section, we said that the probability of state-activation r is conditional on the numerosity of the perceived aggregate. This formulation only applies when an aggregate *is* perceived. But we can imagine that a perceiver in a pitch-black room may hallucinate an aggregate of three dots, and that this hallucination may well trigger activations in the ANS with the same statistics as seeing a real three-membered aggregate.[10] In this case there are no environmental aggregates by assumption, so according

[10] This is a fanciful case for illustration. A more serious example comes from Aida et al. (2015). This study utilized a numerosity discrimination task like the ones discussed in §1. The stimuli were aggregates of dots hidden in random-dot-stereograms. With anaglyph glasses, aggregates of dots appeared to hover over the page, and their numerosities were discriminated. Here, there are no environmental aggregates (or at least, none whose numerosities the subjects are discriminating). However, this study did not test for magnitude or distance effects.

to the LGM, the probabilities of all state-activations are undefined. We can account for this case by amending the LGM's definition of the conditional distributions: the probability of state r is conditional on the *number of component representations* linked to the causally driving aggregate representation.

This amendation does not, in itself, require us to postulate representations of numerosity. That 10 dot-representations should be linked to an aggregate representation is a fact about the creature's psychology at a time, but of course, not all facts *about* one's psychology are represented *by* one's psychology. Furthermore, the *process* by which an aggregate representation is formed and the component representations are linked to it need not require enumeration. Suppose I look at a stack of books. I may first parse the complex shape of the overall stack, assigning it certain boundaries and location. Then, parsing processes at finer spatial resolution may begin generating representations of the component books. As the spatial locations of books are found to relate systematically to the spatial elements of the stack-shape, the stack becomes represented as an aggregate. The individual books may then be linked to the aggregate by laws of spatial inclusion: where the book is represented as having spatial location and boundaries that fall within the stack-shape, link it to the aggregate.

Intrapsychological relations of the ANS

Again following Burge's terminology, I will assume that, if any states of the ANS have representational contents, those contents are primarily attributive. An occurrence of a representational state of the ANS then functions to attribute a property or relation to some environmentally perceived entity or entities. Again using Burge's terminology, if an ANS state functions to attribute a property to an entity, I will say the state *indicates* that property. If an ANS state indicates a property, it must be a property that aggregates can have. Similarly, if it indicates a relation, it must be a relation between aggregates.

Aggregate representations (together with the component representations linked to them) stochastically cause activations of ANS states. I assume that an ANS state-activation is *bound* to the aggregate representation that caused it. For example, suppose the representation $\underline{\text{(that } x_1\text{) aggregate}(x_1)}$ causes activation of state r. Then r is bound to x_1 in the psychology. If r has an attributive content, then attribution of its indicated property to x_1 would suffice for binding. But binding would occur even if r had no representational content at all. And similarly, if, simultaneously, $\underline{\text{(that } x_2\text{) aggregate}(x_2)}$ caused activation of state r', then r' is linked to x_2 in the psychology. This is required to account for the production of behavioral response via the ANS. If r and r' were untethered to the discriminated aggregates, the creature would not have a basis for pointing to one aggregate as the larger or smaller.

The ANS interfaces with memory mechanisms. Recall the same-different task of §1. In order to provide a response, the monkey must compare two aggregates across a temporal delay. There are different models of how it might achieve this comparison. One model is that the ANS state activated by the earlier aggregate is stored in memory, and is retrieved and compared to to the state activated by the later aggregate. This introduces further parameters into our model. For instance, once r has been activated and stored, does it undergo further stochastic garbling in memory? If so, with what distribution? Another model is that the earlier aggregate representation and its linked component representations are stored. When a comparison is required, these representations are retrieved, and stochastically cause a new ANS state activation. This also adds further parameters to our model.

The semantic question and methodology

Assuming that the ANS harbors some representational states, our semantic question has three components. First, which kind of ANS state has representational content? Second, what class of properties or relations are indicated by these content-bearing states? Third, what is the exact mapping from the set of content-bearing states to the class of indicated properties?

Semantic proposals for the ANS can be divided into two families, depending on how they answer the first question. Imagine that a creature perceives exactly one aggregate. The single aggregate representation causes the ANS to enter into a state of type r. We called the set of possible state-types the state space of the ANS, or R. Call the members of R the *atomic states* of the ANS. Now suppose that the creature perceives exactly two aggregates. The two aggregate representations cause two state-activations in the ANS of types r and r'. In other words, the ANS enters into a *complex state*, which we may call (r, r'). The set of complex states of the ANS is $R \times R$.[11] A **monadic semantics** for the ANS holds that some atomic states have representational content. This family of semantic proposals holds that sometimes, an occurrence of exactly one atomic ANS state r carries representational content. A **relational semantics** for the ANS holds that *no* atomic state has representational content, but that some complex states do. This family of semantic proposals holds that, so long as the creature is perceiving exactly one aggregate, no state of the ANS is representational; but sometimes, when the creature perceives two aggregates, a complex state of the ANS represents a relation between the aggregates.

I will use '$[\![\cdot]\!]$' to denote a function from states of the ANS (either R or $R \times R$) to representational contents. So, monadic semantics holds that for some r, $[\![r]\!]$ is defined. Relational semantics holds that for no r is $[\![r]\!]$ defined, but for some (r, r'), $[\![(r, r')]\!]$ is defined.

In this paper, I will consider four semantic proposals (two from each family) that differ on the second question, the class of properties and relations indicated by states of the ANS. The two forms of monadic semantics that I shall consider are numerosity semantics and partial-numerosity semantics. The former claims that atomic ANS states represent the numerosities of aggregates. The latter claims that atomic ANS states represent partial numerosities, which are the properties denoted by such natural language expressions as "two and a half bagels." The two forms of relational semantics I shall consider are ratio semantics and comparative semantics. The former claims that complex ANS states represent the ratio

[11] In general, if the creature perceives exactly M aggregates, then it enters into a state $(r_1, \ldots, r_M)$. I count these states as complex as well. So generally, a state s of the ANS is complex iff $s \in \cup_{k=2} R^k$. Intuitively, a state is complex if it is not simple.

between the numerosities of two aggregates. The latter claims that complex ANS states can only represent one aggregate as having a larger, smaller, or same numerosity as another aggregate.

Specifying the content-bearing states of the ANS and the class of properties they indicate leaves open the exact mapping from states to properties. Consequently, each of the four semantic proposals comes in many flavors, corresponding to the different specific mappings of states to properties.

Any specific mapping specifies a semantic individuation of states of the ANS. For instance, consider a (rather implausible) mapping according to which every atomic state in R is mapped to the representational content one-component(x) (which indicates a property had necessarily by all and only one-membered aggregates). On this view (a version of numerosity semantics), activation of any state r in the ANS is an event with an attributive representational content. Hence, when r activates, the creature enters into a complex overall psychological state with the content (that x_1) aggregate(x_1) one-component(x_1). I assume that representational contents mark causal powers of the states that have them. So there must be some causal law of the creature's psychology in which the state-type one-component(x) figures as part of the antecedent. Since, on the implausible mapping, each r is a state with this representational content, the activation of any such r is an event that satisfies the antecedent of the causal law. Hence, there must be some relevant event-type E that is entrained by activation of any ANS state.[12] But we have no empirical evidence to think that there is any such causal law linking mere activation of the ANS to a single effect. There is, apparently, no behavior of the animal that is common to all individual presentations of numerosities. Hence, we have no reason to postulate such a causal law, and hence no reason to postulate

[12] I keep 'relevant' loose, but two assumptions. I assume that E must be individuated psychologically, not (e.g.) neurally. And I assume that E must, in some very loose sense, be specifically and reasonably related to its antecedent. It may be that every ANS state r helps to causally produce the belief "I am answering the numerosity discrimination task right now." But if that were the only common effect of each r, then, since the content of this belief is not specifically related to representation of numerosity-one, it would not suffice to ascribe the content one-component(x) to each r.

a representational content that marks the presence of such a law.

In this way, the empirical evidence constrains the admissible semantics. Although the previous example is of course implausible, I will retain the same methodology in what follows. For each of the four semantic proposals, I will state its most plausible specific mapping. I will then ask whether there are empirical phenomena that are well explained by casting them as the consequents of causal laws whose antecedents embed the representational states as defined by the mapping. If there are such empirical phenomena, then that semantic proposal has prima facie evidential support. If not, then the semantic proposal is rejected as unmotivated. I begin with the most obvious semantic proposal: the ANS represents numerosities.

4.4 Against numerosity semantics

According to numerosity semantics, atomic states of the ANS represent numerosities. Intuitively, this means that the ANS can represent an aggregate as having one component, or two components, or three components, and so on. But it is not entirely clear what properties should be denoted by the term 'numerosity.'

However we understand numerosities, they are properties that bear an intimate connection to cardinalities. Specifically, an aggregate has numerosity n iff its components are in one-to-one correspondence with the nth von Neumann ordinal. One could upgrade this biconditional to an identity, and say that having numerosity *two* just *is* the property *being in one-to-one correspondence with the second von Neumann ordinal.* We could then say that some states of the ANS have the representational content "in one-to-one correspondence with the second von Neumann ordinal (x)." But such an ascription seems far too sophisticated to attribute to most creatures with an ANS. Contents mark causal powers. Presumably, this content marks a causal psychological power to think about ordinals, and to determine relations of one-to-one correspondence. Most animals with an ANS lack these capacities.

Ordinals are abstracta beyond their ken, and there is no evidence that aggregates are judged as the same in size via a process of relating their components one-to-one. This example illustrates the general problem: we, as theorists, must specify the indicated property by way of numbers (or other abstracta), but in a way that does not impute to the creature implausible representational relations to the numbers (qua abstracta) themselves.

A similar problem arises when considering the perceptual representation of continuous magnitudes such as length or distance. We, as theorists, may specify a represented property by using a real number, as when we say "the state represents the flagpole as 2.3333 units long." It is implausible that perceptual systems have representational relations to real numbers as such. Peacocke (2015) provides a plausible solution to this apparent problem. When we say that the flagpole is 2.33333 units long, what property do we attribute to it? Philosophers influenced by the logical empiricists have answered: we attribute a triadic relation, holding between the flagpole, the real number 2.333, and the standard unit (in Paris, say). Adapting an argument from Putnam (1969), Peacocke claims that this construal conflicts with the role of magnitudes in causal explanation. It is the magnitude of the flagpole's length itself that causes its shadow to have a certain length; the fact that one can measure the magnitude with a real number via a standard unit is peripheral to the explanation and irrelevant to the magnitude's causal powers. More generally, laws of nature apparently quantify over magnitudes as such. For that reason, we should adopt them into our ontology as a kind of universal.

The magnitudes of a particular type (such as the specific lengths that an object may have) form a class with a particular structure. Peacocke claims that this structure obeys the axioms for extensive magnitudes put forward by Suppes and Zinnes (1963). Such axioms define, for instance, a relation whereby one magnitude is related to two others as their "sum." By obeying such axioms, it can be shown that a set of magnitudes is isomorphic to the real numbers (under a choice of unit object). This isomorphism is what enables us to pick out the magnitudes by means of real numbers. But the magnitudes are distinct from the real numbers. Applied to the case of perception, we can say that the theorist's ascription "the state represents the flagpole as 2.3333 units long" actually ascribes to the the state a representation of the continuous magnitude itself, without representation of any real numbers. And again, the existence of the magnitude itself is established by finding it mentioned in a true statement of natural law. This ensures that the magnitude itself has causal powers, and so is apt, in principle, to cause psychological events in the creature.

We may extend Peacocke's argument to the present case. Recall from chemistry that $PV = nRT$. I will (uncritically) assume that this is a causal law. So understood, it implies that the gas in the box has the pressure and volume it has partly because of its temperature, and partly because of the *number* of molecules that compose it. Hence, the 'n' in $PV = nRT$ denotes a property of the aggregate of gas molecules. We as theorists use numerical expression to pick it out, but, as with continuous magnitudes, the property itself is distinct. We may call the denoted property a *discrete magnitude*, and claim that numerosities are discrete magnitudes.[13] Henceforth, I will use 'N' and Arabic numerals to denote numerosities (understood as discrete magnitudes), and use '$\underline{\mathrm{N}}$' to denote a representational content indicating the numerosity N.

According to numerosity semantics, atomic states of the ANS indicate numerosities. However, there is presumably a largest numerosity N that the ANS can represent. A natural

[13] There is a complication here that I elide, but hope to explore in further work (see footnote 9). Suppose the box actually contains two aggregates: an aggregate of gas molecules, and an aggregate of *pairs* of gas molecules (where pairhood does not require, e.g., spatial proximity). These two aggregates are spatiotemporally coincident, and plausibly have identical causal powers. But they are in one-to-one correspondence with different von Neumann ordinals, and hence have different numerosities. Consequently, it becomes unclear that numerosities are causally efficacious. A natural reply is to say: 'n' in $PV = nRT$ denotes the numerosity of an aggregate of gas molecules, not of *pairs* of gas molecules. In fact, we have no reason to think the aggregate of pairs even exists. T. Burge (1977) follows this tactic, claiming that aggregates can only be composed of entities that count as individuals from the perspective of empirical science; and unrestricted pairings of gas molecules are not individuals from that perspective. But there is a prima facie problem here. On the one hand, there are apparently perceivable aggregates of pairs, where the number of pairs n causes ANS state r with probability $P(r|n)$. See, e.g., Kirjakovski and Matsumoto (2016). On the other hand, the pairings are too arbitrary and "up to the perceiver" to count as individuals from the perspective of empirical science. If we are motivated to say that the ANS can veridically represent numerosities in these cases, we need a more sophisticated account of what aggregates and numerosities are.

thought is to determine this upper bound via discrimination performance. For instance, we might find the least numerosity N such that the creature can discriminate N from $N+1$ only 50% of the time.[14] But suppose those numerosities are 10 and 11. Then the animal cannot represent any numerosity greater than 10. But the animal can easily distinguish 10 from 15, and 20 from 40. On this proposal, the animal would represent all of these numerosities as 10—an implausible consequence for the numerosity semanticist. There is a better proposal. As N increases, the creature will shift from representing an aggregate of N items as an aggregate to representing it as an individual. Consider, for instance, that as the number of dots on a display increases, it begins to look like a densely textured surface, rather than an aggregate of dots. We may then fix the largest perceivable numerosity N_{max} as the average numerosity at which an aggregate of size N ceases to be represented as an aggregate at all.

If we let $\mathbb{N} = \{\underline{1}, \underline{2}, \ldots, \underline{N_{max}}\}$, we can say that $[\![\cdot]\!]$ is a mapping from R to $\mathbb{N}$. We also suppose that there is an ordering relation over $\mathbb{N}$, defined in the obvious way. But what is the exact mapping? One proposal is as follows: for all $r \in R$, $[\![r]\!] = \underline{N}$ iff $r = \mu_N$. Recall that μ_N is the mean of the distribution $P(R|N)$. On this proposal, the overwhelming majority of atomic states have no representational content at all. Consequently, the probability of representing an aggregate as having *any* numerosity is vanishingly small (theoretically zero). This proposal is clearly misguided.

A more natural mapping is to partition the state space of the ANS into ordered subintervals, and assign all the atomic states in an interval to the same numerosity representation. There are various partitions one could adopt. A natural partition to seek is one that maximizes the probability that a state with the content $\underline{N}$ occurs, given that an aggregate with N components has been perceived. Of course, there are tradeoffs. The partition that maximizes the probability of $\underline{1}$ given 1 is the implausible mapping from the last section, which assigns $\underline{1}$ to every atomic state. The tradeoff is that, for all $N \neq 1$, the probability

[14] Note that, in theory, there are no such numerosities. 50% discrimination occurs only if the distributions $P(R|N)$ and $P(R|N+1)$ perfectly overlap. But according to LGM, although these distributions become closer and closer as N increases, they are always separated by some amount.

of N̲ given N is zero. A compromise, then, is to center the interval of atomic states representing N at μ_N, and to compress the intervals logarithmically (to "keep pace" with the distributions). A specific proposal to that end is to place the boundary between successive intervals at the midpoint between the nearest means. Then, for all $r \in R$, the mapping is:[15]

$$[\![r]\!] = \underline{N} \quad \text{iff} \quad \frac{\mu_{N-1} + \mu_N}{2} < r < \frac{\mu_{N+1} + \mu_N}{2}$$

This mapping is depicted graphically in Figure 4.5 below.[16]

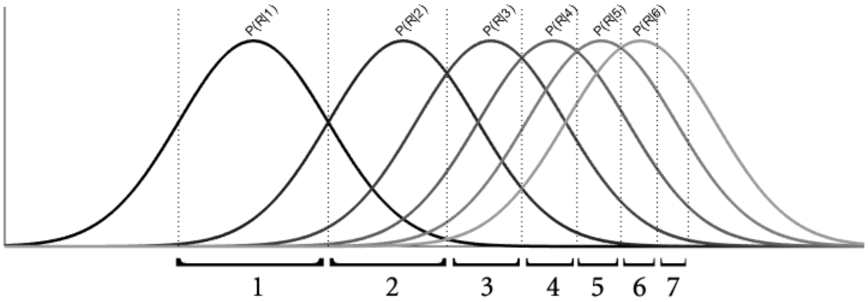


Figure 4.5: The dashed lines mark boundaries between successive intervals. An atomic state falling in the interval marked "1" has the representational content 1̲, etc.

The proposed mapping constitutes a proposal about when numerosity representations occur. Our question now is to see whether there is any explanatory advantage to postulating such representational states. Are there empirical results that motivate postulating causal laws with these representational states as antecedents? I will look at the four core paradigms used in numerosity discrimination tasks. For each task, I will define a *continuous decision rule* that maps individual state-activations to behavioral responses directly. These decision rules are typically ones that have been used in the modeling literature to derive predictions from the LGM, and to confirm the model. I also define a *discrete decision rule* that maps instances of representational contents to behavioral responses. Such a decision rule corresponds to a

[15] Let $\mu_0 = 0$ by stipulation

[16] Note that, although the lower boundary for numerosity 1 is depicted as being halfway between μ_1 and μ_0, in my simulations (see text), I took it to be exactly at $\mu_0 = 0$.

putative causal law. For both rules, I run MATLAB simulations of performance in the discrimination task and compare them qualitatively to empirical results. All MATLAB code is available in the Appendix.

First, consider Experiment 2B from Brannon and Terrace (2000), a *larger-smaller task*. In this task, monkeys were repeatedly shown pairs of numerosities. Each pair contained a numerosity between 1 and 9, and all pairs contained distinct numerosities. Each possible pair from the range 1-9 was presented. The design and order of stimuli controlled for non-numerical properties, as described in §1. In each case, monkeys had to indicate which numerosity was the larger.

On a trial, the topmost aggregate generates an atomic ANS state of type r_T. Similarly, the bottom-most aggregate generates a state of type r_B. To make a decision, the monkey must use some decision rule that maps all possible pairs (r_T, r_B) to one of two decisions: "respond top" or "respond bottom." In this case, the continuous decision rule D_C is

$$\text{Respond "top" if } r_B \leq r_T.$$

$$\text{Respond "bottom" if } r_B > r_T.$$

And the discrete decision rule D_D is

$$\text{If } [\![r_B]\!] = [\![r_T]\!], \text{ defer to } D_C.$$

$$\text{Otherwise: Respond "top" if } [\![r_B]\!] < [\![r_T]\!].$$

$$\text{Respond "bottom" if } [\![r_B]\!] > [\![r_T]\!].$$

Rule D_C is intuitive: if the decision value r_T is greater than r_B, then assume that's because r_T was caused by a larger stimulus. Since r_T was generated by the top stimulus, respond "top." The rule D_D is more complicated. Imagine that, on a trial, r_T and r_B both represent the same numerosity N (i.e. $[\![r_B]\!] = [\![r_T]\!]$). The monkey must decide which aggregate is larger, but two tokens of the same representational type do not provide a basis for such discrimination. In this case, the monkey adverts to the order relations between the different atomic states underlying the identically-typed representations. When r_T and r_B represent different numerosities, D_N says: choose the aggregate bound to the greater representation.

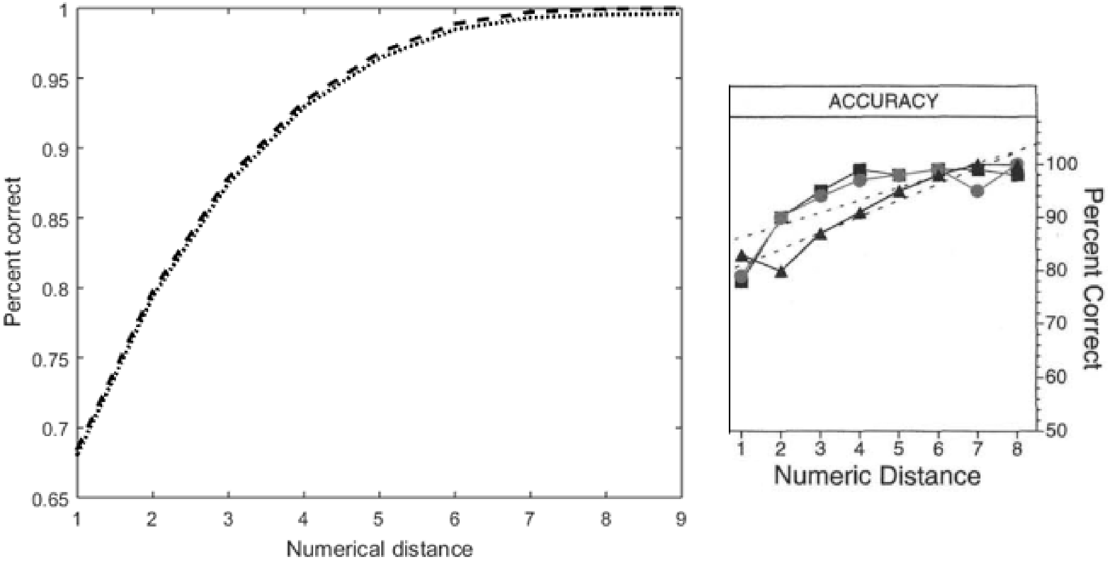


Figure 4.6: Left: Results of a MATLAB simulation comparing the predictions of the two decision rules described in the text. The dark dashed line denotes the predictions of D_C, and the lighter dotted line denotes the predictions of D_D. Right: The empirical results of Brannon and Terrace (2000). This simulation and all others used $\sigma = 0.15$, which is an empirically reasonable value (see Dehaene (2007, p. 529)).

Figure 4.6 depicts the results of a simulation comparing the predictions of the two rules. The x-axis indicates the numerical distance between two stimuli. The y-axis indicates the percent of correct responses made, on average, to stimuli having a particular numerical distance. As the figure illustrates, the two rules make identical predictions. (The dotted line in the Figure has been artificially moved slightly lower, to make the curves visually distinguishable.) It is intuitive why this is so: D_C and D_D are extensionally equivalent rules. Notice that if $[\![r_1]\!] \neq [\![r_2]\!]$, then $[\![r_1]\!] < [\![r_2]\!]$ iff $r_1 < r_2$.

Figure 4.6 also illustrates that the LGM predicts the distance effect. As numerical distance between stimuli increases, discrimination performance improves. The figure also shows the results from Brannon and Terrace (2000). Both decision rules predict results qualitatively similar to those actually observed. Note that Brannon and Terrace themselves conclude that their results confirm the LGM.[17]

The other experiments in Brannon and Terrace (2000) demonstrate, quite interestingly, that monkeys have an ability to *order* stimuli by their numerosities. Monkeys were trained to order aggregates of numerosities 1-4. Subsequently, they were exposed to novel sets of stimuli numbering 5-9. Upon first encounter with such stimuli, monkeys immediately transferred the ordering behavior they had learned for the smaller numerosities. Such task, however, can be viewed as a generalized version of the larger-smaller task. The decision rule D_D in that case would be: if distinct states r and r' both have the same representational content, defer to rule D_C; otherwise, order the stimuli according to the order of the numerosities attributed to them by the ANS. Here too, a continuous rule and a discrete rule make identical predictions.

Consider now a *same-different task*. Recall the Nieder experiment described in §1. Monkeys must release a lever iff the sample and test aggregates match in their numerosities. Let r_s be the atomic state derived from the sample, and let r_t be the atomic state derived from the test. The near-optimal decision rule (see Macmillan and Creelman (2004, pp. 221-223)), and one that has been found to provide excellent fit to the empirical data of the Nieder experiment (see Dehaene (2007, p. 539)), is a continuous decision rule D_C that dictates a "same" response only if r_s and r_t are sufficiently close. More precisely, the system maintains a *decision criterion* δ, and responds "same" only if the difference between atomic states falls below this criterion. So our continuous decision rule D_C is: Respond "same" iff $|r_s - r_t| < \delta$. Our discrete decision rule D_D is the obvious one: respond "same" iff the aggregates are represented as having the same numerosities. That is, respond "same" iff $[\![r_s]\!] = [\![r_t]\!]$.

Figure 4.7 depicts the empirical predictions of each decision rule. The predictions of D_C are depicted in black, and the predictions of D_N are depicted in gray. These rules differ drastically. Moreover, comparison to the actual results of Nieder's experiment (Fig. 1

[17] Literally, they conclude that it confirms the model of Dehaene and Changeux (1993), which, as specified in fn. 6, is a mechanistic implementation of LGM.

above) reveals that D_D is empirically inadequate. D_D predicts a rapid decline in accuracy on trials where the sample and test *do* match, but no such decline is observed.[18] The difference between the rules is due to the fact that, as N increases, the size of the sub-interval $\{r : \llbracket r \rrbracket = \underline{\mathrm{N}}\}$ approaches the size of an individual atomic state. In this limit, D_D becomes the rule: response "same" iff $r_s = r_t$. The probability that $r_s = r_t$ is vanishingly small. Moreover, the poor prediction does not depend on logarithmically compressing the sub-intervals. If the sub-intervals were linearly spaced, the frequency of responding "same" to *non*-matching pairs will drastically increase, beyond what the data show.

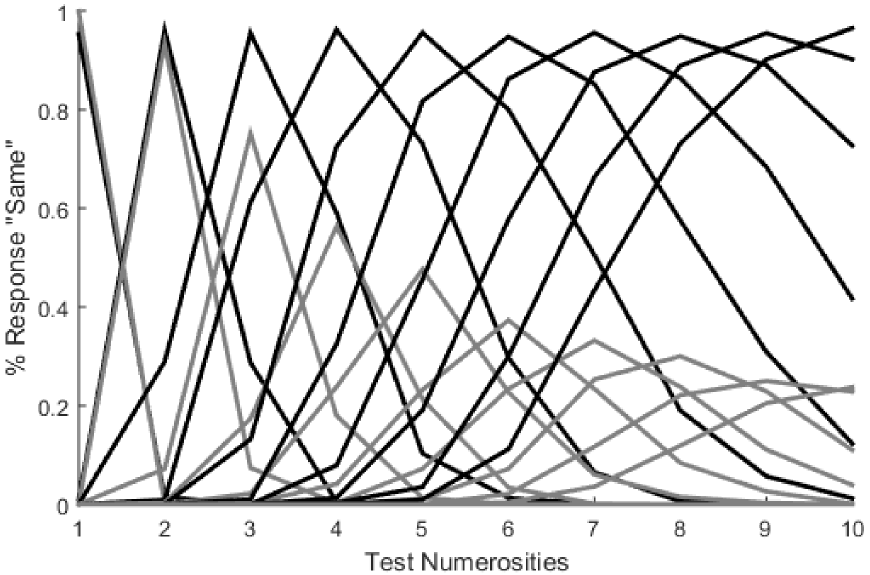


Figure 4.7: Predictions of the continuous rule are in black, while predictions of the discrete rule are in gray. Although not labeled, each curve depicts the behavioral results for a different sample numerosity, as in Fig. 1.

Finally, consider the experiment of Platt and Johnson (1971). Rats were trained to repeatedly press a lever with their paws. When they had pressed the bar some target T

[18] Again, I am only making a qualitative comparison. But if you are skeptical, because you observe in Figure 1 a slight downwards trend in accuracy on matching trials, see the behavioral results in Figure 3 of Nieder and Merten (2007). This study reproduces the experiment of Nieder and Miller (2003), but with numerosities from 1-30. The downward trend in matching accuracies is nowhere near as drastic as D_D predicts.

number of times, a food-dispenser becomes active, and entry into the food-hole leads to reward. Premature entry is penalized by a 10 second time-out, but the response counter is not reset. Platt and Johnson trained rats to perform this task, for different values of T (4, 8, 16 and 24).

Modeling rat behavior in this task is considerably more complicated than the experiments we have so far reviewed. The rat is clearly storing some value in memory. In §3, we discussed several of the choices one must make in modeling the relationship between the ANS and memory. The modeling literature on response-counting tasks does not provide unequivocal guidance on these questions. In any case, Figure 4.8 compares the empirical predictions of a continuous rule (black) and a discrete rule (gray) for this task. It also shows the original data of Platt and Johnson (1971). The simulation assumed that the Nth lever-press draws from $P(R|N)$, that memory is sampled per trial, and that memory is Gaussian but with a slightly higher variance. It does not model any impact of the time-out. Neither set of predictions is particularly good. Both overrate the frequencies of entry to the food-hole given a particular number of executed lever presses. Both apparently underestimate the means of the response distributions. Even so, the discrete rule underestimates the dispersion of the response distribution, and consequently overestimates the frequency with which "correct" responses are made.

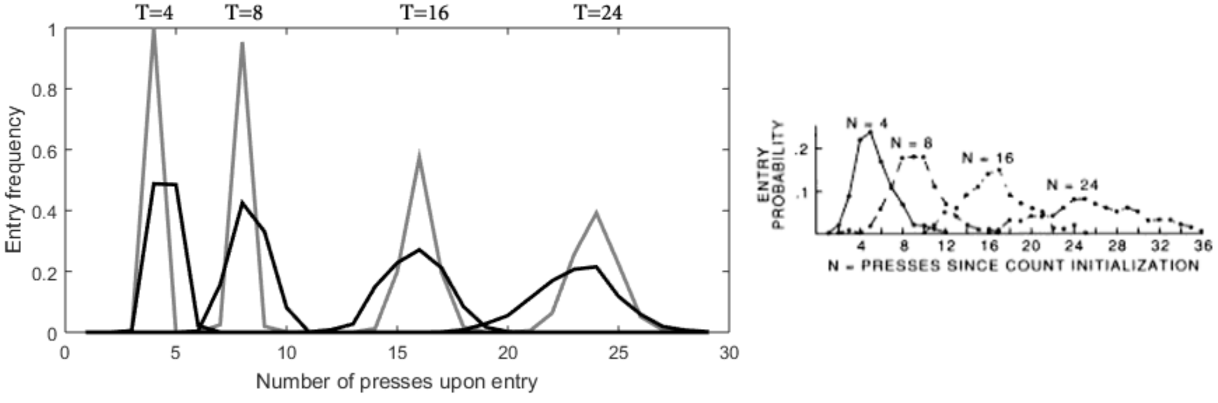


Figure 4.8: Left: Continuous prediction in black, discrete prediction in gray. Right: The empirical results of Platt and Johnson (1971); figure as redrawn in C. R. Gallistel (1990, p. 322).

From these qualitative comparisons between simulations and empirical results, I conclude that there is no empirical reason stemming from the four core paradigms of numerosity discrimination to postulate numerosity representations as described above. In each case, postulation of such states make either identical or empirically worse predictions than postulating only the atomic states of the ANS. There are, of course, other experimental protocols.[19] I hope to discuss these in more depth in future extensions of this work. Preliminary results indicate that numerosity representations are not motivated by these paradigms either.

There are, however, two further experimental findings that might seem to motivate numerosity representations. The first is that monkeys can be trained to associate Arabic numerals with the numerosities of aggregates. Consider the monkey, named Ai, in Tomonaga and Matsuzawa (2002). On a trial, an aggregate of dots was presented (with a numerosity between 1 and 9), and Ai was required to press the Arabic numeral corresponding to the presented numerosity. Results showed the magnitude and distance effects. Prior to this study, Ai had received extensive training in labeling scenes with linguistic expressions (e.g. labeling a shoreline with 'BEACH'), and also in labeling numerosities with Arabic numerals (see Matsuzawa (1985)). Controls indicated that Ai's labeling behavior was based on an estimate of numerosity, not some correlated continuous magnitude. In a further experiment, Ai demonstrated an ability to order the numerals themselves (see Nobuyuki and Tetsuro (2000)).

It seems that, by the end of training, Ai achieved a mapping from atomic ANS states to categorizations with a formal structure akin to our mapping above. In this paper, I will not address the issue of whether these categorizations should be seen as numerosity representations. Instead, I want to emphasize that this ability required extensive training, first to learn to correlate symbols with other stimuli, then more to correlate numerals with numerosities of specific aggregates, and more to finally generalize the ability to aggregates

[19] Two of the most popular that I have not addressed are violation of expectancy and various "addition/substraction" paradigms. See Flombaum et al. (2005) and Xu and Spelke (2000).

of more general constitution. Without such extensive training, animals do not exhibit the classification behavior of Ai. It is possible that Ai's ANS had states representing numerosities prior to training, and the difficulty of training stems from the difficulty of associating these representations with a memorized list of symbols. Alternatively, it's possible that the difficulty of training stems from establishing a learned mapping from Ai's state space to a set of categorizations, where no such mapping existed before.

Second, recall that response times in numerosity discrimination tasks also exhibit the distance and magnitude effects. The most natural way to model response times is to add a temporal dimension to the LGM, by supposing that a perceived aggregate causes *several* ANS states to activate over time. Effectively, the animal samples the conditional distributions. Highly different numerosities cause samples that are immediately recognizably distinct, whereas measuring overlap between the similar samples from similar numerosities requires more samples, and hence more time. The most popular formal extension of a model like the LGM that predicts response times is called the drift-diffusion model.[20] According to such models, the probability $P(R|N)$ is itself represented (or at least encoded) by the ANS. Similarly, on a trial in a discrimination task, experimenters may ask not just for a response, but for the creature's *confidence* in the response they made. Studying confidence is difficult in animals, and few studies have examined it for numerosity discrimination (but see Beran et al. (2006)). The most popular formal extensions of models like the LGM that predict confidence judgements similarly assume that the probability $P(R|N)$ is represented (or encoded) by the animal.[21] On their face, these models seem to assert that, for any ANS state r, the ANS also represents $P(r|1)$, $P(r|2)$,.... Read literally, such a representation represents the probability of a state, given that *numerosity* n has been presented. Thus, it may appear that these models commit us to numerosity representations. I do not believe

[20] See Gold and Shadlen (2007) for a general review of such models. See Dehaene (2007, pp. 548-550) for a sketch of its application to response times in numerosity discrimination.

[21] For a review, see Kepecs and Mainen (2012).

that these elements of the model should be read literally or representationally. I will argue this in other work.

Note that the above argument tells equally against intervallic semantics, which assigns to intervals of atomic states representational contents like between 4 and 6 members(x). Presumably, such a semantics would widen the intervals of the state space that get mapped to contents, and this would only make the failure of empirically adequate predictions worse.

4.5 Against partial numerosity

I take the previous section to show that no atomic semantics is viable that maps the state space into a discrete set of representations. If any atomic semantics is viable, it must map the state space into a continuous set of representations. The control experiments mentioned in §1, however, rule out the most obvious such sets. For instance, we could not map the state space into a continuous set of density representations.

We need a continuous magnitude whose possession by an aggregate is invariant to various transformations of the aggregate. We may take inspiration from the semantics literature on partial counting.[22] Some aggregates can be partially counted. One may say, apparently felicitously and truly, that two and a half bagels are on the table. Moreover, this sentence remains true under various transformations of the aggregate: move its components apart, blow them up in size, change their colors, and still there are two and a half bagels. I will call the property attributed to an aggregate by such a sentence a *partial numerosity*.

The partial numerosity of an aggregate is defined only relative to a property whose "partial satisfaction" by the components is being measured. The partial count of two and a half bagels is 2.5 relative to the property *bagel*, but 3 relative to the property *edible thing*. Consequently, counts can only be partial relative to fractionable properties—that is, properties that admit degrees of satisfaction. The property *bagel* is fractionable, while

[22] For introduction and references, see Liebesman (2016).

the property *spatiotemporally continuous object* is not. Moreover, it's plausible that partial counts are defined only for aggregates with components of the same type. It's unclear what partial count one would give to a heterogeneous collection comprising $2\frac{1}{2}$ cars, $3\frac{1}{6}$ bagels, and $1\frac{1}{6}$ socks. (It is not, for instance, $6\frac{5}{6}$ *things*; there are 9 *things*.) Finally, as N. Salmon (1997) notes, it's plausible that partial counts are only defined when the "sum" of partial objects is less than one. If we see a showroom floor with two whole Toyotas and 16 front-halves of Toyotas, there doesn't seem to be a way to answer "For which X is it true that there are X Toyotas on the showroom floor?"

According to partial numerosity semantics, some state of the ANS has representational content of the form, e.g., $\underline{\text{two and two-thirds components}(x)}$. Having this representational content would require an ability to track the degree to which a component partially satisfies some fractionable property. It is not clear, however, that creatures with an ANS are generally capable of representing any fractionable properties. Animals can perceptually represent bodies and surfaces, but neither of these properties is fractionable—the notion of "half a body" is incoherent. We may assume that animals can perceptually represent things as predators or conspecifics, but these are not plausibly fractionable properties, at least for the animal. From the monkey's perspective, half a lion is either a predator (if represented as living) or not (if represented as non-living).

Let's assume that a monkey can represent something as a banana. The property *banana* is fractionable: being a biological kind, it sets a standard for completeness. Let's also assume, more dubiously, that the monkey can represent something as a half-banana, or a quarter-banana. To conclude that ANS states represent partial numerosities, we need empirical evidence that, e.g., the monkey responds differently to comparisons between two and three bananas, and two and two and a half bananas. If the monkey doesn't respond differently, we should conclude that in both cases, the ANS is being driven by an aggregate representation linked to three component representations (e.g., in both cases, three *edible* *thing* representations). I have found no unequivocal empirical test of this.[23] Given this lack of empirical evidence, and given the empirical implausibility that monkeys represent fractionable properties as such, I conclude that partial numerosity semantics is unmotivated.

4.6 Ratio semantics

Our review of the empirical literature suggests that the ANS functions primarily as a *comparative* capacity. The paradigms we have reviewed assay animals' abilities to compare the numerosities of aggregates. Moreover, the ecological relevance of numerosity for free-ranging animals seems to be in making comparisons: which cache of food offers the greater dividends, which group (mine or theirs) has the greater numbers and is therefore more likely to win in a fight.[24] We saw that monadic semantics has trouble empirically accounting for such comparisons. This is the starting point for a relational semantics. Representational contents mark causal powers, and as we have seen, only *pairs* of ANS states seem to have causal powers, so perhaps only such pairs have representational content.

The first proposal from relational semantics is ratio semantics. Intuitively, this proposal holds that a complex ANS state represents two aggregates as having numerosities standing in a certain ratio. Prima facie, ratio semantics seems well-motivated by the empirical evidence. It's an empirical fact that numerosity discrimination performance is ratio-dependent: one can distinguish numerosity a from b equally reliably as one can distinguish numerosity c from d, if $a/b = c/d$. So if an animal can distinguish 5 from 6 dots 75% of the time, it will also distinguish 10 dots from 12 dots 75% of the time. In both of these cases, a ratio semantics may claim, the animal represents one aggregate as having five-sixths as many components as the other.

[23] Though see Experiment 5 of Flombaum et al. (2005).

[24] For references to the ecological literature on numerosity, see Chapters 1-4 and Chapter 13 in Geary et al. (2015).

As with the other semantic proposals, there are many ways to provide a specific ratio semantics. The first issue is determining the set of ratios that may be represented. The specific proposal I consider is as follows. I assume that the ANS cannot represent the ratio 0:1. I assume that the ANS can represent the ratio 1:1. This is plausibly what is represented in comparing very close numerosities. I also suppose that the largest ratio less than 1:1 that the ANS can represent is the ratio between numerosities at which the animal reaches 75% discrimination accuracy. Let's suppose the animal distinguishes 5 dots from 6 dots 75% of the time. Then the animal can represent aggregates as being in a 5:6 ratio, but cannot represent any greater ratio (such as 6:7). Finally, I will suppose that the set of representable ratios are all integer multiples of the lowest nonzero ratio in the set. A creature whose greatest representable ratio (less than 1:1) is 5:6 would be able to represent the ratios in the set $\mathbb{Q} = \{$1:6, 2:6, 3:6, 4:6, 5:6, 1:1$\}$.

What is the exact mapping from the set of complex states $R \times R$ to $\mathbb{Q}$? Again, there are several options. I will suppose, however, that "same" responses in same-different tasks are driven by the representation $\underline{1{:}1}$. In our running example, the animal reaches 75% discrimination performance on ratios of 5:6. Presented with numerosities in such a ratio, the animal will incorrectly respond "same"—that is, incorrectly represent the ratio as 1:1—25% of the time. So, our mapping from a complex state (r, r') to $\underline{1{:}1}$ must ensure that $\underline{1{:}1}$ occurs erroneously on 25% of trials with a stimulus ratio of 5:6. We can accomplish this by saying that $[\![(r, r')]\!] = \underline{1{:}1}$ iff $|r - r'| < \delta$, and fixing δ so that $P(|r - r'| < \delta|5, 6) = 0.25$. Given that r and r' are drawn from Gaussians, the δ satisfying this condition may be analytically derived.[25]

This leaves our mapping unspecified for complex states when $|r' - r| > \delta$. Intuitively, as r' gets larger than r, the ratio a:b between the aggregate linked to r and the aggregate linked to r' should decrease. Hence, supposing $r < r'$, we want to specify how much larger

[25] r is drawn from $\mathcal{N}(\mu_5, \sigma^2)$ and r' is drawn from $\mathcal{N}(\mu_6, \sigma^2)$. It follows that $r' - r$ is drawn from $\mathcal{N}(\mu_6 - \mu 5, 2\sigma^2)$. Hence, probabilities of the form $a < r' - r < b$ can be calculated from the density function of this distribution.

r' must be if the pair (r, r') represents a particular ratio. Again, there are different ways to specify this. For our running example, I propose a linearly spaced mapping: for integer k, if r' is at least $k\delta$ away from r, then $[\![(r, r')]\!]$ is at least $\underline{\text{(1-}(k/6))}$. Specifically, assuming that $r < r'$, and for $1 \leq i \leq 6$:

$$[\![(r, r')]\!] = \underline{i\text{:}6} \quad \text{iff} \quad (6 - i)\delta < r' - r < (6 - i + 1)\delta$$

This proposal is illustrated graphically below in Figure 4.9. Suppose that aggregate A generates state r on the ANS, as depicted. The dotted boundaries in the figure depict the intervals that r' would have to fall in, relative to r, for (r, r') to have a certain ratio content. If r' falls in the second interval from r, then $[\![(r, r')]\!] = \underline{5\text{:}6}$. Note that, unlike the boundaries set up by numerosity semantics, these boundaries are not fixed to the ANS. They are determined relative to r, and so they "jump around" due to the underlying stochasticity of r. Note that, as drawn, for $r' >> r$, $[\![(r, r')]\!]$ is undefined. This could be remedied by extending the interval for $\underline{1\text{:}6}$ to occupy the remainder of the state space. This extension would have the consequence that, even when an aggregate is much much larger than another, their ratio is nevertheless represented as 1:6.

There is obviously a psychological difference between seeing an aggregate that is twice as numerous as another, and seeing an aggregate that is four times as numerous as another. This psychological difference manifests itself in response time and confidence. I may be quicker to judge an aggregate as larger when it is four times as numerous as a comparison aggregate than when it is only twice as numerous. A chimpanzee may more confidently incite a conflict when its group is four times as numerous as the rival group than when its group is only twice as numerous. Ratio semantics offers a straightforward explanation of these psychological difference: I represent the aggregates respectively as twice and four times as numerous as the other. But there is an alternative explanation: my response times and confidence are functions of the difference between the components of the complex state. As $|r' - r|$ increases, my average response time goes down, and my confidence goes up. These

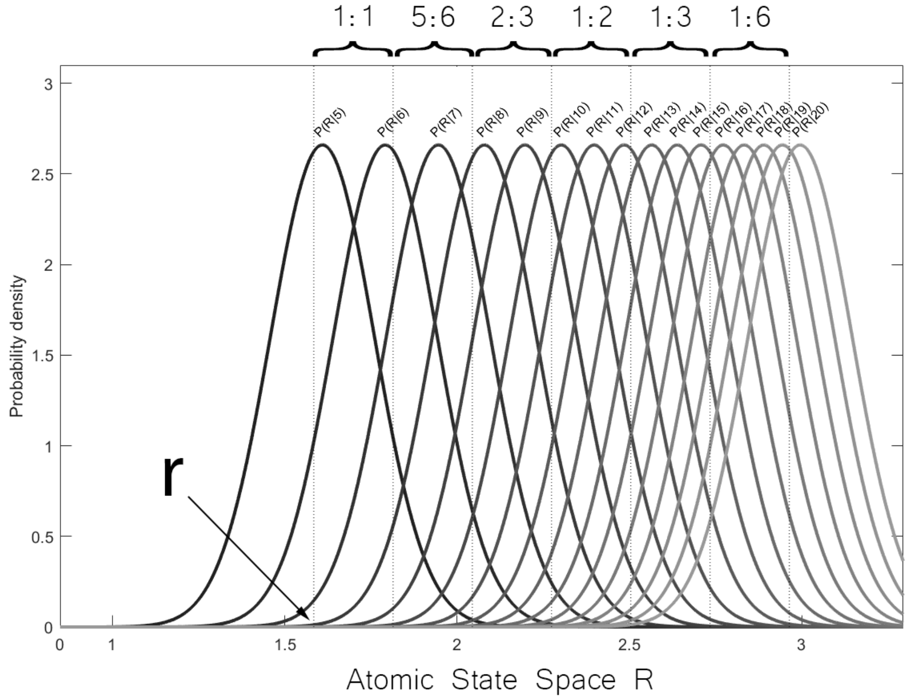


Figure 4.9: Illustration of ratio semantics.

two explanations are likely to make distinct predictions. As r' moves from less than $r + \delta$ to greater than $r + \delta$, ratio semantics predicts a sudden shift in confidence and response times, while the alternative explanation predicts a smooth change in confidence and response times. Empirical curves for response time and confidence tend to be smoothly varying functions, and so ratio semantics is prima facie disconfirmed. However, as I mentioned in §4, modeling response times and confidence with the LGM is complicated, and so a systematic comparison of these two explanations must wait until a future extension of this work.

4.7 Comparative semantics

The final proposal I will discuss is comparative semantics. Comparative semantics is straightforward. It holds that complex states of the ANS can represent one aggregate as being more, less or equally as numerous as another aggregate. Suppose ANS state r_1 is bound to

$\underline{\text{(that } x_1\text{) aggregate(}x_1\text{)}}$, and r_2 is bound to $\underline{\text{(that } x_2\text{) aggregate(}x_2\text{)}}$. And suppose $r1 << r_2$. According to comparative semantics, when the ANS enters into the complex state (r_1, r_2), the creature enters into an overall psychological state with a representational content we might write as

$$\underline{\text{(that } x_1\text{)(that } x_2\text{) aggregate(}x_1\text{) aggregate(}x_2\text{) more numerous(}x_2, x_1\text{)}}.$$

Moreover, the creature represents two aggregates as having the same numerosity when they cause ANS states r_1 and r_2 that are within a certain threshold δ, in accordance with our discussion in §4 and §6. Specifically, let '$x \prec y$' denote the relation between aggregates x and y when y is more numerous, and let '$x \approx y$' denote the relation when x and y are equally numerous. Then we may define our mapping as $[\![(r, r')]\!] = \underline{\approx}$ if $|r' - r| < \delta$, and $\underline{\prec}$ otherwise.

The other semantic proposals discussed in this paper postulated novel states of the ANS. Postulation of novel states requires empirical evidence for those novel states to abductively explain and predict. For each proposal, no such evidence was found. Comparative semantics, by contrast, proposes a representational gloss on states that are already known to be causally efficacious. According to the LGM, a complex state (r, r'), where $|r' - r| < \delta$, causally drives "same" responses. Comparative semantics holds that this complex state represents aggregates as being equally numerous. According to the LGM, a complex state (r, r'), where $r' > r$, causally drives a "larger" response to the aggregate linked to r'. Comparative semantics holds that this state represents that aggregate as larger. The semantics ascribed to the state are appropriate given that state's causal role. More generally, these states support ascription of representational content. The states are differentially sensitive to relations between numerosities as such, abstracting from the other contingent features that aggregates may have. In this respect, the states bear comparison to perceptual constancies. It therefore becomes substantial to describe the animal's performance in terms of veridicality. The animal really was representing one aggregate as being more numerous, not as greater along some other dimension. In one case, it succeeded, and in another, it failed. Moreover, even if ratio semantics is empirically supported, we should still accept comparative semantics. An animal can be correct that one aggregate is more numerous than another, but be wrong about the degree to which it is more numerous.

4.8 Conclusion

As alluded to throughout, the argument of this paper requires further development. To truly rule out numerosity semantics, and to fairly evaluate ratio semantics, a more rigorous examination of models of response time and confidence is required. Puzzles about the metaphysics of aggregates remain. Each semantic proposal should be evaluated in the light of experimental paradigms that are "peripheral" from the core paradigms I discussed here.

Still, I conclude that monadic semantics is empirically ungrounded. I am skeptical about ratio content. That leaves a fairly minimal semantics for the ANS: comparative semantics. Assuming that's right, it is interesting to wonder what becomes of appeals made to the ANS in the philosophical and psychological literature. In explaining the acquisition of numerical concepts in humans, Carey (2009) postulates a central role for the representations of the ANS, and several developmental psychologists follow her. Laurence and Margolis (2005) argues that ANS representations are unfit to play the postulated roles. This debate deserves to be clarified and resolved. Additionally, appeals are made to the ANS in the literature on the epistemology of mathematics. I would like to clarify whether the ANS can in fact support these appeals. I am interested in seeing how a clear and empirically grounded account of numerosity discrimination can shed light on these and other issues.

Bibliography

Aida, S. et al. (2015). Overestimation of the number of elements in a three-dimensional stimulus. *Journal of vision* 15.9, pp. 23–23.

Barthelme, S. and Mamassian, P. (2010). Flexible mechanisms underlie the evaluation of visual confidence. *Proceedings of the National Academy of Sciences* 107.48, pp. 20834–20839.

Barthelmé, S. and Mamassian, P. (2009). Evaluation of Objective Uncertainty in the Visual System. *PLoS Computational Biology* 5.9.

Bayne, T. (2010). *The Unity of Consciousness*. Oxford University Press.

Beck, J. et al. (2007). Probabilistic population codes and the exponentialfamily of distributions. *Progress in Brain Research* 165, pp. 509–519.

Beck, J. (2019). On perceptual confidence and 'completely trusting your experience'. *Analytic Philosophy* 61.2, pp. 174–188.

Beck, J. M. et al. (2008). Probabilistic population codes for Bayesian decision making. *Neuron* 60.6, pp. 1142–1152.

Beran, M. J. et al. (2006). Rhesus macaques (Macaca mulatta) monitor uncertainty during numerosity judgments. *Journal of Experimental Psychology: Animal Behavior Processes* 32.2, p. 111.

Block, N. (2018). If perception is probabilistic, why does it not seem probabilistic? *Philosophical Transactions of the Royal Society B* 373.1755.

Brannon, E. M. and Terrace, H. S. (2000). Representation of the numerosities 1–9 by rhesus macaques (Macaca mulatta). *Journal of Experimental Psychology: Animal Behavior Processes* 26.1, p. 31.

Brannon, E. M., Wusthoff, C. J., et al. (2001). Numerical subtraction in the pigeon: Evidence for a linear subjective number scale. *Psychological Science* 12.3, pp. 238–243.

Burge, J. and Geisler, W. S. (2015). Optimal speed estimation in natural image movies predicts human performance. *Nature Communications* 6, pp. 1–11.

Burge, T. (2003). Perceptual entitlement. *Philosophy and Phenomenological Research* 67.3, pp. 503–548.

Burge, T. (1977). A theory of aggregates. *Nous*, pp. 97–117.

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

Burge, T. (2010b). Origins of Perception. *Disputatio* IV.29, pp. 1–38.

Burge, T. (2020). Epistemic Entitlement. In: ed. by P. J. Graham and N. J. L. L. Petersen. Oxford University Press. Chap. Entitlement: The basis for empirical epistemic warrant, pp. XX–XX.

Camp, E. (2009). A Language of Baboon Thought? In: R. Lurz, ed. *Philosophy of Animal Minds.* Cambridge University Press, pp. 108–127.

Carey, S. (2009). *The origin of concepts.* Oxford University Press.

Cody, M. (1966). A general theory of clutch size. *Evolution* 20.2, pp. 174–184.

Colombo, M., Elkin, L., and Hartmann, S. (2016). Bayesian Cognitive Science, Monopoly, and Neglected Frameworks. *Psychology.*

Colombo, M. and Hartmann, S. (2017). Bayesian Cognitive Science, Unification, and Explanation. *The British Journal for the Philosophy of Science* 68.2, pp. 451–484.

Colombo, M. and Series, P. (2012). Bayes in the Brain: On Bayesian Modelling in Neuroscience. *British Journal for the Philosophy of Science* 63.3, pp. 697–723.

Dakin, S. C. et al. (2011). A common visual metric for approximate number and density. *Proceedings of the National Academy of Sciences* 108.49, pp. 19552–19557.

Dehaene, S. (2001). Subtracting pigeons: logarithmic or linear? *Psychological science* 12.3, pp. 244–246.

Dehaene, S. (2007). Symbols and quantities in parietal cortex: Elements of a mathematical theory of number representation and manipulation. *Attention & performance XXII. Sensori-motor foundations of higher cognition, ed. P. Haggard & Y. Rossetti*, pp. 527–74.

Dehaene, S. (2011). *The number sense: How the mind creates mathematics*. OUP USA.

Dehaene, S. and Changeux, J.-P. (1993). Development of elementary numerical abilities: A neuronal model. *Journal of cognitive neuroscience* 5.4, pp. 390–407.

Diaconis, P., Holmes, S., and Montgomery, R. (2007). Dynamical Bias in the Coin Toss. *SIAM Review* 49.2, pp. 211–235.

Ernst, M. O. and Banks, M. S. (2002). Humans integrate visual and haptic information in a statistically optimal fashion. *Nature* 415, pp. 429–433.

Flombaum, J. I., Junge, J. A., and Hauser, M. D. (2005). Rhesus monkeys (Macaca mulatta) spontaneously compute addition operations over large numbers. *Cognition* 97.3, pp. 315–325.

Frankish, K. (2009). Partial Belief and Flat-Out Belief. In: F. Huber and C. Schmidt-Petri, eds. *Degrees of Belief*. Springer, pp. 341–354.

Gallistel, C. R. (1990). *The organization of learning*. The MIT Press.

Gallistel, C. et al. (2014). The perception of probability. *Psychological Review* 121.1, pp. 96–123.

Geary, D. C., Berch, D. B., and Koepke, K. M., eds. (2015). *Evolutionary Origins and Early Development of Number Processing.* Vol. 1. Mathematical Cognition and Learning. Elsevier.

Gelman, A. et al. (2013). *Bayesian Data Analysis.* 3rd ed. Chapman and Hall.

Gillies, D. (2000). Varieties of Propensity. *The British Journal for the Philosophy of Science* 51.4, pp. 807–835.

Gold, J. and Shadlen, M. (2007). The Neural Basis of Decision Making. *Annual Review of Neuroscience* 30, pp. 535–574.

Grimson, W. (1981). *From Images to Surfaces: A Computational Study of the Human Early Visual System.* MIT Press.

Gross, S. (2020). Probabilistic representations in perception: Are there any, and what would they be? *Mind and Language* 35.3, pp. 377–389.

Hacking, I. (1965). *Logic of Statistical Inference.* Cambridge University Press.

Hajek, A. (2007). The reference class problem is your problem too. *Synthese* 156, pp. 563–585.

Halpern, J. (2017). *Reasoning about Uncertainty.* MIT Press.

Hawthorne, J. (2005). Degree-of-Belief and Degree-of-Support: Why Bayesians Need Both Notions. *Mind* 114.454, pp. 277–320.

Henik, A., ed. (2016). *Continuous Issues in Numerical Cognition: How Many or How Much.* Academic Press.

Hillis, J. H. et al. (2002). Combining sensory information: Mandatory fusion within, but not between, senses. *Science* 298.5598, pp. 1627–1630.

Icard, T. (2016). Subjective Probability as Sampling Propensity. *Review of Philosophy and Psychology* 7, pp. 863–903.

Joyce, J. M. (1999). *The Foundations of Causal Decision Theory.* Cambridge University Press.

Joyce, J. (2009). Accuracy and Coherence: Prospectsfor an Alethic Epistemology of Partial Belief. In: F. H. C. Schmidt-Petri, ed. *Degrees of Belief.* Synthese Library.

Kahneman, D. and Tversky, A. (1979). Prospect Theory: An Analysis of Decision under Risk. *Econometrica* 47.2, pp. 263–292.

Kepecs, A. and Mainen, Z. F. (2012). A computational framework for the study of confidence in humans and animals. *Philosophical Transactions of the Royal Society of London B: Biological Sciences* 367.1594, pp. 1322–1337.

Kirjakovski, A. and Matsumoto, E. (2016). Numerosity underestimation in sets with illusory contours. *Vision research* 122, pp. 34–42.

Knill, D. C. (1998). Surface orientation from texture: Ideal observers, generic observers and the information content of texture cues. *Vision Research* 38.11, pp. 1655–1682.

Knill, D. C. (2003). Mixture models and the probabilistic structure of depth cues. *Vision Research* 43.7, pp. 831–854.

Knill, D. C. and Richards, W., eds. (1996). *Perception as Bayesian Inference.* Cambridge.

Knill, D. C. and Saunders, J. A. (2003). Do humans optimally integrate stereo and texture information for judgments of surface slant? *Vision Research* 43.24, pp. 2539–2558.

Körding, K. P., Beierholm, U., et al. (2007). Causal Inference in Multisensory Perception. *PLoS ONE* 2.9, e943. DOI: `10.1371/journal.pone.0000943`.

Körding, K. P. and Wolpert, D. M. (2006). Bayesian decision theory in sensorimotor control. *Trends in Cognitive Sciences* 10.7, pp. 319–326.

Kyburg, H. (1974). *The Logical Foundations of Statistical Inference.* Springer.

Landhuis, E. (2004). Magician-turned-mathematician uncovers bias in coin flipping. *Stanford Report* 7-4-2004.

Larry Wasserman (2000). *All of Statistcs.* Springer.

Laurence, S. and Margolis, E. (2005). Number and Natural Language. In: S. S. Carruthers Peter Laurence Stephen, ed. *The Innate Mind: Structure and Contents.* Oxford University Press. Chap. 13, pp. 216–238.

Levi, I. (1977). Direct inference. *The Journal of Philosophy* 74.1, pp. 5–29.

Lewis, D. (1980). A Subjectivist's Guide to Objective Chance. In: R. Jeffrey, ed. *Studies in Inductive Logic and Probability.* Vol. 2. Berkeley: University of California Press, pp. 83–132.

Liebesman, D. (2016). Counting as a Type of Measuring. *Philosopher's Imprint* 16 (12), pp. 1–25.

Ma, W. J. et al. (2006). Bayesian inference with probabilistic population codes. *Nature Neuroscience* 9.11, pp. 1432–1438.

Ma, W. (2019). Bayesian Decision Models: A Primer. *Neuron* 104.1, pp. 164–175.

Ma, W., Beck, J., and Pouget, A. (2011). Sensory Cue Integration. In: ed. by M. L. J Trommershauser K Kording. Oxford University Press. Chap. A Neural Implementation of Optimal CueIntegration, Chapter 21.

Macarthur, R. and Levins, R. (1967). The limiting similarity, convergence, and divergence of coexisting species. *The American Naturalist* 101.921, pp. 377–385.

Macmillan, N. A. and Creelman, C. D. (2004). *Detection theory: A user's guide.* Psychology press.

Makinson, D. (2009). Levels of Belief in Nonmonotonic Reasoning. In: F. Huber and C. Schmidt-Petri, eds. *Degrees of Belief*. Springer, pp. 341–354.

Maloney, L. and Mamassian, P. (2009). Bayesian decision theory as a model of human visual perception: Testing Bayesian transfer. *Visual Neuroscience* 26.1, pp. 147–155.

Maniscalco, B. and Lau, H. (2012). A signal detection theoretic approach for estimating metacognitive sensitivity from confidence ratings. *Consciousness and Cognition* 21, pp. 422–430.

Matsuzawa, T. (1985). Use of numbers by a chimpanzee. *Nature* 315.6014, pp. 57–59.

Mayo, D. (1996). *Error and the Growth of Experimental Knowledge.* University of Chicago Press.

Meck, W. H. and Church, R. M. (July 1983). A mode control model of counting and timing processes. *Journal of Experimental Psychology: Animal Behavior Processes* 9.3, pp. 320–334.

Mills, S. and Beatty, J. (1979). The propensity interpretation of fitness. *Philosophy of Science* 46.2, pp. 263–286.

Miyazaki, M., Nozaki, D., and Nakajima, Y. (2005). Testing Bayesian Models of Human Coincidence Timing. *Journal of Neurophysiology* 94.1, pp. 395–399.

Morrison, J. (2016). Perceptual Confidence. *Analytic Philosophy* 57.1, pp. 15–48.

Munton, J. (2016). Visual Confidences and Direct Perceptual Justification. *Philosophical Topics* 44.2, pp. 301–326.

Munton, J. (2018). The Eye's Mind: Perceptual Process and Epistemic Norms. *Philosophical Perspectives* 31.1, pp. 317–347.

Munton, J. (Forthcoming). Visual Indeterminacy and the Puzzle of the Speckled Hen. *Mind and Language.*

Najemnik, J. and Geisler, W. S. (2005). Optimal eye movement strategies in visual search. *Nature* 19.2004, pp. 1–12.

Nieder, A. (2013). Coding of abstract quantity by ‘number neurons’ of the primate brain. *Journal of Comparative Physiology A* 199.1, pp. 1–16.

Nieder, A., Freedman, D. J., and Miller, E. K. (2002). Representation of the quantity of visual items in the primate prefrontal cortex. *Science* 297.5587, pp. 1708–1711.

Nieder, A. and Merten, K. (2007). A labeled-line code for small and large numerosities in the monkey prefrontal cortex. *The Journal of neuroscience* 27.22, pp. 5986–5993.

Nieder, A. and Miller, E. K. (2003). Coding of cognitive magnitude: Compressed scaling of numerical information in the primate prefrontal cortex. *Neuron* 37.1, pp. 149–157.

Nobuyuki, K. and Tetsuro, M. (2000). Numerical Memory Span in a Chimpanze. *Nature* 403, p. 6765.

Peacocke, C. (2015). Magnitudes: Metaphysics, Explanation, and Perception. In: A. C. Danièle Moyal-Sharrock Volker Munz, ed. *Mind, Language and Action: Proceedings of the 36th Annual Wittgenstein Symposium.* de Gruyter, pp. 357–390.

Pearl, J. (2000). *Causality: Models, Reasoning, and Inference.* Cambridge University Press.

Platt, J. R. and Johnson, D. M. (1971). Localization of position within a homogeneous behavior chain: Effects of error contingencies. *Learning and Motivation* 2.4, pp. 386–414.

Putnam, H. (1969). On properties. In: *Essays in Honor of Carl G. Hempel.* Springer, pp. 235–254.

Ramsey, F. P. (1926). The Foundations of Mathematics and other Logical Essays, in: ed. by R. B. Braithwaite. London Kegan, Paul, Trench, Trubner New York Harcourt, Brace and Company. Chap. Truth and Probability, Ch. VII, 156–198.

Reichenbach, H. (1949). *The Theory of Probability. An Inquiry into the Logical and Mathematical Foundations of the Calculus of Probability.* University of California Press.

Rescorla, M. (2019). A realist perspective on Bayesian cognitive science. In: A. Nes and T. Chan, eds. *Inference and Consciousness.* Taylor and Francis.

Roitman, J. D. and Brannon, E. M. (2003). Nonverbal Representations of Time and Number in Animals and Human Infants. In: *Functional and Neural Mechanisms of Interval Timing.* CRC Press.

Salmon, N. (1997). Wholes, parts, and numbers. *Nous* 31, pp. 1–15.

Salmon, W. (1971). *Statistical Explanation and Statistical Relevance.* University of Pittsburgh Press.

Sato, Y. and Körding, K. P. (2014). How much to trust the senses: learning likelihoods. *Journal of Vision* 14.2014, pp. 1–14.

Saunders, J. A. and Knill, D. C. (2001). Perception of 3D surface orientation from skew symmetry. *Vision Research* 41, pp. 1–21.

Savage, L. (1954). *The Foundations of Statistics.* New York: Wiley.

Serwe, S., Kording, K., and Trommershauser, J. (2011). Visual-haptic cue integration with spatial and temporal disparity during pointing movements. *Experimental Brain Research* 210.1, pp. 67–80.

Seydell, A., Knill, D. C., and Trommershauser, J. (2010). Adapting internal statistical models for interpreting visual cues to depth. *Journal of Vision* 10.4, pp. 1–27.

Shafir, E. and Tversky, A. (1995). Decision making. In: E. Smith and D. Osherson, eds. *An invitation to cognitive science (2nd ed.). Thinking: An invitation to cognitive science.* MIT Press, pp. 77–100.

Shea, N. (2018). *Representation in Cognitive Science.* Oxford University Press.

Shea, N. (2020). Representation in Cognitive Science: Replies. *Mind and Language* 35.3, pp. 402–412.

Siegel, S. (Forthcoming). How can experiences explain uncertainty? *Mind and Language.*

Sober, E. (1984). *The Nature of Selection.* University of Chicago Press.

Spirtes, P., Glymour, C., and Scheines, R. (2001). *Causation, Prediction, and Search.* 2nd ed. Cambridge, MA: MIT Press.

Strevens, M. (2006). *Bigger than Chaos: Understanding Complexity through Probability.* Harvard University Press.

Strevens, M. (2011). Probability Out Of Determinism. In: C. Beisbart and S. Hartmann, eds. *Probabilities in Physics.* Oxford University Press.

Suppes, P. and Zinnes, J. (1963). Basic Measurement Theory. In: R. Luce, R. Bush, and E. Galanter, eds. *Handbook of Mathematical Psychology, Vol. 1.* John Wiley and Sons, NY.

Tassinari, H., Hudson, T., and Landy, M. (2006). Combining Priors and Noisy Visual Cues in a Rapid Pointing Task. *The Journal of Neuroscience* 26.40, pp. 10154–10163.

Tomonaga, M. and Matsuzawa, T. (2002). Enumeration of briefly presented items by the chimpanzee (Pan troglodytes) and humans (Homo sapiens). *Animal Learning & Behavior* 30.2, pp. 143–157.

Vance, J. (2020). Precision and Perceptual Clarity. *Australasian Journal of Philosophy.*

Xu, F. and Spelke, E. S. (2000). Large number discrimination in 6-month-old infants. *Cognition* 74.1, B1–B11.

Zorzi, M. and Butterworth, B. (1999). A Computational Model of Number Comparison. In: *Proceedings of the Twenty First Annual Conference of the Cognitive Science Society.* Lawrence Erlbaum Associates.

Zylberberg, A., Barttfeld, P., and Sigman, M. (2012). The construction of confidence in a perceptual decision. *Frontiers in Integrative Neuroscience* 6, p. 79.