· The Kid, the Clerk, and the Gambler ·

Critical Studies in Statistics and Cognitive Science

Mathias Winther Madsen

· The Kid, the Clerk, and the Gambler ·

Critical Studies in Statistics and Cognitive Science

ILLC Dissertation Series DS-2015-02

INSTITUTE FOR LOGIC, LANGUAGE AND COMPUTATION

For further information about ILLC-publications, please contact

Institute for Logic, Language and Computation
Universiteit van Amsterdam
Science Park 107
1098 XG Amsterdam
phone: +31-20-525 6051
e-mail: illc@uva.nl
homepage: http://www.illc.uva.nl/

Cover design by Julie Cetti and Marie-Anne Soyez
after a watercolor by Mathias Winther Madsen.

Printed and bound by Ipskamp Drukkers, Enschede.

ISBN: 978-94-6259-947-5

· The Kid, the Clerk, and the Gambler ·

Critical Studies in Statistics and Cognitive Science

Academisch Proefschrift

ter verkrijging van de graad van doctor aan de
Universiteit van Amsterdam
op gezag van de Rector Magnificus
prof.dr. D.C. van den Boom
ten overstaan van een door het college voor
promoties ingestelde commissie, in het openbaar
te verdedigen in de Agnietenkapel
op diensdag 1 december 2015, te 10:00 uur

door

Mathias Winther Madsen

geboren te Kopenhagen, Denemarken.

Promotores: Prof. dr. M.J.B. Stokhof
Prof. dr. M. van Lambalgen

Overige leden: Prof. dr. A. Betti, Unversiteit van Amsterdam
Prof. dr. G.J. Steen, Unversiteit van Amsterdam
Dr. W.H. Zuidema, Unversiteit van Amsterdam
Prof. dr. J.W. Romeijn, Rijksunversiteit Groningen
Prof. dr. R.A. Zwaan, Erasmus Unversiteit Rotterdam
Dr. D.H. Lassiter, Stanford University

Faculteit der Geesteswetenschappen

Contents

Acknowledgments and Sources

This dissertation is the product of four years of research and teaching. During this time, I have had help and support from a number of people.

I thank my supervisors, Martin Stokhof and Michiel van Lambalgen, for their encouragement and solidarity in bleak times. I would also like to thank the students who took my two project courses at the ILLC, the 2014 project course in information theory[1] and the 2015 course in statistical learning theory[2]. Peter Grünwald, Ronald Meester, and Nina Gierasimczuk also gave interesting guest lectures in the course on statistical learning theory, and I am very grateful for that.

I spent March 2015 at Stanford, guest lecturing on information theory in Dan Lassiter's course on probability theory for linguists[3]. I owe a very special thanks to Dan for inviting me to CSLI, and to the director of the institute, Chris Potts, for agreeing to fund this visit. I had a great time in California.

During my time at Stanford, I met with Thomas Icard and Noah Goodman, both of whom gave interesting comments and perspectives on what would later become Chapter 7 of this dissertation. I also had very enjoyable conversations with Brad Efron and Peter Norvig about Fisher's early papers, clean Python code, Chomsky's philosophical project, the job market in statistics, and many other things.

A number of people also diserve thanks for helping me during the final phases of the writing process. Simon Daniel Eiriksson, Johannes Emerich, and Jelle Bruineberg read versions of Chapters 6 and 9 and pointed out problems and errors. Jelle Bruineberg, Kelly Mostert, and Iris van de Pol gave comments on the Dutch summary of the dissertation. Marie-Anne Soyez and Julie Cetti kindly agreed to design the cover under an unreasonable time pressure. Linnea Uddenmyr made the sun shine and talked me through Chapter 9.5.

[1] `http://informationtheory.weebly.com/`

[2] `http://statisticallearningtheory.weebly.com/`

[3] `http://web.stanford.edu/~danlass/courses/prob-and-stats-winter15/`

Much of this dissertation consists of reworked versions of materials that have been published elsewhere. The sources are as follows:

- Chapter 2 will be published shortly in *Cognitive Science.* I am grateful to my reviewer Cristina Cacciari for pushing me to dig deeper into the interesting historical background of the topic, here recounted in Chapter 2.1.2.
- Chapter 3 is previously unpublished. It is based on my presentation at the inaugural conference of IACS in Lund, 2014.
- Chapter 4 is an extended version of a paper from the recently published collection *Bayesian Natural Language Semantics and Pragmatics* (Madsen, 2015a).
- Chapter 5 was published in 2014 as part of the collection *Pristine Perspectives on Logic, Language, and Computation* (Madsen, 2014b).
- Chapter 6 is a massively revised version of a paper presented at NASSLLI 2014 in Maryland. I also presented related material at the RAOM workshop in Groningen (Madsen, 2014a).
- Chapter 7 was recently published in *Künstliche Intelligenz* (Madsen, 2015b).
- Chapter 8 is a slightly modified version of a paper that I delivered at the 2014 Trends in Logic conference in Gent. Katie Steele and Richard Pettigrew gave thoughtful and detailed commentaries on the paper at that conference.
- Chapter 9 is previously unpublished. It is loosely based on material I presented on three different occasions in 2014: At the LoLa day at the ILLC, at the History of the Humanities conference in Rome, and at the workshop on the Logic and Philosophy of Mathematical Practices in Brussels.

I want lastly to thank the reviewers from which I have received comments and critiques over the last four years. Although some of my experiences with the peer review system have been truly nightmarish, I have also received useful, constructive, precise, scholarly criticism which changed my research and my writing for the better. I am grateful to the anonymous reviewers who provided this feedback.

Amsterdam
November, 2015.

Mathias Winther Madsen

Chapter 1

Introduction

This dissertation is about the concept of thought. It contains series of case studies that critically examine how this concept is put to work in human sciences such as linguistics, psychology, logic, statistics, and artificial intelligence.

These case studies span across a variety of topics, but they share a consistent philosophy and methodology. Throughout the dissertation, I see each of these sciences as an expression of a specific conception of what a human being is. Attending to the simplifying assumptions that go into creating this idealized cartoon image can often help us crack open the hard question of where a particular theory goes wrong and what kind of observations it precludes itself from making.

Consider for instance the following little story:

> The link between neurons and physical behavior is easy to see in a well-understood case such as the knee-jerk reflex—the lifting of your leg when your doctor taps below your kneecap. Neural connections run from the sensing neurons in the knee, through one link in the spinal cord, back to motor neurons that drive the leg muscles. . . .
>
> Now imagine that your doctor, instead of tapping your knee, asks you to lift your leg. The link between an input signal (the doctor's words) and the output (lifting) now does involve the brain and is much more complex, but the neural information processing perspective is still the key to understanding the behavior. (Feldman, 2006, p. 4)

According to this story, an analogy with the knee-jerk reflex is not only useful for understanding complex social behavior; it is "the key."

My point is not that this take on the issue is right or wrong, but rather that it represents a highly specific perspective on the phenomenon. Recognizing the selective blindness that comes with such a perspective can help us better notice the ways in which it fails to explain important aspects of reality.

Although all human sciences work with some idealized notion of a human being, the distinguishing features they attribute to this cardboard figure often differ substantially between theories. Some authors emphasize intuitive and imagistic thought; others rule-governed rationality; and others again competent handling of uncertainty.

Each of these emphases bring to mind a different kind of prototypical behavior, and indeed a different kind of prototypical person. Such stereotypes illuminate some things and obscure others, and they suggest different sorts of evidence. In each of the case studies in this dissertation, I take up a particular theory, examine the evidence supporting it, and ask how our conclusions might change if we change the anthropological assumptions behind it.

The first case in point is cognitive metaphor theory. This is the influential theory proposed by Lakoff and Johnson (1980a) which claims that metaphors like *attacking an argument* are expressions of deeply rooted cognitive habits:

> It is important to see that we don't just *talk* about argument in terms of war. We can actually win or lose arguments. We attack his position and we defend our own. We gain and lose ground. We plan and use strategies. (Lakoff and Johnson, 1980a, p. 4)

This argument leans heavily on a certain story about what "we" do. It draws much of its strength from a certain stereotypical image of how normal people normally behave, and this image in turn generates certain psychological hypotheses and expectations. This makes it a prime candidate for the kind of critical examination I have in mind.

In the first two chapters of this dissertation, I thus survey the vast literature on cognitive metaphor theory, trying to assess the quality of the "mountains of evidence" which allegedly "fills the pages of our discipline to overflowing" (Johnson and Lakoff, 2002, pp. 251 and 261). As it turns out, the situation is a bit more complicated than that, and I discuss a number of problematic issues related to individual differences, conceptual ambiguities, and cognitive flexibility. Much of this counter-evidence cannot be explained without taking the normalizing force of language seriously, and this raises important issues about where the "cognitive" in "cognitive metaphor theory" comes from.

Chapter 3 also examines a set of time-honored concepts from linguistics, although from a completely different corner of the discipline. Here, the topic is the distinction between syntax and semantics, and a specific strand of evidence from the brain sciences that allegedly shows that this distinction is based on neurological fact.

To a large extent, this finding relies on a certain philosophy of language much more than it supports one. As we open the black box and start looking at the details of the experimental record, we find that the facts only suggests a distinction between syntax and semantics if we decide to force the data into that conceptual scheme in the first place. Once we swap that vocabulary for another one, the same data may suggest a very different picture, centered around concepts of sense-making, expectation, and action.

The next case study also takes up a concept that has played an operative role in linguistic theory but often remained unanalyzed. In their book *Relevance*, Sperber and Wilson (1986) claimed that a whole range of pragmatic phenomena in natural language could be explained in terms of a the concept of relevance, but their definition of the concept was rather vague, and they admitted that the theory

> ... is very speculative and, as it stands, too general to determine directly either specific experimental tests or computer simulations. (Sperber and Wilson, 1987, pp. 709–710).

In a response to this observation, I use Chapter 5 to take a closer look at the concept of relevance. I suggest one possible way of spelling out its meaning in more precise terms, drawing on ideas from the golden age of mathematical statistics in the postwar period.

Continuing on the topic of probability, Chapters 6 and 7 raise the question of how to design a formal calculus for reasoning about other people's reasoning. Like many problems in epistemology and artificial intelligence, this question has been approached from two different angles, probability theory and logic.

These two traditions have often been seen as competing and contradictory, but I argue that their insights can be integrated quite seamlessly if we are sufficiently pragmatic in our attitude towards the formal modeling of reasoning. By separating concerns about realism from concerns about design, we can built the benefits of both of those traditions into a single system.

The last two chapters deal more exclusively with statistics, and specifically with its relation to the concept of rational thought.

In Chapter 8, I review a well-known decomposition of statistical estimation error into two components, one due to an unwillingness to learn from data, and one due to an excessive sensitivity to the data. The observation that some measure of prior opinion can be beneficial seems to provide an argument in defense of Bayesian statistics, which is often derided for its use of unjustified prior assumptions. As I explain, however, this argument draws on notions that are antithetic to fundamental principles of Bayesian statistics, and the issue can therefore not be settled this way.

In the last case study, I dive deeper into these issues, taking a historical perspective of the controversies surrounding statistical inference. The emotions run surprisingly high in this debate, and there seems to be a disconnect between its content and the polemic tone in which it is conducted. I explain this discrepancy by looking at statistics in the wider context of the history of rationality, and I exhibit both internal and external evidence indicating that the debate about statistics reflects the mood swings of the recent history of rationality.

As this overview shows, this dissertation spans a number of diverse topics in linguistics, psychology, and statistics. It can, however, be read as a coherent attempt to interrogate the concept of thought: In all of the disciplines that I cover, this concept provides a focal point which tends to draw out the most fundamental assumptions about what human beings are. By pulling this conceptual thread, we can unravel the whole theoretical fabric.

The purpose of that exercise is not to argue that we should cleanse the human sciences of all philosophical baggage. Their anthropological assumptions are not trespassers that happened to slip into the otherwise rigorous language of science; on the contrary, disciplines like linguistics, psychology, and statistics are by their very nature formal expressions of a philosophical anthropology. Once we recognize this connection between the roots and the leaves, we may come to better understand the limitations of those sciences and, in some cases, to overcome them.

Chapter 2

Cognitive Metaphor Theory and the Metaphysics of Immediacy

One of the core tenets of cognitive metaphor theory is the claim that metaphors ground abstract knowledge in concrete, first-hand experience. In this chapter, I argue that this grounding hypothesis contains some problematic conceptual ambiguities, and under many reasonable interpretations, empirical difficulties. I present evidence that there are foundational obstacles to defining a coherent and cognitively valid concept of "metaphor" and "concrete meaning," and some general problems with singling out certain domains of experience as more immediate than others. I conclude from these considerations that whatever the facts are about the comprehension of individual metaphors, the available evidence is incompatible with the notion of an underlying conceptual structure organized according to the immediacy of experience.

2.1 Introduction

In many of its contemporary versions, one of the central postulates of cognitive metaphor theory is that all abstract concepts ultimately derive from our immediate experience of physical action and perception: "Meaning is grounded in our bodily experience" (Johnson, 2007, p. 12; Kövecses, 2005, p. 18).

The thinking which underlies this claim is that many abstract concepts — time, love, causality, and the like — are understood in terms of analogies with more direct, physical experiences. For instance, you might understand the abstract notion of *swallowing a compromise* because you project intuitions about the actual, bodily act of swallowing onto the abstract domain of negotiation.

Other aspects of knowledge and understanding are, according to most versions of the theory, grounded indirectly through a chain of metaphors that connect very

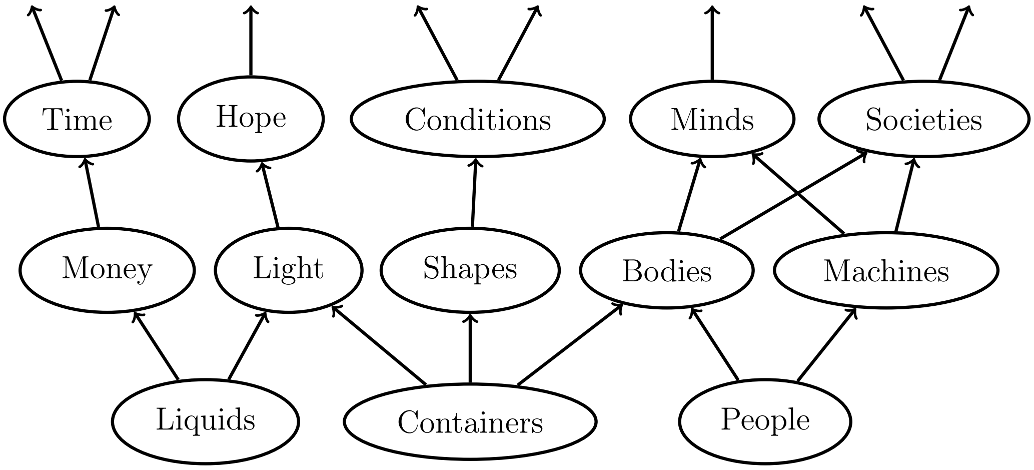


Figure 2.1: A sketch of a hypothetical "conceptual system," with bodily experience at the bottom, and more abstract domains higher up.

abstract concepts with increasingly concrete ones:

> Once a domain of knowledge becomes well known, it can itself serve as a source domain (basis) for understanding more novel concepts (Feldman, 2006, p. 209).

After you have learned to think of light as a kind of water (that *flows*, *pours in*, *fills the room*, etc.), you can then learn to think about hope or reason as a kind of light. Ultimately, your knowledge of any abstract domain is thus grounded in your direct, first-hand, physical experience.

If this grounding hypothesis is correct, it suggests that the "conceptual system" postulated by the theory could be seen as a large, ordered structure (Lakoff and Johnson, 1980b). Its roots would be immediate, sensory experiences, and its leaves would be abstract concepts. Figure 2.1 shows one way of sketching such a conceptual system, with nodes representing concepts, and the connections representing cognitive analogies. (The links shown in the picture are based on Lakoff et al. (1991), but the specifics are not particularly important here.) This image, in turn, suggests that careful analysis might trace back any particular piece of abstract knowledge — about, say, mathematics, politics, or poetry — to its roots in bodily experience.

2.1.1 Beyond the Metaphysics of Immediacy

The purpose of this chapter is to point out some conceptual and empirical problems with this picture of a "tree of knowledge," and with the notion that careful analysis of language can reveal how the individual member of a speech community experiences the world.

I will argue that there are strong reasons to doubt that everybody has the same sense of immediate experience, and that it simply is not true — even on the most charitable reading — that "people have the same bodies and basically the same relevant environments, and so will have very much the same experiences" (Lakoff, 2008, p. 26). It is thus nonsensical to draw conclusions about how "we" think on the basis of linguistic observations, or to expect a public language to reflect a private experience.

To make this point, I will first discuss some problems with the project of constructing a coherent cognitive model out of a collection of linguistic observations. I then present some more direct evidence that experience is more loosely related to language than cognitive metaphor theorists generally acknowledge. Based on these discussions, I conclude that there are a number of questionable leaps of faith involved in drawing conclusions about individual cognition based on the behavior of the language community as a whole.

I should emphasize, however, that this does not entail that analogical thought is irrelevant to linguistic comprehension. It is possible, for instance, that some people understand phrases like *inflation rate* by, say, forming a mental image of a balloon. There are ways that such a claim could be spelled out as an explicit processing theory, including some that may have some empirical merit. What I want to challenge is the notion that linguistic metaphors imply analogical reasoning like smoke implies fire, and that we need a certain kind of first-hand experience to understand a metaphor. Comprehension does not require grounding, and in many cases, people do just fine without it.

This leaves partly open the more difficult question of how people understand abstract concepts like love, argument, causality, and so on. I will not attempt to answer this questiom here, only note that it is probably quite safe to assume that they use a number of different strategies, that these strategies differ from person to person, and that they change with age, circumstance, and experience. The point of this chapter is to argue that this multitude of strategies cannot be unified into a single "conceptual system," much less one that stands in a one-to-one relationship with language.

2.1.2 A Brief History of Cognitive Metaphor Theory

The abundance of metaphors in everyday language was widely recognized among linguists in the first half of the 20th century. Bloomfield (1933), for instance, noted that phrases like *a head of cabbage* were strictly speaking metaphorical and presented a large collection of such cases, commenting that one can "add examples practically without limit" (p. 149).

However, with the increased interest in cognitive theories of psychology in 1940s and 1950s, the interpretation of this observation started to change. A number of scholars started specualating that thought itself might be "radically metaphoric" (Richards, 1938, p. 48). Most prominently, of course, Whorf (1956) tirelessly pointed out the metaphysical assumptions built into the English metaphors of time, thought, and movement, and he repeatedly suggested that these linguistic differences might reflect or even cause a completely different perspective on the world.

In psychology too, other authors made similar observations. For instance, Asch (1955) noted that English-speakers often describe people in terms of sensory concepts like *warm*, *bright*, and *bitter*, and he collected linguistic evidence that seven other (unrelated) languages showed similar tendencies. Asch noted that these patterns could be explained either in terms of "resemblances we experience between particular physical and psychological data" or in terms of conventional associations, but he warned against jumping to conclusions from linguistic evidence alone (p. 30–31).

Werner (1954) also found experimental evidence that the generalization errors children exhibited in the lab had numerous similarities with the principles of semantic growth described by Bréal (1900) and Bloomfield (1933), giving further weight to the claim that the etymological principles of metaphorical meaning extension had a cognitive or perceptual basis (cf. Werner, 1919; Werner and Kaplan, 1963).

The relationship between synesthesia and poetic language was also an early object of fascination among psychologists. Downey (1912) thus attempted to reconstruct the perceptual skills necessary to unpack the synesthetic metaphors in English poetry, and Ullmann (1951) compiled a corpus of just over 2000 synesthetic metaphors from poetic text in order to investigate the topic more rigorously (pp. 277–284). Ullman also found that some sensory domains were more likely to be source domains for the metaphorical description of others, suggesting a linear ordering of sensory domains according to how "differentiated" there were:

(2.1) Touch $<$ Heat $<$ Taste $<$ Scent $<$ Sound $<$ Sight

This idea resurfaced in the work of Williams (1976), who suggested, based on an inspection of English and Japanese dictionary entries, that the sensory domains could be placed in a partial order slightly more complicated that the one proposed by Ullman (cf. Fig. 2.2). Marks (1978) further speculated that the psychological basis of this process was a tendency for perception to share substrates and routines across the senses.

The comparison between synesthesitic perception and ordinary metaphor was also the starting point for the impressive book by Osgood et al. (1957), which presented statistical evidence that sensory oppositions tended to exhibit very strong cross-modal correlations. For instance:

> A happy man is said to feel "high," a sad man "low"; the pianist travels "up" and "down" the scale from treble to bass; souls travel "up" to the

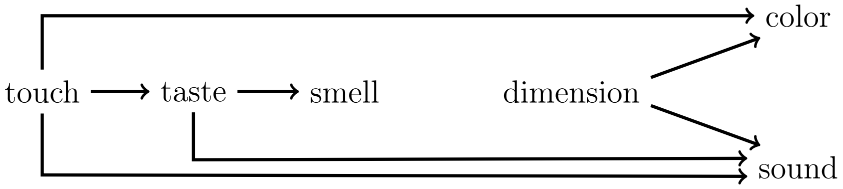


Figure 2.2: The partial ordering of the senses proposed by Williams (1976, p. 463). His concept of "dimension" refers to features like thickness, hollowness, position, size, and the like.

> good place and "down" to the bad place; hope is "white" and despair is "black." (Osgood et al., 1957, p. 21)

These impressionistic observations were tested in several different experiments, and the authors consistently found that the actual dimensionality of semantic space was much smaller than its nominal dimensionality, due to correlations between different facets of experience.

In parallel to these and other developments in psychology (cf. Billow, 1977; Brown, 1958), the interest in metaphor and analogical reasoning also grew considerably in philosophy (Cassirer, 1946; Langer, 1948; Leatherdale, 1974), linguistics (Stern, 1931; Thomas, 1969; Levin, 1977), and other disciplines. Just to mention one particularly interesting example, Schon (1963), an industrial consultant, argued in great detail that the "displacement of concepts" was an important force in creative thought and illustrated his point by detailed analyses of the visual metaphors of comprehension (pp. 170–176).

In the light of these developments, it would be hard to agree with Lakoff and Johnson that they had overthrown "central assumptions" of "the Western tradition since the Greeks" after spending a week compiling everyday metaphors (Lakoff and Johnson, 1980a, pp. ix–x). In fact, a wide variety of researchers in a wide variety of disciplines had already presented the same data and and proposed very similar cognitive explanations for a least a quarter century. Lakoff and Johnson did, however, go much further than their predecessors in terms of how strong conclusions they drew from their linguistic evidence, thus sparking another round of discussion oddly reminiscent of the heated debates that followed Whorf's provocative statements in the 1950s.

2.1.3 Critical Literature on Cognitive Metaphor Theory

Perhaps not surprisingly, Lakoff and Johnson's work attracted criticism from the very outset. A number of reviewers of the book noted immediately that it rested on the same circular reasoning for which Whorf was originally criticized (Black, 1981; Smith, 1982; Strang, 1982). This criticism has since been reiterated many times, underscoring the need for psychological evidence in support of psychological claims (Murphy, 1996; McGlone, 2001; Haser, 2005).

This call did not go unheeded, however. By now, a number of psychological studies have compiled various sorts of evidence related to cognitive metaphor theory, with mixed conclusions. A few representative samples from this literature are worth discussing.

In a very classical paradigm, cued recall studies have consistently found that retrieval cues based on intended figurative meaning work better than cues based on literal word meanings or other superficial characteristics (Verbrugge and McCarrell, 1977; McGlone, 2001). This could possibly reflect the fact that the comprehension of a metaphor like *food for thought* involves little or no analogical reasoning about actual food. A series of studies have also investigated whether such conventional metaphors prime people for the corresponding literal meanings, e.g., whether exposure to the metaphor *a prolific researcher* helps you read the word *fertile* faster. The results have been inconsistent, and the findings seem to depend very heavily on the design of the materials (Keysar et al., 2000; Thibodeau and Durgin, 2008).

In the reverse direction, Boroditsky and Ramscar (2002) has found that priming people with literal movement tends to bias their reading of movement metaphors, with the prime accounting for about an eighth of the total variance. Wilson and Gibbs (2007) found similar effects after teaching people to perform bodily movement cues. Their results are problematic, however, due to the unconventional nature of the metaphors they used (e.g., *stretch for understanding*) as well as the fact that almost half of their subjects tacitly named the movements they were asked to perform, raising some doubts about the non-verbal nature of the task (Wilson and Gibbs, 2007, p. 727).

In another non-verbal priming study, Meier et al. (2004) found that the valence of words like *ugly* or *angel* is assessed about 3% slower on average when the words are written in a color incongruent with its meaning, i.e., white for *ugly* and black for *angel* (however, see also Brandt et al., 2014). The opposite did not hold, that is, valence did not prime color. This seems to support the notion that color is more primordial than valence, and the latter is understood in terms of the former.

However, the same researchers also found in another study that word valence primes vertical position (priming you with the word *wise* makes you look up), while the opposite is not true (Meier and Robinson, 2004). This made the same connection, but in the opposite direction: Now the abstract target domain unilaterally primed the source domain, while the other study had found a source domain to unilaterally prime a target domain. No explanation has yet been given for this discrepancy.

Lastly, in a striking confirmation of cognitive metaphor theory, Núñez and Sweetser (2006) found that the remarkable Aymara metaphors for time, in which the future is “behind” and the past is “in front” of the speaker, cohered with the gestures the Aymara-speakers used during interviews (held mostly in Spanish). The weight of this evidence is reduced somewhat, however, by the fact that English-speakers systematically gesture left-right rather than front-back when they talk about time (Casasanto and Jasmin, 2012). In light of these two conflicting observations, it is natural to ask whether we should see the case of Aymara as confirming cognitive metaphor theory and the case of English as refuting it, or reject the relevance of both observations.

These are only a few examples from a much larger literature. More studies are reviewed by Gibbs (2011), who tends to read the evidence as supporting cognitive metaphor theory, and McGlone (2011), who tends to read the evidence as refuting it.

It should be noted, however, that most of these studies deal with what we might call the validity question of cognitive metaphor theory: Are there or are there not cognitive analogies behind our linguistic metaphors? That is, does language provide a realistic picture of cognition? In this chapter, I take a somewhat different approach and instead question the notion that the omnibus concepts of “language” and “cognition” have a single coherent referent at all. If this is not the case, then the question of whether one resembles the other will, of course, be hard to get off the ground.

2.2 Reliability Issues in Metaphor Identification

In this section, I will discuss some issues related to the definition of a metaphor. I will argue that a naive approach to the question is infeasible, but also that some of the recent attempts to design more rigorous protocols have problems of their own.

2.2.1 The Limits of Intuition

In *Metaphors We Live By*, Lakoff and Johnson relied on introspective evidence to discover a hidden metaphorical structure behind phrases like *unemployment went up*. In examples like these, the sentence often contains a word which has an obviously physical meaning and an obviously non-physical one. The word *up*, for instance, has one rather concrete, physical meaning, and one abstract. Most people thus accept without question that phrases like *the balloon goes up* should be given cognitive, historical, and analytical priority over phrases like *unemployment went up*.

Not all cases are this uncontroversial, though. Lakoff and Johnson themselves had to take some rather obscure rhetorical detours in order to explain how the conceptual mapping ARGUMENT IS WAR could be grounded in first-hand experience for people who had never fought in wars (Lakoff and Johnson, 1980a, ch. 13).

Similar problems could have been pointed out for the many English metaphors deriving their meaning from sailing, gardening, farming, or other activities that are no longer as widely practiced as they used to be.

For some years, this tension posed a real threat to the coherence of the theory. The issue seemed to be explained away, however, when Grady (1997) introduced the notion of "primary metaphors." His idea was to restrict the scope of cognitive metaphor theory so that its claims about direct grounding in first-hand experience would only apply, in terms of the picture in Figure 2.1, to the bottom layer of the conceptual system.

Even within this weaker theory, however, a number of polysemies still cause a bit of trouble. Consider for example the following English words:

bulb electrical lamp; seed of an onion

cap lid on a bottle; soft hat

knot fastening on a string; tumor; protuberance on a branch or root.

cup drinking bowl (noun); to round one's hands (verb).

Being motivated in similarity, these polysemies are almost paradigmatic examples of metaphors. Yet it seems quite unnatural to claim, for instance, that one of the two meanings of *cap* is more concrete, familiar, or intuitive than the other. And more problematically, the phrase *cup one's hands* is derived historically from the noun *cup* even though cognitive metaphor theory generally takes hand movements to be one of the most fundamental, familiar, and immediate experiences we have (e.g., Rohrer, 2001). The arrow thus seems to be pointing in the wrong direction in this case.

These and other issues related to the introspective methodology of cognitive metaphor theory have caused concern among a number of researchers within recent decades. Attempts have consequently been made to come up with more precise definitions of what exactly it means for one word sense to be more "concrete" than another, and thus what a metaphor is.

2.2.2 The Pragglejaz Definition

The most systematic attempt at pinning down the concept of metaphor is probably the one by the Pragglejaz Group (2007), an international research group consisting of ten prominent cognitive metaphor theorists.[1] According to the definition developed by this group, a word is used metaphorically when it is not used in its most "basic contemporary meaning." The crucial notion of "basic" meaning is then defined by the following four features (Pragglejaz Group, 2007, p. 3; brackets in original):

[1] Peter Crisp, Raymond Gibbs, Alan Cienki, Graham Low, Gerard Steen, Lynne Cameron, Elena Semino, Joseph Grady, Alice Deignan, and Zoltán Kövecses.

—More concrete [what they evoke is easier to imagine, see, hear, feel, smell, and taste];

—Related to bodily action;

—More precise (as opposed to vague);

—Historically older;

Such a pluralistic definition raises some questions: How should we handle disagreement between these criteria? Are they equally important? If not, how should we prioritize them?

In order to answer these questions, I have devised a new method for obtaining qualitative evaluations of word meanings and collected some data pertaining to these four dimensions of "basicness." This method, which will be described in more detail in the following subsection, allowed me to assess whether these four dimensions yield consistent answers that can be used to define a precise notion of metaphor.

The purpose of this investigation was to assess the reliability of the Pragglejaz scheme, not its validity. That is, I wanted to investigate whether the Pragglejaz method can be construed in such a way that, given two word senses, it can consistently pick out one of them as being more "basic." This is distinct from asking what "basic" means and whether the method measures what it purports (Franzen, 2000). It should be noted, though, that the validity question is moot for a very unreliable decision procedure: If you cannot trust your measurements, it does not matter whether you are measuring the right thing.

Since my purpose was to probe the possibility of defining a precise scientific notion of metaphor, I was not interested in correlating the four Pragglejaz criteria with any intuitive notion of "metaphoricity." The idea was not to quantify the folk notion of a metaphor, but to assess whether the four Pragglejaz criteria could be used to define useful and transparent concept of "basicness."

I should also emphasize that the results reported below do not question the potential cognitive reality of processes like mental simulation, mental images, or analogical reasoning. Such processes may sometimes play a part in sentence comprehension. What I am critical of is the notion that the dictionary senses of a word are "radial" projections from a core meaning, and that this tree structure should be a realistic picture of online comprehension (Lakoff, 1987, Ch. 6). As mentioned above, I doubt that the thought processes of most people follow such a pattern, and I suspect that any attempt to shoehorn a list of dictionary senses into such an order will meet strong resistance from the data.

2.2.3 A Reliability Assessment Method

In order to assess the amount of agreement or disagreement between the four Pragglejaz criteria, I needed a way of quantifying pairwise comparisons by each of the four criteria. For some of these criteria, human judgments seemed inevitable — but they also posed a danger, since most people have at least some rudimentary level of theoretical knowledge of their own language, and this might bias their responses (cf. Keysar and Bly, 1995).

For this reason, I chose to collect my human judgments based on "masked" lexical ambiguities, by which I mean pairs of target words that are translations of a single source word in some underlying language. For instance, the Danish verb *synke* means both "to swallow" and "to sink," but no Dutch or English word covers both of these senses. The ambiguity of *synke* in Danish is thus "masked" in a translation into Dutch or English. If I ask a Dutch speaker to compare some features of the Dutch words for *swallow* and *sink*, I should thus expect to get a response not biased by prior linguistic knowledge.

Based on this idea, I designed a questionnaire with 22 masked ambiguities, and I recruited 34 Dutch liberal arts students, none of whom spoke Danish (mean age 18.91 years, standard deviation 1.38; 14 male and 20 female). The questionnaire presented pairs of word senses in disambiguating contexts like the following (here translated from the Dutch):

(2.2) (a) *Workout aims at making the muscles stronger.*

(b) *Add chili powder in order to make the dish hotter.*

More examples can be found in Appendix A. Both the order of the pairs and the internal order of the two options were randomized between questionnaires. The sentences were adapted from real examples from various Dutch websites, and care was taken to make them as similar as possible in terms of length, readability, and structure.

Each student was given one of two written instructions:

- One group was instructed to tick off the sentence with the "more concrete" underlined word, where "concrete" explained as "easier to imagine, see, hear, feel, smell, and taste."
- The other group was told to tick off the sentence with the word which was "more related to bodily action," where "related to bodily action" was exemplified with "walk, bite, hold, jump, and kick."

In both cases, I also allowed subjects to refrain from answering if they found it "completely impossible to make a meaningful decision." This options was used occasionally, so the number of responses were not the same for each pair of words.

The responses thus provided a "concrete" and a "bodily" pair of votes for each pair of words (cf. Table 2.1). In addition to this data, I looked up the historical age of each word sense in two comprehensive standard dictionaries of Danish (Hjorth et al., 2003–2005; Dahlerup et al., 1918–1956). The latter of these can roughly be compared in style and scope to the OED. In all but one case, one of

Word	Sense 1	Sense 2	old		bodily		concrete		precise	
balde	bale	buttock	0	1	6	11	3	11	9	5
blære	bladder	blister	0	1	8	10	3	10	8	9
kegle	idiot	cone	0	1	12	4	8	6	8	5
klappe	slam shut	clap	0	1	4	14	3	9	1	4
knop	pimple	bud	0	1	4	11	15	0	12	8
krøllet	wrinkled	curly	0	1	9	9	4	9	0	0
pande	forehead	pan	0	1	7	8	3	9	6	10
prop	cork	clot	1	0	4	13	9	4	9	6
rage	grope, grab	shave	0	1	17	1	5	9	3	11
rive	tear, rip	rake	1	0	5	10	3	8	2	1
rør	pipe	receiver	1	0	5	8	3	8	9	11
rykke	move, draw	yank, jerk	0	1	4	14	3	11	0	2
skam	pity	shame	0	0	2	15	2	12	7	5
skør	crazy	brittle	0	1	10	6	6	9	0	0
skraber	razor	nap	1	0	1	16	10	4	11	5
skrald	bang, clap	garbage	1	0	18	1	7	8	7	7
snurre	spin	tingle	1	0	12	7	1	14	3	3
stærk	hot, spicy	strong	0	1	1	17	11	3	0	0
synke	sink	swallow	1	0	4	15	4	10	2	1
tabe	lose	drop	0	1	3	14	4	10	0	1
tak	jag, barb	notch	1	0	17	2	6	8	7	4
vask	wash	sink	1	0	9	5	4	9	9	7

Table 2.1: Empirical assessments of the four Pragglejaz criteria for the two meanings of 22 Danish words.

the two word senses had a clear historical priority. I thus had a quantification of the three first Pragglejaz criteria.

The last criterion, "precise as opposed to vague," is more elusive. "Fire truck" is obviously more precise than its superconcept "car," but it is unclear whether, say, "fabric" is more precise than "dust." Some amount of arbitrary choice seems unavoidable in these cases.

In order to minimize my own part in this arbitrary choice, I decided to quantify the notion of precision by using the concept hierarchy in WordNet 3.0 (Fellbaum, 1998), identifying a concept as precise if it was separated from the root of the concept tree by a long string of increasingly narrow superconcepts. For instance, the concept "shame" was given a precision score of 5, since the path connecting the corresponding WordNet concept to the root concept "entity" consists of 5 steps:

(2.3) shame > feeling > state > attribute > abstract entity > entity

For about 82% of the concepts recorded in WordNet, there is only one such path; for the remaining 18% of the cases, some selection or aggregation procedure must be applied in order to convert these depths into a single number. I chose to use the length of the longest path (rather than, say, the average length), since I suspected this might be least sensitive to artifacts of the encoding of the database. In general, however, precision scores are not likely to change very much with changes in the aggregation function, since the variation in path lengths is moderate. Even within the 18% of the word meanings that have multiple paths, the standard deviation of the path lengths is only about 1.46 on average. This compares favorably to average difference between the path lengths in my data set, which was 2.22, with a standard deviation of 0.27.

I acknowledge that this precision measure still is rather arbitrary, and that I could equally well have used any number of others. But since the Pragglejaz authors did not detail how they intended to operationalize the concept of precision, I felt it was fair to pick some reasonable interpretation as long as it did not bias results in the direction of unduly large disagreements.

2.2.4 The Reliability of the Pragglejaz Criteria

The methodologies described above resulted in the data set shown in Table 2.1. This data set does not immediately give rise to a measure of correlation or agreement between the four Pragglejaz criteria, since it comes in the slightly unusual format of pairs of integers. This is not the format assumed by most agreement statistics: For instance, Cohen's Kappa and Fleiss' Kappa require the judgments to be categorical labels, and Cronbach's alpha and Cronbach's standardaized alpha require them to be real numbers (Gwet, 2012; Kline, 2013).

Replacing the pairs of votes by a label indicating the "winner" of each vote seems unnecessarily reductive, and it would potentially inflate the agreement

	bodily	concrete	precise
old	0.1043	−0.0413	0.4965
bodily		−0.1475	0.0528
concrete			0.3156

Table 2.2: Correlation coefficients between the four Pragglejaz criteria, estimated from the data in Table 2.1.

statistics, by erasing the quantitative differences between, for instance, $(16, 17)$ and $(0, 34)$. It thus seems more appropriate to map these pairs (a, b) into some real number $t(a, b)$ in order to conserve at least some quantitative information.

A reasonable candidate for the transformation t is the fraction of the votes favoring one particular choice,

$$t_0(a, b) \;=\; \frac{a}{a+b}, \qquad \text{with } t_0(0,0) \;=\; \frac{1}{2},$$

or a smoothed version of the ratio, such as

$$t_\gamma(a, b) \;=\; \frac{a+\gamma}{a+b+2\gamma}.$$

Using the unsmoothed ratio yields the correlation coefficients shown in Table 2.2. For moderate values of γ, these numbers differ only very slightly, and only in the direction of weaker correlations.

As the table shows, the correlations are not overwhelming. The largest one is the correlation between historical age and precision — but even that one only achieves a correlation coefficient of 0.5, indicating that knowing the value of one of the two variables only decreases the variance of the other by about $0.5^2 = 25\%$ (Blackwell, 1969, Ch. 7).

Correlation data such as these can be aggregated into an agreement score by means of a number of different statistics. One option is Cronbach's alpha, which compares the variance within the individual groups to the variance we get get by pooling all of the data into a single super-group (Gwet, 2012, p. 243). Another possibility is Cronbach's standardized alpha, which is the average of all the correlations with a few minor modifications (Gwet, 2012, p. 245).

In the present case, these two statistics evaluate to 0.3742 and 0.3555, respectively. These scores decrease further if the smoothing parameter γ is increased. Since alpha scores below 0.7 are generally considered indicators of poor agreement (Kline, 2013, p. 15), these scores are certainly in the poor end of the scale, by a

	bodily	concrete	precise
old	*synke, skraber, rage*	*snurrende, knop*	*stærk, rør, balde, knop*
bodily		*stærk, knop, skraber, rage*	*stærk, synke, rage, skam*
concrete			*skam*

Table 2.3: Some words whose meanings are particularly difficult to sort due to conflicts between one or more of the Pragglejaz dimensions.

quite large margin. We can thus be quite confident that the Pragglejaz criteria will frequently disagree with each other when used on words similar to those used in this experiment.

A closer inspection of the data set also allows us to pinpoint words that cause particular trouble. For instance, the respondents tended to think that the concept of "grope or fondle" was more strongly related to bodily action than the concept of "shaving," but they made the opposite judgment with respect to concreteness. Hence, the Danish word *rage*, the underlying source word, represents a conflict between the bodily dimension and the concrete. By collecting such conflicting examples, we can construct a table of dissociations, as shown in Table 2.3.

These results show that the Pragglejaz criteria cannot be combined to form a coherent measure of "basicness" without strong additional assumptions. Too often, they disagree too much to suggest a clear answer. Since the criteria are virtually completely uncorrelated, a weighted average of them will also not fare much better: It will either add up to random noise (if the weights are roughly uniform) or effectively allow a single criterion to dominate the decision (if the weights are highly skewed). Our only choices are thus a method of very poor reliability or a simplistic one-dimensional measure, which in the absence of strong empirical evidence seems unlikely to have any cognitive validity.

This does not necessarily mean that there are no objective answers to the question of whether certain people comprehend certain concepts by means of some kind of analogical inference. However, it does indicate that the lexicon of a language like Danish does not suggest a single, coherent picture of semantic structure which can be used as a hypothesis about cognition. It might thus be the case that cognitive metaphor theory has given us a fairly realistic image of the comprehension of words like *grow*, *grasp*, *sting*, and *step*, but it does not follow that this can be extended to *act*, *bulb*, *cap*, *curb*, *cycle*, *drop*, *flap*, *knot*, *plane*, *right*, *stool*, *tap*, or the rest of the lexicon of any real language.

2.3 Further Problems for Word Sense Ordering

In the previous section, I tried to indicate that there are a number of unresolved issues built into the Pragglejaz method for identifying a metaphor, even though their method is one of the most explicit ones in the literature. This does not necessarily mean that it could not succeed if more details were provided. Perhaps some more explicit, more fleshed-out method could actually tell us reliably when a word is comprehended by means of a cognitive analogy. Perhaps a different set of independent variables could do the job.

In this section, I want to argue that there are some fundamental obstacles to constructing such a diagnostic, even in principle. I want to argue that whatever such a method tells us, it either has to give us contradictory results or radically shrink the set of examples that can be considered metaphorical.

2.3.1 Opposing Metaphors

The problems with devising a notion of "basic" meaning which makes stable judgments can be seen from a phenomenon that has sometimes showed up in the margins of metaphor theory under the name of "bidirectional metaphors" (Ortony, 1979; Carroll, 1994; Morgan, 2008; Forceville, 2008). This phenomenon concerns metaphors like the following:

(2.4) BIOLOGY IS POLITICS

(a) *The animal kingdom*

(b) *Dinosaurs ruled the earth*

(c) *The lion is the king of the beasts*

(d) *The law of the jungle*

(2.5) POLITICS IS BIOLOGY

(a) *Ben-Gurion turned from a hawk into a dove*

(b) *Perry is leading the pack in new poll*

(c) *It's is a sheep's vote*

(d) *The GOP is playing alpha male*

By any usual standard of cognitive metaphor theory, both of these metaphors are instances of non-verbal conceptual mappings: They are productive, they cohere with behavior, they appear in pictures as well as language, etc.[2] There is a large number of examples of such metaphor pairs, as the following small sample shows:

(2.6) (a) PEOPLE ARE ANIMALS (*a wild teenager*)
(b) ANIMALS ARE PEOPLE (*a social insect*)

(2.7) (a) ORGANISMS ARE MACHINES (*hearts pump*)
(b) MACHINES ARE ORGANISMS (*computers think*)

(2.8) (a) COUNTRIES ARE FAMILIES (*founding fathers*)
(b) FAMILIES ARE COUNTRIES (*my mom's a dictator*)

(2.9) (a) COUNTRIES ARE BODIES (*head of state*)
(b) BODIES ARE COUNTRIES (*immune defense*)

(2.10) (a) LANGUAGE IS MUSIC (*learn German by ear*)
(b) MUSIC IS LANGUAGE (*I wrote that song*)

(2.11) (a) FAT IS RICH (*rich cake*)
(b) RICH IS FAT (*fat profit*)

These, too, are multimodal and structurally "deep." For instance, the typical cartoon robot will not only look like a human being, but also work, play, and fall in love like one. Reversely, any textbook in physiology will depict the human body as a mechanical device. Scores of pictorial material exemplify both analogies.

Hence, a metaphor between two domains cannot in general be taken as evidence of an ordering of those domains. If that were the case, then biology would both be "more basic" and "less basic" than politics. The domains of people, machines, plants, animals, bodies, and so on would then collapse into a single all-encompassing and thus useless equivalence class.

If we take the methodology of cognitive metaphor theory seriously, then such opposing metaphor pairs exist even for "primary" metaphors such as MORE IS UP. Vertical altitude can, like most other qualities, also *grow* and *shrink*, be *big* and *small*, be *lacking* and *overwhelming*, etc.:

(2.12) UP IS MORE

(a) *The image has a too big/small height*

[2] Note that I do not claim that these expressions are instances of a single bidirectional metaphor. The debates about bidirectionality, reversibility, and symmetry of metaphors are thus not directly relevant to my point here.

(b) *There was not enough height to stand*

(c) *There was only a couple of feet left*

(d) *These shoes will add an inch to your height*

(e) *At what age will a boy stop growing height?*

I do not mean to suggest by these examples that heights are actually comprehended by analogy with sizes. Rather, my point is that the standard methods of cognitive metaphor theory would force that awkward conclusion upon us in the light of such examples. Similarly contradictory examples can be given for other "primary" metaphors such as AFFECTION IS WARMTH (e.g. storms can *rage*, and weather can be *mild*, etc.). In other words, the problem of opposing metaphor directions cannot be marginalized by separating shallow resemblance from deep analogy — as suggested by Grady (1999) — even if we could operationalize that distinction.

In the lexicon of the English language, even concrete, bodily, experiential subject matters thus act as both target domains and source domains. This undermines the whole enterprise of identifying the ultimate sources of meaning by moving "downwards" from targets to sources. It is thus both empirically and conceptually problematic to claim that "the source domains of the metaphors come from our bodily, sensorimotor experience" (Johnson, 2007, p. 195).

2.3.2 Opposing Sensory Metaphors

There is another batch of examples which emphasizes this problem even more. These concern sensory modalities like vision, hearing, touch, and taste (Williams, 1976). Consider for instance how hues of color are conceptualized as sounds, even though sounds are also conceptualized in terms of colors:

(2.13) SIGHT IS SOUND

(a) *A tone of red*

(b) *A muted green*

(c) Danish: *skrigende lyserød* (= "screaming pink" = "shocking pink")

(2.14) SOUND IS SIGHT

(a) *A bright sound*

(b) *White noise*

(c) German: *Klangfarbe* (= "sound color" = "timbre")

By collecting and systematizing such observations, we can construct a table of metaphors directions as shown in Table 22. Some of these examples require some additional explanation:

Src\Tgt	TOUCH	FLAVOR	SOUND	SIGHT
TOUCH	—	*sharp taste, hot salsa*	*soft music, wall of sound*	*soft hue, dense color*
FLAVOR	*harsh surface, savor someone's (sweet) touch*	—	*spiced chord, sweet sound*	*unsavory pictures, sweet sight*
SOUND	"deaf nettle," "pain-deafening"	"hear the smell," "over-deafen"	—	*muted colors*
SIGHT	*flash of pain, radiating pain*(?)	*clear taste*	*bright tone, clear voice*	—

Table 2.4: The directions of sensory metaphors. Phrases in quotation marks are translated from other languages than English (see text). The question mark indicates an ambiguous example.

- The phrase "deaf nettle" which occurs in the TOUCH IS SOUND cell is a literal translation of the Danish, Dutch, and German name for *Laminum Album*, a nettle that despite its appearance does not sting.

- Similarly, "pain-deafening" is a translation of *pijnverdovend*, the Dutch word for *anesthetic.*

- "Hear the smell" is a literal translation of the Cretan Greek idiom *ακουω τη μυρωδια*, a TASTE IS SOUND metaphor literally meaning "notice the smell" or "pick up the smell."

- "Over-deafen" is a literal translation of the Danish verb *overdøve*, which literally refers to the drowning out of sound, but metaphorically also of taste. (Another TASTE IS SOUND examples not shown in the table is *a hum of chili*, an expression often used by British TV chef Jamie Oliver to describe a mild taste or "touch" of chili.)

- Reversely, *spiced chord* is a SOUND IS TASTE metaphor sometimes used by jazz musicians to describe complex harmonies with wide tonal extensions.

- Lastly, *harsh surface* is included here because the English word *harsh* historically derives from a Scandinavian word meaning "rancid" or "rotten." It thus represents a TASTE IS TOUCH metaphor.[3]

The picture that emerges from these observations is thus that is is not by accident that the concept of "basic" experience is unclear. The same evidence which suggests that metaphors map the abstract onto the concrete suggests the opposite too.

Consequently, we cannot tell by the inspection of linguistic metaphors how two domains of experience compare in terms of immediacy. Any attempt to construct such an ordering will either contradict the data or contradict itself. This further means that no average or other aggregation of the Pragglejaz dimensions can lead to a correct interpretation of all of these metaphors.

2.4 Differences in Immediacy

In the previous sections, we saw that there are immense difficulties connected with the attempts to formalize the notions of "concrete," "abstract," "literal," "metaphorical," and "basic" as they are used in metaphor theory. There are no objective criteria that reliably pick out certain parts of experience as more "basic" than others, and attempts to isolate experience from language often bring the two in conflict.

One explanation for this could be that we were not trying hard enough: Perhaps in a few years, cognitive metaphor theory will have gotten its conceptual house in order, and its hypotheses will be crisp and empirically substantial. Another explanation is that the project rests on false assumptions: If there is no stable entity hidden underneath the term "immediate experience," then it should not come as a surprise that attempts at defining it produce confusing and inconclusive results. In this section, I will pursue this latter option by means of three case studies.

2.4.1 Cognition, Meaning, and Language Use

The assumption that underlies the case studies in this section is that people experience the world differently, but have to talk using the same language. A consequence of this assumption is that the metaphors of a language like English may not reflect the way any single person thinks or perceives the world.

Communicative pressures dictate that you should use your language in a way that is intelligible to your peers, not in a way that faithfully expresses your private experience. The on-average facts about language that we find recorded in

[3] I owe this example to an anonymous reviewer of the paper on which this chapter is based.

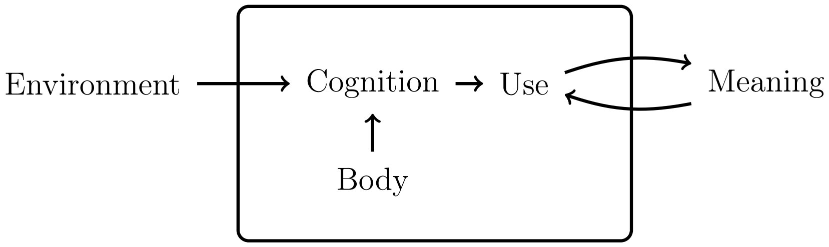


Figure 2.3: A schematic diagram of the relationship between shared meaning, individual cognition, and language use (cf. Pearl, 1988). Variables inside the plate differ between individuals, variables outside are shared (cf. Buntine, 1994).

a dictionary are thus the result of a compromise between cognitive and communicative factors. A dictionary is likely to contradict the experience of many or even all the members of a speech community.

Differences in personal experience is one source of variability that makes an inference from shared language to individual psychology problematic. Another source of variability is the time dimension: The world changes, and so do the skills, strategies, and cognitive styles of the people in it. Your relationship to wind, rain, salt, hunger, horses, roads, and fields is very different from that of a late medieval peasant or sailor. It would therefore be a bit of a stretch to say you had the same "grounding" for metaphorical uses of *plow through*, *nip it in the bud*, *know the ropes*, or *anchor*.

The situation can be schematically depicted as in Figure 2.3. According to this informal model, meaning is determined by aggregate language use. Individual language use is determined by two forces, shared meaning and individual cognition. Cognitive and linguistic styles thus vary between individuals, while meaning is assumed to be shared.

According to this model, meaning is thus a summary of usage styles, like a purple formed out of my red and your blue. Individual language use is a compromise between this aggregate and the individual's idiosyncrasies.

In the terminology suggested by this diagram, cognitive metaphor theory is the simplified model we get if we remove the "plate," that is, if we assume that all variables are shared across individuals. In such a situation, we would expect meaning, use, and cognition to converge, so that there would be no need to distinguish between them.

Instead of using this cyclic diagram, we could also make the time dimension explicit by unfolding it into an acyclic chain. This requires us to index the nodes by points in time, as sketched in Figure 2.4 (cf. Spirtes, 1995). In this diagram, cognition and its upstream influences are for convenience packed into a single

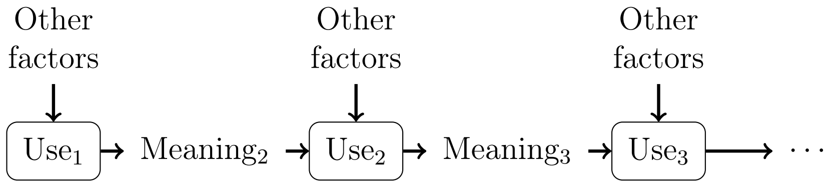


Figure 2.4: Usage patterns can change in response to non-linguistic factors, but the effect of these changes are moderated by the conservative effect of past meaning.

node called "other factors."

The form of this chain indicates how changes in the environment (e.g., technological changes) might influence meaning: In each iteration, a new use variable is created, but its value is influenced both by the previous meaning and by the "other factors." Thus, even if a change in some external factor brings about a radical change in cognition, the old meaning may partly cancel out this change before it is absorbed into the new meaning.

A statistical analogy might clarify these dynamics. Suppose we model the cognitive biases of a member of a speech community as a bag of red and blue marbles, and the pool of shared meaning as another bag, perhaps a larger one. We can then model each individual's choice of language use as the drawing of a marble, either from the shared meaning-bag or from the individual's own cognition-bag. The relative probabilities of these two sources of language use can determine how strong the normalizing effect of the meaning is.

The influence in the opposite direction — use upon meaning — can be modeled by replacement: After selecting a use-marble, the individual replaces a marble in the meaning-bag by a marble of the same color as the selected use-marble (cf. Mahmoud, 2008). This will have the effect that the composition of the meaning-bag gradually comes to resemble the average cognition-bag, even if the variance across the cognition-bags remains high. Only a change in a causally upstream variable (body or environment) can change the contents of a cognition-bag.

The upshot of this story would then be that the shared meaning of this language would gradually come to resemble the kind of object that cognitive metaphor theory studies: A stable, predictable system with systematic patterns. However, the explanation would be different, and we would do away with the assumption that there was a one-to-one correspondence between shared language and individual cognition.

This rather speculative discussion hopefully clarifies the theoretical assumptions underlying this chapter. In the following two subsections, I will substantiate them a bit by taking up a couple of empirical issues. These issues will provide some evidence that we really do need the additional complexity suggested here in order to properly understand the phenomenon of metaphor. The first two examples illustrate how deviations from the "average body" need not change language use, since the force of normalization may counteract the effects of the deviation. The third example shows that the environment might change the cognition of most members of a speech community, triggering slow changes in meaning.

2.4.2 Sensory Loss and Language Conservativity

Languages like English are full of visual metaphors like *I'll see you around*, *I see what you mean*, etc. Cognitive metaphor theory claims that these metaphors are "grounded" in first-hand visual experience and comprehended by means of analogies between vision and other domains.

This claim sits uneasily with the fact that blind English-speakers use and understand such phrases too. Even children who are born blind spontaneously use words like *see* and *look* in roughly the same way as sighted children: They go to *see their friends*, *see what you mean*, and *see how things work* (Landau and Gleitman, 1985, p. 89). This contradicts the "grounding" hypothesis: Visual metaphors do not always rely on visual experience.

Puzzled at being confronted with expectations to the contrary, a theologian who lost his sight in his forties has noted:

> When I use expressions like these, some of my sighted friends are surprised. They laugh, perhaps teasing me, and say, 'You don't really mean that, do you John?' I explain that, when I say I am pleased to see you, what I mean is that I am pleased to meet you, pleased to be with you, glad to be in your presence. I explain that this is surely what anybody, blind or sighted, would mean by that expression. (Hull, 1990, p. 21)

Thus, for the individual speaker, visual metaphors do not require first-hand experience. They can be intelligible after that experience fades away, or even if it never existed in the first place.

The problem, however, extends beyond overtly visual language, due to the many ways in which visual capacities effect lived experience. For instance, people who were born blind are generally both faster and more accurate at assessing the emotional state of a speaker based on the sound of the voice (Klinge et al., 2010).

This way of experiencing the world does not fit very well into the sweeping generalizations about "our" conceptual system one finds in the literature on cognitive metaphor theory. Consider for instance Lakoff and Johnson's claims about the cognitive roots of *face* metonymies:

> ... the metonymy THE FACE FOR THE PERSON is not merely a matter of language. In our culture we look at a person's face—rather than his

> posture or his movements—to get our basic information about what the person is like. (Lakoff and Johnson, 1980a, p. 37)

Contrast this with the following introspective report from a marine biologist who lost his sight at age three:

> ... the face is only part of the whole person. The voice—its quality, its intonation, the use of language—is unique and every bit as informative as the face. I can detect surprise, disgust, pleasure, boredom, dishonesty, thoughtfulness, and a hundred other states of mind from the voice. ... Even physical movements provide a fund of information. A quick, steady gait leaves me with a very different impression from a slow, uncertain shuffle. (Vermeij, 1997, p. 17)

As this example illustrates, the effects of blindness on experience cannot be contained within a neatly delineated domain. At the very least, it profoundly effects the experience of sound and movement too (cf. Hull, 1990, pp. 19–20).

Lastly, the loss of sight also notoriously changes people's relationship to the sense of touch. One study with sighted subjects showed that even a temporary deprivation of vision substantially improves a person's ability to read Braille characters. In fact, simply wearing a blindfold for five days turned out to be a more effective way of improving Braille discrimination skills than a full-time intensive Braille course (Kauffman et al., 2002). Once the blindfold was taken off, these effects disappeared almost entirely within a single day, as did the corresponding changes to the participants' brains (Merabet et al., 2008).

Again, these observations undermine many of the speculative ideas put forward by cognitive metaphor theorists. For instance, the mapping MORE IS UP is supposed to be

> ... grounded via the correlation between quantity and verticality – you pour more water in the glass and the level goes up; (Lakoff, 2008, p. 24)

but for many visually impaired people, sound and weight are more important indications of quantity than height. They thus live in a world in which MORE is correlated with HEAVY or HIGH-PITCHED, rather than with UP. However, this does not mean that they have to use an exotic "blind English" sociolect in which they talk about *heavy numbers* or fail to understand phrases like *a high salary.*

Many, many more examples of this kind could be given. People change their behavioral strategies in innumerable ways in response to sensory loss, causing corresponding changes in their bodies and experiential worlds (Macpherson, 2009; Pascual-Leone et al., 2011). Hence, people who speak the same language do not necessarily "see" the world the same way.

2.4.3 Handedness and Language Conservativity

As a second example, consider the difference between left and right. In some languages, particularly in English, *right* generally comes with positive connotations, while *left* comes with negative ones:

right morally good, justified, or acceptable; true or correct

left gone away from; abandoned; allowed to remain behind

Cognitive metaphor theory could explain this phenomenon as an expression of our inherent bias for the dominant hand, which we might subconsciously feel is better at tasks that require strength or precision, while the left side of the body feels relatively more "dead," clumsy, and awkward.

The problem with this explanation, of course, is that not everybody is right-handed. This makes a bigger cognitive difference than one might think; a number of recent studies have illustrated several subtle ways in which differences in handedness may affect how people experience the world.

For instance, Casasanto and Henetz (2012) asked 126 children to point out which of two stuffed toy animals they liked the most. They found that a significant majority of the right-handed children preferred animals on the right, while left-handed children tended to prefer the ones on the left (Casasanto and Henetz, 2012, pp. 7–8). These children thus intuitively projected their sense of comfort with the dominant hand onto the world itself, judging objects graspable by the dominant hand as inherently more likable. This subjective experience is at odds with the norms explicitly expressed by the English language, as indicated above.

Similar phenomena also occur with adults. In a different study, Casasanto and Jasmin (2010) recorded the hand gestures that the presidential candidates used while making negative and positive utterances during the US presidential debates of 2004 and 2008. Again, they found a significant association between dominant side and positive valuation: The two left-handed "subjects," Obama and McCain, tended to gesture with their left hand when they made positive utterances; the two right-handed "subjects," Bush and Kerry, did the opposite.

Along with the well-known differences between the metaphor systems of closely related languages, this amounts to a double dissociation of language and experience: Shared experience does not reliably imply shared language, and shared language does not reliably imply shared experience.

There is thus something deeply misguided about seeing language as a "window to the mind" (Gibbs, 1996, p. 310; Handl and Schmid, 2011). Even the way we talk about our own hands, heads, hearts, and bones is shaped by the normalizing forces of language. When the denial of those effects is elevated to the status of a psychological methodology, it is bound to obscure differences between people and their bodies.

2.4.4 Cultural Change and Its Semantic Consequences

In the first part of this chapter, I argued that that the seemingly unproblematic notion of "basic meaning" is difficult to define independently from linguistic considerations. However, one might suspect that this largely is a theoretical matter, and that the etymological record speaks its own clear language.

And indeed, many words do have etymologies that conform to intuitions about "basicness," as shown by Sweetser (1990). For instance, a number of English modal verbs have, according to current historical reconstructions, developed from root meanings that pertain to physical ability. (Consider for instance the ambiguity of the English word *might.*)

In other cases, however, things are not so clear. Consider for instance the English word *office.* According to the OED, *office* has 11 different meanings, of which five are particularly central. In very abbreviated form, these are:

OED 1 Regularly repeated Christian ritual (*Office for the Virgin*)

OED 2 Official position (*run for office*)

OED 3 Duty; assigned function (*perform one's office*)

OED 6 Workroom; business or government branch (*post office*)

OED 9 Ceremony or rite (*the last office for the dead*)

From the perspective of cognitive metaphor theory, the hypothesis would be that one of these word senses is more basic than the others, and that the more abstract ones are derived from this root sense by metonymical and metaphorical extensions (Lakoff, 1987). Presumably, a hypothetical cognitive metaphor theorist would judge "workroom" to be the most basic of all the senses and then go on to explain how the "duty" and "position" senses were derived from that.

In order to get a sense of how well this hypothesis fares, I collected all the examples cited in the OED entry for *office* and categorized them according to century. This amounted to 333 example sentences, including the compounds that referred to a single, unambiguous sense (e.g. *office piano*). The full data set is given in Appendix B.

This data set gives us an indication of how frequently the different word senses were used in various centuries, as shown in Figure 2.5. Of course, this estimation method rests on the assumption that the distribution of examples in the OED follows the frequency of the word sense in actual usage. This is not entirely true, but the inclusion of the compound phrases probably mitigates this problem somewhat, since frequent word senses probably tend to spawn more compounds.

Keeping these caveats in mind, the data does not seem to be consistent with the prediction one would get from cognitive metaphor theory. The "abstract" senses of *office* associated with duties or religious service (1, 3, and 9) were

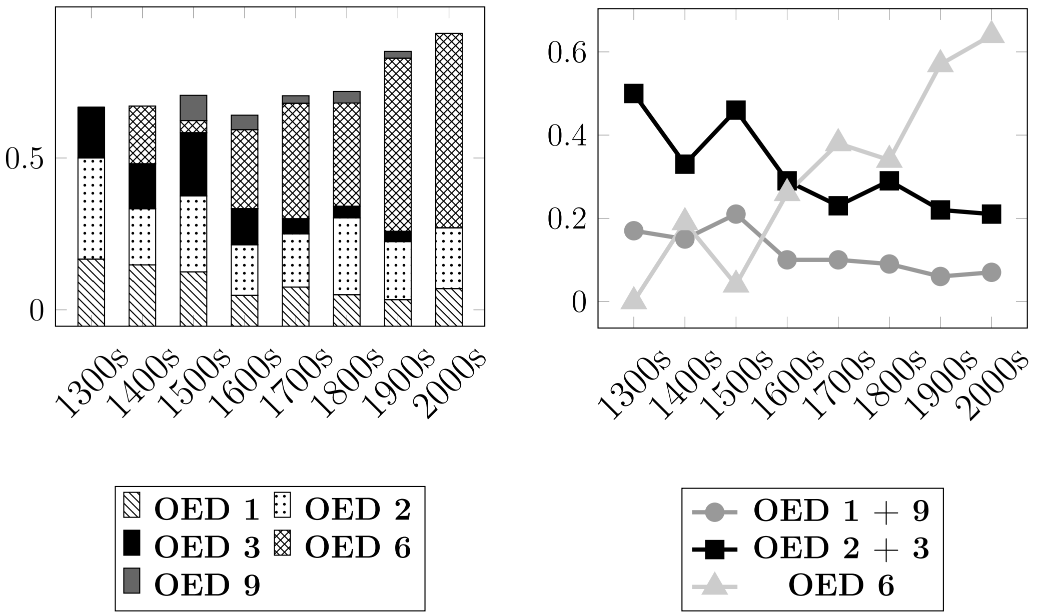


Figure 2.5: Frequency changes in the use of the English word *office*.

quite frequent in the early centuries but decreased in frequency over the charted period. The senses related to non-manual labor (6), on the other hand, increase in frequency. The sense denoting an official position (2) remains stable.

Neither of these patterns make very much sense in terms of extensions from more "basic" to less "basic" domains. Instead, it appears that the underlying mechanism driving the change in the frequencies is the gradual decline in the importance of Christianity in England, and the increasing prevalence of banking and urban capitalism from about 1600 onwards.

Similar points could be made with respect to other words. A small selection of informal examples are shown in Table 2.5, but more could quite easily be produced. Due to the conceptual problems discussed in the first part of this chapter, it is difficult to quantify exactly how large this problem is, but if environment and cognition are related as suggested in Figure 2.3, then this empirical phenomenon is based on a general and productive mechanism, not a freak accident.

2.5 Conclusion

According to cognitive metaphor theory, language is grounded immediate, first-hand experience (Gibbs, 2005; Johnson, 2007; Lakoff and Johnson, 1999). This grounding hypothesis holds the promise of explaining all our abstract knowledge

word	Older meaning	Newer meaning
bill	written statement	paper money
code	collection of laws	encryption key
cure	take care of	treat medically
design	intention	shape
document	teaching, instruction	written record
plastic	malleable (adj.)	a synthetic material (n.)
rifle	scratch or groove	large firearm
train	delay; trail; entourage	line of railroad carts

Table 2.5: A selection of semantic extension that may, depending on one's assumptions, be difficult to construe as moving from the concrete to the abstract.

in terms of familiar bodily experiences like walking, eating, lifting, breathing, and so on.

At the level of the individual speaker, however, this equation between language and experience does not hold. People have different bodies and different environments, and they do not gear into the world in the same way.

This fact is easy to forget for a linguist studying the products of communication, since communication by definition requires us to overcome our differences. Language has a tendency to steamroll any idiosyncrasies that you and I have. For this reason, a psychology built on linguistic observation will always have a tendency to replace the cognition of the actual individual with cognition of an idealized average individual, and this fictional person might not have much in common with any single member of the speech community.

Cognitive metaphor theory thus has trouble explaining facts that have to do with language conformity and conservativity, as opposed to the immediate expression of inner experience. It fails to explain why certain metaphors are not extended in certain ways; why they differ so much between the languages of closely related cultures; how they can refer to experiences that were forgotten generations ago; and, as I have discussed at length in this chapter, it misleadingly suggests that the notion of "immediate experience" can be defined without reference to the contingencies of who you are and what you spend your life doing.

This does not mean that analogical reasoning is uninteresting, never happens, or has no place in metaphor comprehension. It probably plays a large part in many cases. But it does mean that a dictionary of English cannot be read as a textbook in psychology. What happens on the page will often not be what happens in the head.

Cognitive metaphor theory would thus be better off dropping the counterproductive dogma that linguistic behavior is a direct expression of private experience. We do not need to agree in order to talk, and we do not need first-hand experience in order to understand a metaphor. Acknowledging this point will rule out a number of generalizations about "how we think," but on the other hand, it might also rule a number of people back into the definition of who "we" are.

2.6 Appendices

A Sample Translations of Questionnaire Materials

The experiment reported in the main text used 22 different pairs of sentences in Dutch. Below is a translation of 12 of these pairs.

Danish	Translations
knop	Plant lice are small insects which feed on the buds of roses. Acne is a common skin condition that causes pimples on the face.
balde	After a tetanus shot, I was unable to sit on one buttock for 10 days. I was thinking about buying a bale of hay.
rage	The woman biked from the station into Stationsstraat and was groped by the man. Today, we encounter more and more men with their head shaved bald.
skam	She thought that was a real pity. She was filled with shame.
skraber	During shaving of the legs with a razor, the razor will frequently have to be washed. I wonder why a healthy man or woman need to take nap in the middle of the day.
skrald	Every time she slams the door, a loud bang can be heard through the house. It is difficult to prove that the stench comes from the garbage from the restaurant.
snurrende	You've fallen asleep on the couch, and you have a tingling sensation in your leg. The double-traction wheel is named after its traction belt which runs twice around the spinning wheel.
synke	The most important symptoms of a sore throat are hoarseness and pain when you swallow. Older eggs float on water, fresh eggs sink.
takker	You can also see the difference by looking at the jags on the edges of the leaves. I consequently turned the QPI voltage up a couple of notches.

rør	In modern telephones, the loudspeaker in the receiver often also works as bell. During showering, running cold water is heated up in the pipe.
prop	More and more wines are sealed with a screw cap instead of a cork. The blood clot can be big or small, and there may be more than one clot.
rykke	The player moves a piece. The dog yanks at the leash.

B Frequency of *office* by Word Sense and Century

The OED lists 11 meanings of the word *office* and provides examples ranging in time from the 14th to the 21st century. The following table shows the number of examples provided for each words sense and each century:

	Century							
Sense	14th	15th	16th	17th	18th	19th	20th	21st
1	3	4	3	2	3	4	3	1
2	6	5	6	7	7	20	17	3
3	3	4	5	5	2	3	3	0
4	4	4	2	6	3	4	3	0
5	1	0	2	4	3	4	2	1
6	0	5	1	11	15	27	51	9
7	1	3	1	1	5	6	5	0
8	0	2	2	2	1	1	2	0
9	0	0	2	2	1	3	2	0
10	0	0	0	2	0	0	0	0
11	0	0	0	0	0	7	1	0
Sum	18	27	24	42	40	79	89	14

Chapter 3

Back in Time is Not Back in Space: New Evidence from Spontaneous Gesture

The fact that people often gesture in ways that illustrate the metaphors they use has repeatedly been invoked as evidence in support of a cognitive theory of metaphor. Since the gestures illustrate the literal meanings of the metaphors, the argument goes, the metaphors in speech and gesture must derive from a single cognitive source. In this chapter, I evaluate this argument against a relatively large data set of video clips showing university lecturers that gesture while they talk about the past in terms of a metaphorical timeline. The gestures used by these speakers systematically follow a pattern which is inconsistent with the timeline of spoken English. This suggests that either gesture, speech, or both are unreliable indicators of how people actually think.

3.1 Introduction

Pragmatic theories of metaphor comprehension often struggle with the fact that novel metaphors often seem to be understood both rapidly and effortlessly. For instance, the analogy between size and upwards extension is not limited to conventionalized expressions like *prices went up*, but also extend to more colorful variants like *prices soared*, *skyrocketed*, *nosedived*, and so on.

The ease with which we understand such novel expressions is difficult to explain on a dictionary account of metaphor comprehension in which a listener only has two possible strategies, cheap recall or expensive reconstruction. For this reason, an alternative account has been proposed, according to which the unpacking of metaphorical expressions draws on nonverbal analogies that already exist in the head of the listener (Lakoff and Johnson, 1980a; Gibbs, 2005; Lakoff, 2008).

The meaning of *prices went through the roof* is thus transparent to us because we already think about amounts in terms of heights, so that no decoding or guesswork is involved. Under this account, the words on the page, like *prices went up*, “are not dead metaphors, but reflect active schemes of metaphorical thought” (Gibbs, 2011, p. 532). This displaces part of the comprehension effort from online thought to long-term memory, and this could explain why comprehension can often be achieved so smoothly.

However, such an explanation evidently relies very heavily of the power of a certain unobservable entity, the cognitive analogy in the head of the hearer. This naturally raises the question of what independent evidence we have of the psychological reality of this hypothesized cognitive infrastructure.

A number of sources of evidence pertaining to this question have been examined in the course of the last decades. This chapter focuses on one of the most interesting and convincing ones, the spontaneous gestures that accompany metaphorical speech (Lakoff and Johnson, 1999; Gibbs, 2005). If these gestures can be shown to reflect the same imagery as the spoken word, then that would provide strong evidence of a common nonverbal cause behind these two channels of expression.

I accept the force of this argument, and I think there are many pieces of evidence that do support its conclusion. However, as I will argue in this chapter, there are also other pieces of evidence which point in the opposite direction, that is, cases in which hands and words speak different languages (cf. Cienki, 2008). These problematic cases raise doubts about the straightforward conclusion we could otherwise draw from the supportive evidence.

From a logical and methodological perspective, this leaves us with an important choice to make. We can either continue to believe that gestures tap directly into the depths of the mind, with the mixed empirical conclusions that this entails; or we can insist on the soundness of the cognitive theory of metaphor, leaving us with the conclusion that the gesture evidence should really not have been taken seriously in the first place.

3.2 Background

The notion that gestures as well as speech could express metaphorical content has been around since the work of McNeill (1992), who pointed out that the topic of a discussion often is treated as a physical object that can be held up, inspected, put forward, swatted away, and so on.

Observations of this kind have later been placed more systematically within the framework of cognitive metaphor theory by Núñez, Sweetser, Müller, Cienki, and others. A paradigmatic example of this cognitive-metaphorical approach is Núñez’ examination of video recordings of lectures on college mathematics.

In mathematics, the distinction between formal content and informal imagery is in principle clear-cut, with imagery playing only a secondary role. However,

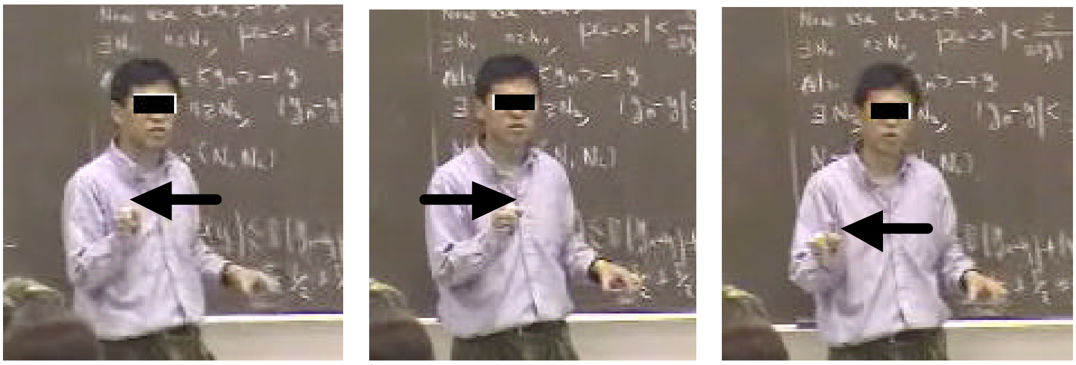


Figure 3.1: Figure 5 from Nunez (2006), showing a math professor performing an "oscillating" gesture as he talks about an "oscillating" function.

as Núñez pointed out, the clean distinction between definition and illustration is frequently violated when math teachers gesture in the classroom.

He could for instance exhibit an example in which a professor of mathematics told students that a function "oscillated," moving his hand back and forward to illustrate the metaphor (cf. Fig. 3.1). This is technically speaking in conflict with the set-theoretic definition of a function, which represents the function as a static set of input-output pairs. By looking for such tell-tale gestures that revealed bits of metaphorical thought, Núñez argued that the informal metaphors of mathematics, such as functions that can *move*, *grow*, and *oscillate*, in fact are active patterns of thought in the heads of mathematicians, even if the mathematicians themselves would deny the importance of such imagery.

In a similar type of argument, Núñez and Sweetser (2006) presented evidence that speakers of the American language Aymara gesture in a way that cohere with their peculiar conceptualization of time.

Aymara is an indigenous language spoken mainly in Bolivia and Peru. As in English, Aymara-speakers talk about time as a linear sequence of events, spread out in a one-dimensional space like beads on a string. However, unlike in English, the Aymara timeline places the past in front of the speaker and the future behind (Tarifa 1969). In Aymara, something "far back" in the past thus happened

(3.1) *ancha* *mayra* *pachana*
a-lot front time-in
"a long time ago"

Similarly, events in the future are behind the present, as in

(3.2) *qhipa* *marana*
behind year-in
"next year"

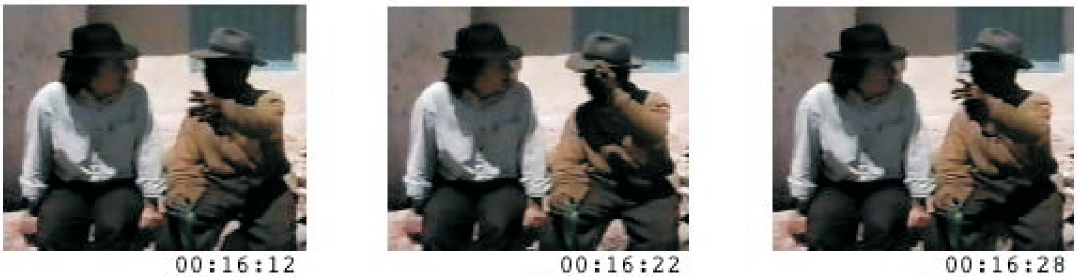


Figure 3.2: Figure 6 from Núñez and Sweetser (2006, p. 428), showing an Aymara-speaker gesturing forwards while using the Spanish phrase *tiempo antiguo* ("ancient times"). The time stamps show minute, second, and frame.

If this system of time reference is indeed an ingrained part of the cognitive furniture of an Aymara-speaker, we would expect them to gesture backwards when talking about the future, and ahead when talking about the past. This system of gestures was in fact exactly what Núñez and Sweetser found when interviewing their Aymara-speaking informants (in a mixture of Spanish and Aymara).

For instance, one interviewee would use the Spanish phrase *tiempo antiguo* ("ancient times") while gesturing forward in a chopping motion (cf. Fig. 3.2). Another would use the phrase *tiempo futuro* ("future times") while swinging a hand backwards, as if throwing something over his shoulder.

The preference for this "Aymara timeline" neatly followed the preference for speaking in Aymara. Of the ten informants who consistently used forward-backward gestures when talking about time, the ones who were more fluent in Spanish tended to use the European system of directions, while those who were more fluent in Aymara used the Aymara system (cf. Table 3.1).

This seems to confirm the cognitive-metaphorical hypothesis. A single system of time references appeared in both speech and gesture, suggesting a common psychological principle behind them both. The reversed timeline in Aymara might thus be a matter of nonverbal though rather than merely words. This provided a compelling argument in favor of the hypothesis that metaphors reflect online analogical thought.

Núñez and Sweetser implicitly contrasted the Aymara system of gestures with a presumed European system which associates the past and future with the back and front, respectively. More recent evidence, however, suggests that the situation in English is more complicated than that.

In one pilot study, Casasanto (2008) found that English-speakers in fact do not gesture forwards and backwards when they talk about time, but rather in a sideways fashion. Rather than depicting the past as being behind them, as suggested by the English metaphors, they gesture as if they they imagine time

Predominant language	European timeline	Aymara timeline	Total
Spanish	5	1	6
Aymara	0	4	4
Total	5	5	10

Table 3.1: Out of Núñez and Sweetser's 30 informants, 10 consistently used sagittal gestures (front-back) when talking about time. Among these 10 people, the direction in which they gestured depended significantly on which language they spoke most fluently (since the probability of drawing 0 black marbles from a jar of 5 black and 5 white is only $p = .0238$).

passing by in front of them, from left to right.

This is somewhat surprising for many people. In fact, Casasanto and Jasmin (2012) have found that English-speakers systematically misrepresent their own time gestures when they are explicitly asked to demonstrate them. When asked to show how they gesture while saying things like *way back in the past* or *a generation after that*, Casasanto and Jasmin's subjects tended to use forward-backwards gestures.

However, in a different experiment (involving a story-telling task), this pattern flipped around. Rather than observing front-back and left-right gestures in a ratio of about 3:2, as was the case in the deliberate demonstration experiment, the ratio was closer to 1:3 in the story-telling experiment. The gestures were also more erratic in the directions they used, particularly among the front-back gestures, which were practically random in the story-telling experiment (Casasanto and Jasmin, 2012, pp. 655–56).

These findings raise some serious question about Núñez and Sweetser's findings. If English-speakers do not gesture in accordance with the metaphorical system in they language, how much importance can we then assign to the observation that Aymara-speakers do? It would seem that there are only two options left, either deciding that the evidence from Aymara is not particularly compelling after all, or finding some way of marginalizing the evidence from English as irrelevant, misinterpreted, or wrong.

This seems to place a lot of logical weight on a relatively little data. In the following sections, I will thus present a new data set consisting of spontaneous gestures produced by English-speaking academics in teaching situations. A closer inspection of this data set will brings us one step closer to knowing just how literally we should take a phrase like *back in time.*

3.3 Materials

Yale University publishes a large number of their lectures as videos on the Open Yale Courses website, `http://oyc.yale.edu`. These videos are subtitled and transcribed, and therefore easily searchable. The data used in this chapter are extracted from this resource by searching for phrases that are likely temporal metaphors and then coding the observation according to what gesture the speaker made at the relevant time.[1]

I focused on three phrases that I suspected would provide me with a large set of temporal metaphors:

(3.3) *way back*

(3.4) *far back*

(3.5) *back in time*

In the lectures that I included in my search, spanning the years 2006 to 2011, these phrases occurred 236 times in total. They are mostly spoken by the lecturer, and they occur in courses on finance, chemistry, history, astronomy, game theory, Bible studies, biology, philosophy, and other topics.

These examples were then hand-tagged. I read the context surrounding the use of the phrase to determine whether it was used in a temporal sense or not. This would involve distinguishing between, for instance, cases like the following:

(3.6) *. . . going all the way back to the sixteenth century.* (temporal)

(3.7) *. . . I have to find my way back there . . .* (nontemporal)

For each temporal phrase, I then watched the corresponding clip to see whether the speaker was visible during the relevant time, and whether he or she produced a gesture while using the target phrase. If so, I recorded whether the phrase was directional or not.

This left me with a database of 54 directional gestures. These gestures were categorized according to their direction, that is, left-to-right, right-to-left, back-to-front, or front-to-back.

There is some measure of subjective judgment involved in this categorization, and it would obviously have been preferable to have have several independent coders review the data set. However, since every single data point is available online and can be checked by skeptical readers, I felt less concerned by this possible source of error than I would have been in the case of a private and anonymized dataset.

A summary of the relevant dimensions of the data is shown in Table 3.2. For a more full account of the data set, with precise references to each individual lecture, see Appendix B.

[1] For legal details, see Appendix A.

Totals	
way back	174
far back	30
back in time	32
Total	236

Gesture types	
Not temporal	60
Not visible	19
No gesture	78
Not directional	25
Directional	54
Total	236

Gesture directions	
Front-to-back	9
Back-to-front	9
Left-to-right	10
Right-to-left	26
Total	54

Table 3.2: Overview of the tagged data.

3.4 Analysis

The data is not consistent with a hypothesis which postulates a large overweight of forward-to-backward gestures accompanying the forward-to-backward temporal language. In fact, the forward-to-backward gestures were the least frequent ones in the whole data set. Consistent with Casasanto and Jasmin's findings, the most frequent gesture instead was a right-to-left gesture, suggesting an analogy with process of reading and writing rather than with the English time metaphors.

A more detailed analysis also indicates that the data set is unlikely to have come from a distribution in which front-to-back gestures were the most frequent gesture. This can be seen in two different ways.

The first of these is a Bayesian analysis. We can assume a generative model in which a parameter vector

$$\alpha \;=\; (\alpha_1, \alpha_2, \alpha_3, \alpha_4)$$

is generated according to a uniform distribution on the simplex of four-dimensional probability vectors. A sample

$$x \;=\; (x_1, x_2, x_3, x_4)$$

is then drawn from a multinomial distribution with size parameter $N = 54$ and with the probability parameter α. Since the observation we actually have is $x = (9, 9, 10, 26)$, the posterior probability density over the parameter vector is given by the proportionality

$$p(\alpha \,|\, x) \;\propto\; \alpha_1^9 \alpha_1^9 \alpha_1^{10} \alpha_1^{26}.$$

The constant of proportionality can be found by integrating this function over the whole probability simplex. The probability that α_1 is larger than or equal to the three other parameters is the integral of this function over the set of parameter vectors that satisfy the three size costraints $\alpha_1 \geq \alpha_2$, $\alpha_1 \geq \alpha_3$, and $\alpha_1 \geq \alpha_4$, divided by the integral over the whole parameter space. By numerical integration, I found this number to be

$$P\left\{\alpha_1 \geq \alpha_2, \alpha_3, \alpha_4 \,|\, x\right\} \;=\; .0019.$$

According to this model, the posterior odds in favor of forward-to-backward gestures being the most common direction is thus somewhere in the vicinity of 530 to 1. It is, in other words, highly unlikely that the forward-to-backward gestures are the most common, in contrast to what one might predict based on a cognitive-metaphorical analysis.

The second analysis I will apply relies on maximum likelihood estimates of the parameter vector. The approach here is to see how well the data is explained by the best model that respects the size constraints on α. This optimal model can either be compared to an alternative without restrictions on α, or with the data itself.

The selection of the best model is an optimization problem that can be solved by means of Lagrange multipliers. In the absence of any size constraints on α, the maximum likelihood estimate of the parameter vector is the frequency vector,

$$\hat{\alpha}_{\text{out}} \;=\; \left(\frac{9}{54}, \frac{9}{54}, \frac{10}{54}, \frac{26}{54}\right).$$

Since this vector does not respect the size constraints, the best model that does will lie somewhere on the boundary of the set $\{\alpha_1 \geq \alpha_2, \alpha_3, \alpha_4\}$. Again using Lagrange multipliers to inspect the boundary of this set, we find a maximum likelihood estimate of

$$\hat{\alpha}_{\text{in}} \;=\; \left(\frac{17.5}{54}, \frac{9}{54}, \frac{10}{54}, \frac{17.5}{54}\right).$$

Each of these parameter settings assigns a specific likelihood to the data, about 3.1×10^{-3} and 4.2×10^{-5}, respectively. The data thus favors the unconstrained model by a factor of about 74.3, speaking against the hypothesis that gestures are more likely to move backwards than in other directions.

This conclusion can be quantified in the form of an error probability. When $x = (x_1, x_2, x_3, x_4)$ is drawn from a multinomial distribution whose expected bin counts are $\bar{x} = (17.5, 9, 10, 17.5)$, the marginal distributions of the individual bin counts x_i are almost Gaussian. This means that the sum of the normalized errors, $\sum_{i=1}^{n}(x_i - \bar{x}_i)^2/\bar{x}_i$, closely resembles a χ^2 variable (cf. Fig. 3.3).

In other words, if the cell counts x are drawn from a multinomial distribution with parameter vector $\hat{\alpha}_{\text{in}}$, then the sum

$$\chi^2 \;=\; \frac{(9-17.5)^2}{17.5} + \frac{(9-9)^2}{9} + \frac{(10-10)^2}{10} + \frac{(26-17.5)^2}{17.5} \;\approx\; 8.26$$

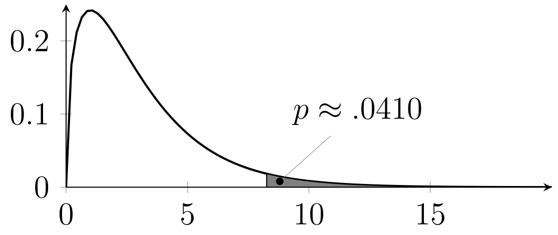


Figure 3.3: When x is drawn from a multinomial distribution with expected cell counts $\bar{x} = \alpha N$, the distribution of each individual cell count x_i is approximately normal, with mean and variance $\bar{x}_i$. Hence, $\sum_{i=1}^{4}(x_i - \bar{x}_i)^2/\bar{x}_i$ and $\sum_{i=1}^{4} 2\,(\ln p(\bar{x}_i) - \ln p(x_i))$ follow almost the same distribution, both resembling a χ^2 variable with 3 degrees of freedom. The probability mass located in the far end of this distribution is thus a reasonable measure of model fitness.

should represent a sample from a χ^2 distribution with $4 - 1 = 3$ degrees of freedom. By using this hypothetical distribution statement, we can estimate the probability of encountering a more extreme observation by measuring the mass of the tail of the χ^2_3 distribution:

$$p \;=\; \int_{8.26}^{\infty} \chi^2_3(t)\,dt \;=\; 0.0410.$$

Direct simulation of the corresponding multinomial distribution confirms the soundness of this number: When sampling from a multinomial distribution with size parameter $N = 54$ and probability vector $\hat{\alpha}_{\text{in}}$, the probability of encountering an observation less probable than $x = (9, 9, 10, 26)$ is about

$$p \;\approx\; 0.0357,$$

with the uncertainty on the last digit. Since the random variables $2(\ln p(\bar{x}) - \ln p(x))$ and $\sum_{i=1}^{n}(x_i - \bar{x}_i)^2/\bar{x}_i$ have almost identical distributions for a data set of this size, these two p-values measure almost the same kind of extreme event.

The data set thus represents an observation which is significantly less probable than we should expect under even the most charitable selection of parameter values. If the gesture directions are indeed distributed exactly in the proportions $\hat{\alpha}_{\text{in}} = ({}^{17}\!/_{54}, {}^{10}\!/_{54}, {}^{10}\!/_{54}, {}^{17}\!/_{54})$, then an observation as extreme as $x = (9, 10, 10, 25)$ should occur only about 1 in 30 times. This suggests once more that a theory which favors front-to-back gestures over other directions is unlikely in the light of this data.

3.5 Examples

As suggested by the data in Table 3.2, there are several examples in the dataset in which speakers gesture backwards while employing a metaphorical timeline to talk about the past. For instance, in one lecture on Milton's poetry, English professor John Rogers clearly waves his left hand backwards, with a flat palm facing forwards, while using the phrase

(3.8) *. . . anything that he was able to imagine way back in 1644 . . .*

(English 220, lecture 22, 51:00)

The stroke of this gesture occurs just briefly before the onset of the phrase *way back*, providing a gestural support for the metaphor. Such examples exist, and they seem to confirm the idea that the speaker is thinking in terms of a temporal space as well as talking about one.

However, as discussed in the previous section, while the theory suggests that such gestures should be in the majority, they are in fact the least common form of directional gesture. The most common family gestures by a large margin are right-to-left gestures.

One example of such a gesture comes from a lecture on the American revolution. History professor Joanne Freeman talks about slang and says that

(3.9) *. . . it actually isn't really modern, it dates all the way back . . .*

(History 116, lecture 4, 21:00)

During the time it takes her to utter the phrase *all the way back*, she rolls her left hand downwards and to her left, finishing the chopping motion just as she says the word *back*.

There is no question about the directionality in the clip; if anything, the gesture moves forward more than backwards. The timing is also matches the phrase as perfectly as one could possibly expect, suggesting that the gesture really is used as an illustration of the temporal metaphor. Yet, it uses a different direction than the linguistic metaphor suggests. As mentioned above, a large number of such examples exist, with lectures waving, chopping, sweeping, or throwing to their left as they talk about the past.

This alternative motion could potentially be interpreted as sign of a differently oriented timeline in the heads of the speakers. However, such a hypothesis is difficult to believe in the light of the erratic use of directions attested in the rest of the corpus. In one case, for instance, economics professor Ben Polak seems to throw an imaginary object forwards as he refers to the beginning of the course, using the phrase

(3.10) *. . . way back on the very first day of the class . . .*

(Economics 159, lecture 24, 0:50)

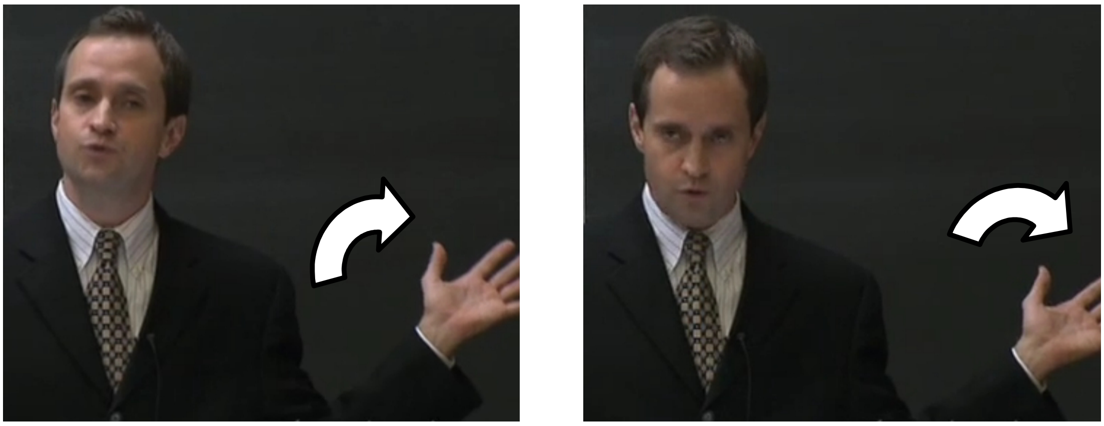


Figure 3.4: Consistent with the theoretical expectation, English professor John Rogers gestures backwards as he utters the phrase *way back in 1644*.

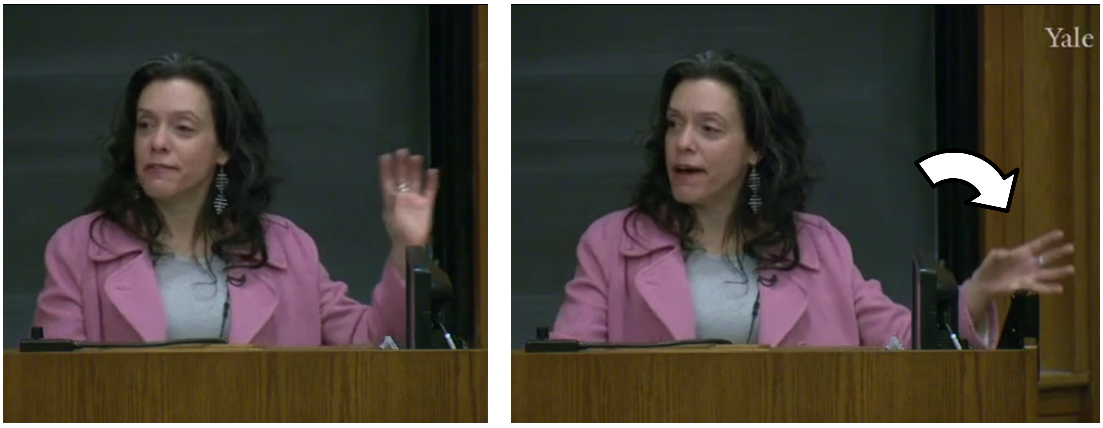


Figure 3.5: History professor Joanne Freeman gestures leftwards while explaining that a certain piece of slang *dates all the way back*.

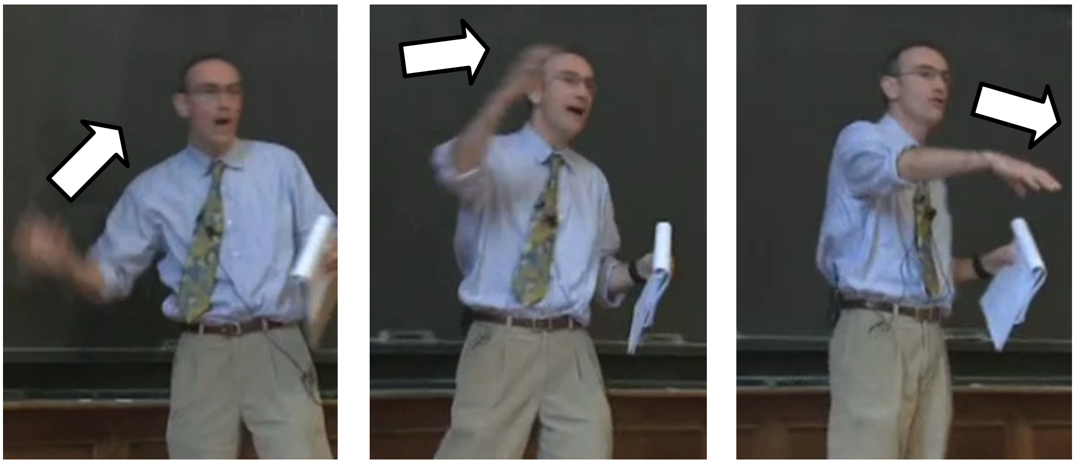


Figure 3.6: Economics professor Ben Polak gestures forwards as he refers to a topic discussed *way back on the very first day of the class*.

The gesture is timed so that his dragged-out pronunciation of *way* coincides with the swinging stroke of the gesture. Again, this finely tuned timing of word and hand makes it difficult to shrug off the example as an accident. By any ordinary standard, this forward motion would be interpreted as supporting the phrase *way back*, in spite of the obvious contradiction between the two information channels.

In a similar example, the political science professor Ian Shapiro swings his hand forwards in a pendulum motion, as if rolling an imaginary bowling ball towards the audience. This gesture is timed so as to start at the same time as the word *all* in the phrase

(3.11) *... going all the way back to the Magna Carta ...*

(Political science 118, lecture 2, 38:35)

Other examples are even more difficult to explain in terms of a mental timeline with a fixed direction. In a remark about the Oedipus myth, English professor Paul Fry uses the mixed time metaphor

(3.12) *... going all the way back to Oedipus' family history and then down through the history of his offspring ...*

(English 300, lecture 9, 19:25)

Surprisingly, he very clearly points upwards at exactly the time he says *all the way back*, and then lets his arm drop down while saying *down through.*

Although the gesture thus illustrates the latter of the two time metaphors, he uses only one motion to illustrate them. Clearly, we cannot explain both his mixed metaphor and his gesture as an expression of a single mental image. (Incidentally, this example was not coded as a directional gesture, since it did not follow one of the four canonical directions.)

A similarly inconsistent example occurs as the history professor David Blight discusses the concept of popular sovereignty:

(3.13) *... it's all the way back there in the epistles of Paul, and even before that in certain kinds of writings ...*

(History 119, lecture 5, 17:20)

At the onset of the first time reference, *all the way back*, he swings his hand to the right, throwing an invisible object away from his body. But then at the onset of the second time reference, *even before that*, he waves it back in front of his body in a clearly leftward swatting motion. (This example was tagged as left-to-right in the data set, since it is the leftward gesture that coincides with the phrase *way back.*)

Within a couple of seconds, he has thus both gestured to his right and to his left to indicate that he is referring to the remote past. Any theory that allege one of these two gestures to be faithfully representing a mental image in his head would thus have a problem explaining the other one.

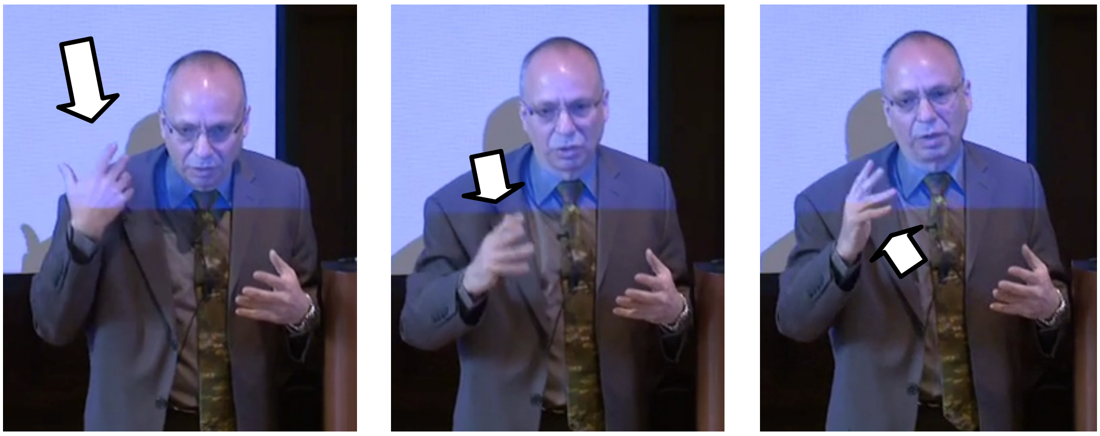


Figure 3.7: History professor Ian Shapiro gestures forwards as he explains that a certain tradition dates *all the way back to the Magna Carta*.

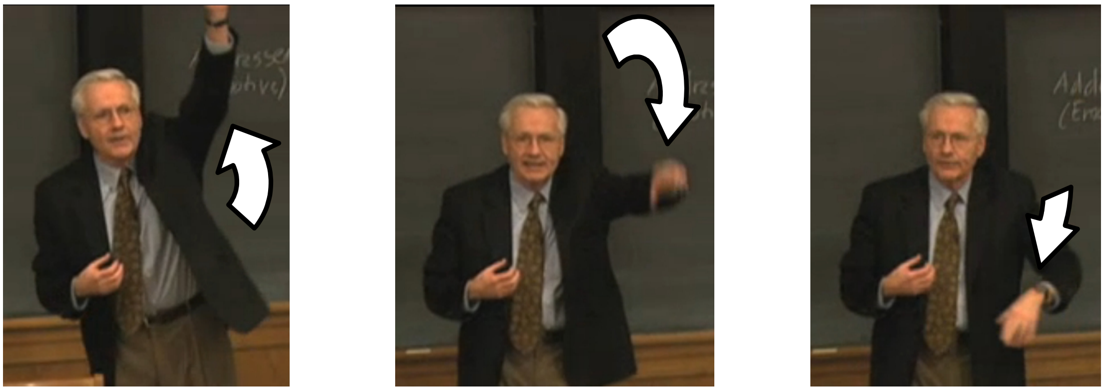


Figure 3.8: English professor Paul Fry points upwards while talking about going *all the way back* in a family history, and then *down through history*.

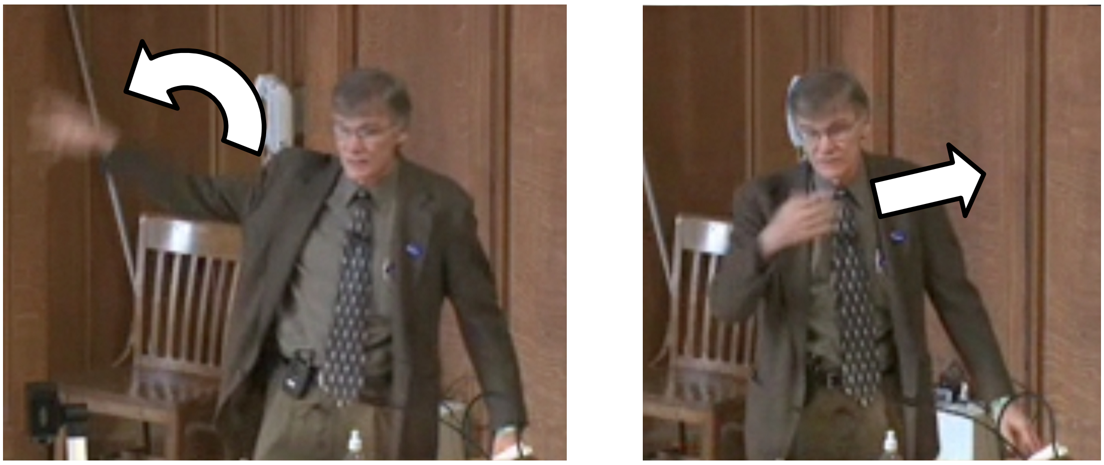


Figure 3.9: History professor David Blight gestures to his right while saying *all the way back there in the epistles of Paul*, and then gestures to his left as he refers to a time *even before that*.

3.6 Conclusion

There is undeniably a striking systematicity to the metaphors in the everyday language. Abstract entities move, wobble, run, and lead the way; behaviors are cold, sharp, sour, and bitter; and plans and projects can take off, fly, crash, or arrive at their destinations. Faced with all this imagery, it is tempting to hypothesize that the way we talk betrays a hidden pattern of thought which existed below the surface all along.

However, much of the evidence supporting this theory itself rests on substantial theoretical assumptions about the relationship between language, thought, and behavior. Often, what first appears as strong supporting evidence turns out on closer inspection to be a matter of a much more complex set of social behaviors that cannot simply be taken as a "window to the mind."

Take for instance the observation by Boers that disease metaphors seem to become more prevalent during the winter months (Boers, 1999). His survey of the opinion pieces in the *Economist* showed a neatly circular pattern in the frequency with which they described states of economic affairs with medical terms like *healthy*, *arthritic*, and *cure* rather than, say, in terms of battle or growth. This seemed to clearly indicate that the direct, personal experience of cold and sickness would be expressed through language.

However, this line of argument only holds up as long as we choose the right examples. Consider instead the linguistic effects of the massive earthquake and tsunami of 26 December 2004. A quick count of the frequency of the two phrases

(3.14) *a tsunami of*

(3.15) *a torrent of*

in *New York Times* archive 1994–2014 shows that the former became significantly more common after the catastrophe of December 2004 (cf. Table 3.3). This is most probably not due to the literal references, since the construction *a tsunami of* occurs almost exclusively in metaphors like

(3.16) *... sinking in a tsunami of red ink ...* (*NYT*, 2003)

(3.17) *... corruption and a tsunami of crime ...* (*NYT*, 2004)

(3.18) *We have a tsunami of donors.* (*NYT*, 2005)

(3.19) *... Asia faces a "tsunami" of diabetes ...* (*NYT*, 2006)

It seems highly implausible that this change in the frequency — from about 7% to about 13% — could have been brought about by an increase in direct, physical experience. Most of the *New York Times* contributors who used phrases like *a tsunami of anti-Americanism*, which occurred three times in 2004, had probably never witnessed a tsunami, much less felt one.

	1994–2004	2005–2014	Totals
a tsunami of	68	271	339
a torrent of	949	1,870	2,819
Totals	1,017	2,141	3,158

Table 3.3: The frequency of two overwhelmingly metaphorical phrases as recorded in the *New York Times* archive, 1994–2014. The result differs significantly from the expectation under an independence hypothesis, with a probability of only $p \approx 1.2 \times 10^{-6}$ in the two extreme tails of the hypergeometric distribution, on the far side of the observation (i.e., by Fisher's exact test).

The topic of this chapter has been another such case: The problematic counter-evidence that undermines the attempt to interpret gestures as a shortcut to the secrets of the mind (Lakoff and Johnson, 1999; Gibbs, 2005). As it turns out, this seemingly convincing source of evidence faces massive problems when we take the full picture into account and inspect the negative as well as the positive evidence.

This raises some serious doubts about one of the most notable sources of evidence in support of cognitive metaphor theory. While it does not necessarily mean that the theory is wrong, it does indicate that its claims need to be formulated more precisely, and that its sweeping generalizations cannot be sustained without new sources of strong corroborating evidence.

3.7 Appendices

A Copyright

The video clips discussed in this chapter are owned and distributed by Open Yale Courses, `http://oyc.yale.edu`. Copyright and other issues related to the use of these materials are discussed in their Terms of Use, found at

`http://oyc.yale.edu/terms`

The images and quotes reproduced in this text are covered by a Creative Commons license, specifically the non-commercial, with-attribution, share-alike license discussed at

`http://creativecommons.org/licenses/by-nc-sa/3.0/us`

This license prohibits the commercial use of the materials, requires the user to give appropriate credit to the source, and to reapply the same license to any derivative works. The images reproduced here (including the arrows, which were added by me) are thus also subject to the same Creative Commons license.

B Data

The data set used in this chapter contained 236 items. The table below contains references to every item that was tagged as containing a directional gesture, either back-to-front (BF), front-to-back (FB), right-to-left (RL), or left-to-right (LR). For brevity, only the 54 items included in the statistical analysis are shown in the table.

The abbreviated course names can be used to locate the original clips. For instance, lecture 3 of the course labeled "astr-160" can be found at

`http://oyc.yale.edu/astronomy/astr-160/lecture-4`

Other addresses are collected in the list at `http://oyc.yale.edu/courses`.

Course	Lct.	Time	Dir.	Course	Lct.	Time	Dir.
astr-160	3	12:22	BF	econ-252-11	23	21:48	RL
clcv-205	3	17:42	BF	engl-220	7	11:31	RL
clcv-205	13	1:06:48	BF	engl-310	22	01:26	RL
econ-159	21	03:46	BF	gg-140	1	23:04	RL
econ-159	24	00:53	BF	gg-140	27	13:36	RL
econ-252-11	10	28:57	BF	gg-140	26	29:12	RL

mcdb-150	2	11:43	BF
phys-201	16	07:01	BF
plsc-114	21	27:14	BF
plsc-118	2	38:31	BF
econ-159	22	06:25	FB
econ-159	21	47:52	FB
econ-159	21	1:07:30	FB
econ-159	6	20:48	FB
engl-220	22	50:55	FB
engl-291	15	42:01	FB
hist-234	21	00:01	FB
phil-176	13	31:31	FB
plsc-118	15	32:25	FB
engl-220	22	33:29	FB
hist-119	5	17:17	FB
hist-234	12	01:12	FB
hist-276	8	50:22	FB
mcdb-150	8	33:47	FB
mcdb-150	4	35:32	FB
plsc-270	12	26:21	FB
psyc-123	11	45:58	FB
psyc-123	11	45:58	FB
astr-160	17	37:25	RL
astr-160	21	39:31	RL
chem-125b	14	18:41	RL
econ-252-11	7	31:37	RL
gg-140	1	23:04	RL
hist-116	4	21:02	RL
mcdb-150	23	1:13:03	RL
mcdb-150	14	01:40	RL
mcdb-150	2	38:58	RL
musi-112	11	26:43	RL
phys-201	16	05:36	RL
plsc-118	3	01:36	RL
psyc-123	4	1:06:20	RL
psyc-123	18	17:03	RL
psyc-123	7	40:59	RL
rlst-152	23	01:36	RL
rlst-152	5	09:50	RL
rlst-152	17	36:51	RL
rlst-152	22	02:23	RL
rlst-152	4	35:01	RL
engl-220	22	33:29	LR
hist-119	5	17:17	LR
hist-234	12	01:12	LR
hist-276	8	50:22	LR
mcdb-150	8	33:47	LR
mcdb-150	4	35:32	LR
musi-112	11	26:43	LR
plsc-270	12	26:21	LR
psyc-123	11	45:58	LR
psyc-123	11	45:58	LR

All time references are approximate.

Chapter 4

Brain Potentials and the Syntax-Semantics Distinction

The N400 and the P600 are two patterns of electrical brain potentials which can sometimes be found when people read or hear unexpected words. They have been widely claimed to be the neurological correlates of semantic and syntactic anomalies, respectively, but evidence accumulated over the last decade has raised some serious doubts about that interpretation. In this chapter, I first review some of this evidence and then present an alternative way to think about the issue. My suggestion is built on Shannon's concept of noisy-channel decoding by tables of typical sets, and it is thus fundamentally statistical in nature. I show that a proper application of Shannon's concepts to the reading process provides an interesting reinterpretation of our notion of "syntax," thus questioning some fundamental assumptions of linguistics.

4.1 Introduction

One of the key ideas that helped establish linguistics as a respectable science in the 1950s was the idea that the processing of syntax is carried out by an autonomous mental module that works independently of the processing of meaning. Chomsky famously pushed this point by presenting sentence pairs that seemed to dissociate violations of syntactic expectations from violations of semantic expectations (Chomsky, 1957, pp. 15–16):

(4.1) *The child seems sleeping* (syntactic anomaly)

(4.2) *Today I saw a fragile whale* (semantic anomaly)

Based on such examples, he concluded that violations of syntactic rules had nothing to do with the meaning of the words in the sentence, and thus

> . . . we are forced to conclude that grammar is autonomous and and independent of meaning, and that probabilistic models give no particular insight into some of the basic problems of syntactic structure. (Chomsky, 1957, p. 17)

While the introspective case from example sentences was Chomsky's strongest argument at the time, he did describe some hypothetical psychological experiments, such as memory and reading tests, that he imagined might corroborate his claim by distinguishing "ungrammatical" sentences from mere nonsense (Chomsky, 1957, p. 16).

This chapter is about one attempt to realize such a research program. As I shall explain shortly, the brain sciences found some new results in the 1990s which seemed to finally put Chomsky's claims on a solid scientific basis. For a period of about ten years, the consensus was accordingly that the anatomical and psychological reality of Chomsky's syntactic engine was a scientific fact. The autonomy of syntax had acquired the authority of being a matter of "brain stuff."

In one sense, this chapter is an elaborate criticism of this idea. I will review a wealth of empirical evidence from the last decade which seriously complicates the picture of the brain phenomena that were originally used to vindicate the Chomskyan view of language. Even a casual look at this literature exposes an abundance of unruly little details that fit very badly into the old pigeonholes.

But this is also a constructive effort: I propose a precise computational model that may explain much of the psychological evidence. This model is based on a seemingly insignificant idea about statistical error used by Shannon in the 1948 paper which founded the field of information theory (Shannon, 1948).

This idea has some deep connections to the foundations of statistical inference and thus to the notion of rational decision-making in the most general sense. Presenting my model will thus require a detour around some basic concepts from information theory and statistics. Once I have introduced those concepts, I will focus on a more specific example of a decoding problem similar to the experimental situations used in the psychological literature. I will then discuss how that toy problem might be scaled up to give accurate, quantitative predictions.

4.2 The N400 and the P600

In this section, I will introduce the brain phenomena that are the pivot of the entire chapter. These phenomena are a couple of distinctive changes in the electrical brain potential that can be measured off the scalp of a person's head during reading, listening, or other activities. Only in the next section will I discuss in more detail how we should interpret these modulations of the brain potential.

4.2.1 The N400

Suppose I present you with a sentence by flashing each word on a computer screen at regular intervals:

(4.3) *I . . . spread . . . the . . . warm . . . toast . . . with . . .*

While you are reading the sentence, I am recording the electrical activity at different places on the top of your head. Since all nerves in your body change the electrical field around them when they are excited, this recording provides an estimate of how active the nerves inside your head are. By using the electrical activity in the head as a measure of cognitive effort, this experimental paradigm can thus reveal whether a sentence requires more effort than usual to read, and where in the sentence the problems begin.

This means that rather specific differences between sentences can be compared. I could, for instance, change a single word in a sentence and then compare the pattern of electrical activity that arises when you read the original sentence with the pattern associated with the manipulated one:

(4.4) *I spread the warm toast with butter* (original)

(4.5) *I spread the warm toast with socks* (manipulation)

This manipulation was the idea behind an experiment which was first carried out by Kutas and Hillyard in 1980, and which has since then been replicated many times. Their experiment involved a set of sentence pairs like the ones above: One completely unproblematic sentence and one that ended in a "semantically inappropriate (but syntactically correct)" word (Kutas and Hillyard, 1980, p. 203).

By averaging over several subjects' responses to the two versions of the sentences, Kutas and Hillyard found that the oddness of the manipulated sentences tended to evoke a characteristic electrical response characterized by an excess of negative electrical potential relative to the control condition (Fig. 4.2.1).

Because this bump on the waveform showed up around 400 milliseconds after the unexpected word was presented, they called the effect the "N400." They suggested that it might reflect a kind of "second look," like a cognitive equivalent of rubbing your eyes in disbelief.

4.2.2 The P600

Kutas and Hillyard's experiment was explicitly designed to manipulate "semantic congruity" while keeping syntactic factors constant. This naturally led to the question of whether syntactic manipulations would lead to different results than semantic manipulations.

This question was answered in 1992 by Osterhout and Holcomb, who compared a different kind of sentence pairs in an otherwise virtually identical experimental paradigm:

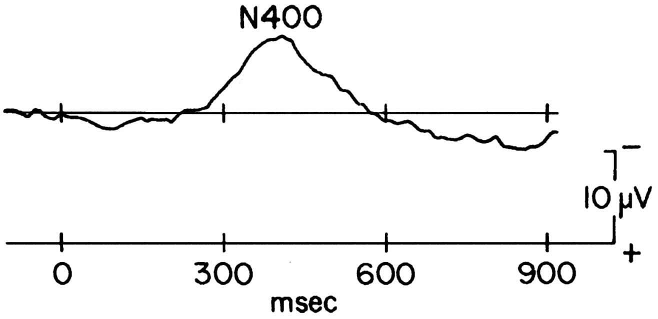


Figure 4.1: A graph from Kutas and Hillyard (1980, p. 204) showing the averaged response to the inappropriate word (e.g., *socks*) minus the averaged response to the appropriate (e.g., *butter*). Both waveforms were recorded by an electrode at the back of the head (the medial parietal position). As is conventional in electrical engineering, negative charge is plotted as up.

(4.6) *The swimmer decided to lose weight* (control)

(4.7) *The swimmer urged to lose weight* (manipulation)

Although the difference in wording between these two sentences lies in the verb (*decided* vs. *urged*), the difference in how odd they are appears only later: The sentence fragment *The swimmer urged ...* can be completed in many perfectly fine ways, but the extension *The swimmer urged to ...* cannot. It is thus only when the word *to* is presented that it gets difficult to come up with a plausible structural interpretation of the sentence.

Consistent with this reasoning, Osterhout and Holcomb found a marked difference between the reaction to the word *to* in the two conditions, with the manipulated sentence provoking an increased positive potential. They found that this positive effect peaked at about 600 milliseconds after the onset of the word and logically named it "P600."

Having thus apparently found a neurological correlate of the mismatch between verb type and complement type, Osterhout and Holcomb speculated that "the P600 and N400 effects are elicited as a function of anomaly type (syntactic and semantic, respectively)," and that "the P600 co-occurs with syntactic anomaly" (Osterhout and Holcomb, 1993, pp. 785 and 798).

Their results thus seemed to support the idea that syntax and semantics are distinct features of language — not only in the mind of the linguist, but also in the brain of the language user.

4.3 The Frayed Ends of Syntax

With the discovery of the P600, the brain sciences produced a strong neurological argument in favor of the Chomskyan view of language. Linguists could now with confidence postulate the "autonomy of syntax" and cite the authority of brain science as support (see e.g. Fernández and Cairns, 2010, ch. 3). The issues raised by Chomsky from a purely introspective perspective finally seemed to be settled by empirical investigations.

The issue is, however, not quite as simple as this. Within the last decade, a number of experiments have documented P600 responses to sentences that are not syntactically anomalous in any usual sense of the word "syntax." I will spend the remainder of this section illustrating these complications in some detail, drawing quite heavily on an excellent review paper by Kuperberg (2007). For a more exhaustive review of the wide variety of contexts in which the N400 is found, including non-linguistic contexts, see Kutas and Federmeier (2011).

4.3.1 Word Shuffling and P600 Effects

One of the first indications that the P600 could not unproblematically be equated with broken syntax came from a Dutch study which manipulated the test sentences by changing a passive construction into an active, causing the logical subject and the logical object swap roles (Hoeks et al., 2004).

In Dutch, the main content verb is often placed at the very end of the sentence, as in *The boy has the ball kicked.* If you read such a verb-final sentence in a word-for-word presentation, a whole string of nouns and auxiliary verbs can thus be presented to you before you finally reach the main content verb (here *kicked*). This means that one can construct a Dutch sentence such that it creates a very strong expectation as to what the final verb will be, and those expectations can then be violated or respected.

This grammatical fact was exploited in the materials used by Hoeks et al. They constructed three different manipulated sentences from a control by substituting either the main content verb, an auxiliary verb, or both:

(4.8) *De eed werd door de jonge artsen afgelegd.*
("The oath was taken by the young doctors.")

(4.9) *De eed werd door de jonge artsen behouden.*
("The oath was kept/retained by the young doctors.")

(4.10) *De eed heeft de jonge artsen afgelegd.*
("The oath has taken the young doctors.")

(4.11) *De eed heeft de jonge artsen behouden.*
("The oath has kept/retained the young doctors.")

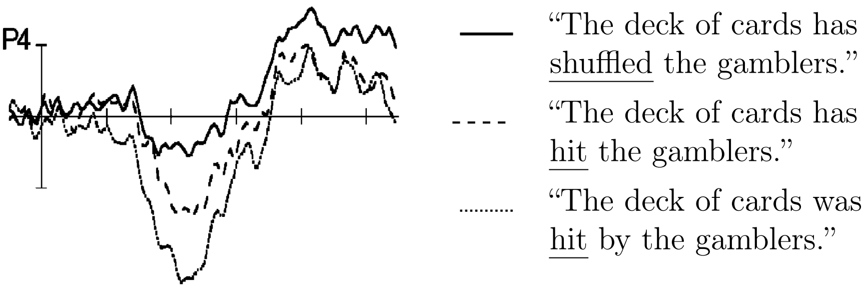


Figure 4.2: Averaged single-electrode recordings from Hoeks et al. (2004, p. 68), with positive charge plotted as up. The three waves show the differences in voltage between the three manipulated conditions and the control condition (“The deck of cards was shuffled by the gamblers”). All three waveforms exhibited a significant P600 effect at this electrode.

The manipulations of the main verb (e.g., substituting *retain an oath* for *take an oath*) tended to produce mere nonsense, and as expected, this manipulation produced an N400 effect. However, the manipulation of the auxiliary verb — which caused the subject and object to change roles — unexpectedly triggered an excess of late positive potential beginning around 600 millisecond after the final verb (cf. Fig. 4.3.1).

Although both of these manipulations resulted in a grammatical sentence, only one of them would produce an N400, and the other would instead produce a very pronounced P600. This ran directly counter to the notion that the P600 exclusively tracks syntactic problems.

Similar effects were reported for English. In one 2003 experiment, Kuperberg et al. compared two different kinds of unexpected verbs and again found that only some of them provoked an N400 response:

(4.12) *For breakfast, the boys would only eat . . .* (control)

(4.13) *For breakfast, the boys would only bury . . .* (N400)

(4.14) *For breakfast, the eggs would only eat . . .* (P600)

Kuperberg et al. suggested that this difference might be explained in terms of the thematic roles prescribed by a verb like *eat*, assuming that “the P600 is sensitive to violations in this thematic structure” (Kuperberg et al., 2003, p. 127).

This line of thought was also supported by another study which compared mere nonsense to bizarre subject-verb combinations (Kim and Osterhout, 2005). This study, too, found marked differences in the kinds of response elicited by the two manipulations relative to a control condition of ordinary English:

(4.15) *The library books had been borrowed by the graduate student.* (control)

(4.16) *The tragic mistake was borrowing in a hurry.* (N400)

(4.17) *The library books had been borrowing the graduate student.* (P600)

Like Kuperberg et al., Kim and Osterhout suspected that the difference between the two target sentences had something to do with how nouns plug into verbs. They formulated this intuition in terms of a "semantic attraction to particular predicate–argument combinations" (Kim and Osterhout, 2005, p. 215).

4.3.2 Animacy and Related Criteria

Looking at these examples, one might come to expect that the crucial issue here had something to do with animacy: *boys* are animate and therefore appropriate subjects for the verb *eat*, but *eggs* are not. However, as several authors have argued, this distinction does not quite capture the differences in the empirical material (Kuperberg et al., 2003; Kuperberg, 2007).

For instance, *library books* and *tragic mistakes* are both inanimate noun phrases, and both are inappropriate as subjects for the verb *borrow*. Still, the contruction *the library books borrowed* elicits a P600, while *the tragic mistakes borrowed* elicits an N400, as mentioned above. Since these constructions are equivalent in terms of animacy violations, this dimension alone is thus an unreliable predictor of P600 effects.

It should also be kept in mind that many earlier experiments had used subject-verb combinations that were clearly illegitimate in terms of animacy and other selection criteria, but still produced N400 effects instead of P600 effects. One German study, for instance, used the following target sentences:

(4.18) *Der Honig wurde ermordet.*
("The honey was murdered.")

(4.19) *Der Ball hat geträumt.*
("The ball has dreamt.")

Both of these sentence types provoked a marked N400 response, but no P600 (Rösler et al., 1993). It would thus appear that inanimate objects can be used with verbs that require animate subjects without provoking a P600 (although this will probably produce an N400).

As for the opposite implication — P600 effects predicting animacy violations — other research groups have also documented P600 responses to sentences in which animate subjects do clearly animate things. One study by Kolk et al. (2003), for instance, found this in Dutch sentences like

(4.20) *De vos die op de stropers joeg ...*
("The fox that was hunting the poachers ... ")

(4.21) *De kat die voor de muizen vluchtte . . .*
("The cat that was fleeing the mice . . . ")

Of course, the roles are in some sense swapped in this sentence — but since both foxes and poachers are animate, the strangeness of these sentences cannot be described in terms of animacy alone.

In another experiment applying essentially the same materials, van Herten et al. (2005, p. 249), to their own surprise, also found P600 effects for sentences such as

(4.22) *De bomen die in het park speelden . . .*
("The trees that played in the park . . . ")

In a later paper, they hypothesized that this P600 effect was caused by the strong semantic relation between the words (one can *play* in a *park* full of *trees*) in combination with the oddness of the sentence as a whole (the *trees played*). One way of describing this pattern could be that it involves nouns and verbs that tend to occur in the same contexts, but rarely in subject-predicate combinations (Van Herten et al., 2006).

This theory would explain differences such as the following:

(4.23) *Jan zag dat de olifanten de bomen snoeiden en . . .*
("John saw that the elephants pruned the trees and . . . " — P600)

(4.24) *Jan zag dat de olifanten de bomen verwenden en . . .*
("John saw that the elephants spoiled the trees and . . . " — N400)

Trees, *elephants*, and *pruning* can easily occur in the same text or even the same sentence, but it is rarely the elephants that are doing the pruning. On the other hand, *spoiling* (as in *spoiling a child*) is semantically unrelated to elephants and trees, and the word tends to occur in very different contexts.

Whether this exact formulation of the theory is correct or not, there are thus quite strong reasons to doubt that the conditions producing the P600 is a matter of plugging an inanimate noun into a verb frame that requires an animate argument, or *vice versa.* A wide variety of word relationships seem to be responsible for turning "semantic" anomalies into potentially "syntactic" ones.

4.3.3 Grammatically Ungrammatical

Given all of this data, it seems that there is no clear and straightforward relationship between the P600 and the arguments that can be plugged into a verb. The issue does not seem to be syntactic well-formedness, nor does it seem to be about animacy or thematic structure. It might still be the case, however, that some other grammatical parameter — say, aspect, mood, gender, number — is the key. But even this vague suggestion has some empirical evidence against it.

One such piece of evidence comes from a study that recycled some sentences which had already previously been shown to produce strong and reliable N400 effects. In this experiment, an elaborate context was prepended to the test sentences, and the brain potentials were recorded in the same manner as in the previous study (Nieuwland and Van Berkum, 2005).

One of stories used in this experiment read as follows, in translation from the original Dutch:

(4.25) *A tourist wanted to bring his huge suitcase onto the airplane. However, because the suitcase was so heavy, the woman behind the check-in counter decided to charge the tourist extra. In response, the tourist opened his suitcase and threw some stuff out. So now, the suitcase of the resourceful tourist weighed less than the maximum twenty kilos. Next, the woman told the suitcase that she thought he looked really trendy. . . .*

Surprisingly, this long build-up to the crucial unexpected word completely canceled the N400 effect found in the original study and instead produced a strong P600 effect. Thus, the manipulation of the context which turned the suitcase into a prop in the narrative dramatically changed the way the subjects read the nonsense sentence *the woman told the suitcase*.

But perhaps the deepest problem with the idea that the P600 can be described in grammatical terms is that it can be provoked by sentences without any syntactic anomalies at all. This was already shown in a study from 1999 (Weckerly and Kutas, 1999) which compared sentences of the following form:

(4.26) *The novelist that the movie inspired praised the director for . . .*

(4.27) *The editor that the poetry depressed recognized the publisher of the . . .*

Strictly speaking, there is nothing semantically or syntactically anomalous about movies that inspire novelists, or about poetry that depresses editors. Still, these constructions tended to produce a P600 response.

Weckerly and Kutas used these findings in order to make a point about "the use of animacy information in the processing of a very difficult syntactic structure" (Weckerly and Kutas, 1999, p. 569). Their findings may also have more specific causes, though, such as the strong, conventional expectations carried by word clusters like *novelist*, *movie*, and *inspire*. However, since their experimental materials were not reprinted in the paper, it is hard to assess how strong the case for such an explanation might be.

However, regardless of these points of interpretation, it seems clear that there are some serious problems with seeing the P600 as the exhaust pipe of a syntactical processor in the brain: There are several kinds of semantic oddness that are quite strongly related to the P600; it can be quite heavily modulated by discursive context; and it can even be elicited by grammatical sentences. The standard tools from the toolbox of theoretical linguistics thus seem to have some problems getting to grips with this phenomenon.

I take this as a sign that we need to back up a bit and take a fresh look at our assumptions. In the next section, I will consequently suggest that in order to understand the brain processes behind the N400 and the P600, we need to go all the way back to Chomsky, and then go a little further back.

4.4 The Statistics of Decoding

In the preceding section, I have presented a number of empirical phenomena that seriously challenges the traditional identification of the N400 and the P600 with semantic and syntactic anomaly, respectively. As the examples have shown, the P600 in particular crops up in a number of different circumstances that cannot be equated with broken syntax without bending our notion of syntax completely out of shape. This raises two questions.

- First, we might wonder whether there is any system at all to where the P600 is present. Indeed, after looking at the wide variety of examples in the previous section, one might get the impression that brain potentials just aren't the kind of phenomena that can be adequately predicted. It is an open question whether we can even describe the occurrence of these brain potentials in intuitive terms, and whether we can look at a sentence and guess what kind of brain response it will evoke.

- Second, assuming that there is some kind of system, the question is what formalism would be most suited to articulate it. It is tempting to pull down the grammatical vocabulary from the shelf, since we are after all talking about language; but it is not a given that a theory of the N400 and the P600 should come in such a form, or that it should involve any concepts from Chomskyan linguistics.

In this section, I will suggest that there is indeed a way of understanding the two brain components of the reading-related brain potentials, but that they have little to do with language as such. Instead, I will draw on some classical statistical insights from information theory (Shannon, 1948) and approach the activity of reading as a kind of decoding problem. The two brain responses will then show up as correlates of two particular kinds of decoding error.

This suggestion is consistent with an intuition shared by many researchers in the field, namely, that the P600 is some kind of "error-correction signal" (Gibson et al., 2013, p. 8055) or the electrical counterpart of a "monitoring process triggered by a conflict between two or more incompatible representations" (Van Herten et al., 2006, 1195; see also Frenzel et al., 2011).

Similarly, Kolk et al. (2003) note that both the N400 and the P600 are triggered by unexpected events, and they add:

> The problem with such events is that they can have two sources. They can be real, in the sense that an unexpected event has indeed occurred. On the other hand, they can also stem from a processing error. (Kolk et al., 2003, p. 31)

To paraphrase this in information-theoretic terms, strange things can either happen in the source or in the channel. I agree completely with this intuition, and the purpose of this section is to spell it out in some more mathematical detail. The model that I will present is very similar to the one discussed by Gibson et al. (2013, p. 8055), both in its general philosophy and in the mathematical detail. I will, however, be a bit more explicit in my claims about the N400 and the P600 (which they only mention in passing), in particular when it comes to identifying their exact computational-level counterparts.

I will claim that the two electrical responses correspond to two different kinds of decoding error: The P600 is present when there are too many candidate causes that match an observed effect, and the N400 is present when there are none. This hypothesis draws on an idea that Shannon used in the proof of his discrete channel coding theorem (Shannon, 1948, Th. 11), and which was later made more explicit by others (Cover, 1975; Cover and Thomas, 1991, ch. 7.7). Shannon operationalized the notion of a "match" between a cause and an effect in terms of a concept called the "typical set." This set contains, as I shall explain in more detail below, the events that are about as probable as one would expect an event to be (Shannon, 1948, Th. 3).

Before I can formulate my hypothesis about the difference between the N400 and the P600, I will first have to discuss this concept and a few other basic notions from information theory. I will then illustrate the idea behind my theory by showing how it works in a computationally simple case, and then hint at what it would look like in a more realistic model.

4.4.1 Typical Sets and Most Probable Sets

One of building blocks of information theory is the concept of a "typical set" (Shannon, 1948, Th. 3). The typical set associated with an experiment is the set of outcomes whose logarithmic probabilities are close to the expected logarithmic probability.

To put this differently, suppose we identify the experiment in question with a random variable X and that we let $p(x)$ be the probability of the event $X = x$. As suggested by Samson (1951), we can then think of the number $-\log p(x)$ as a quantification of how "surprising" the observation $X = x$ is. This "surprisal" is a number between 0 and ∞, with less probable events having a higher surprisal values (cf. Fig. 4.3).

A random variable X which can take different values with different probabilities thus always corresponds implicitly to a "surprisal" variable $-\log p(X)$

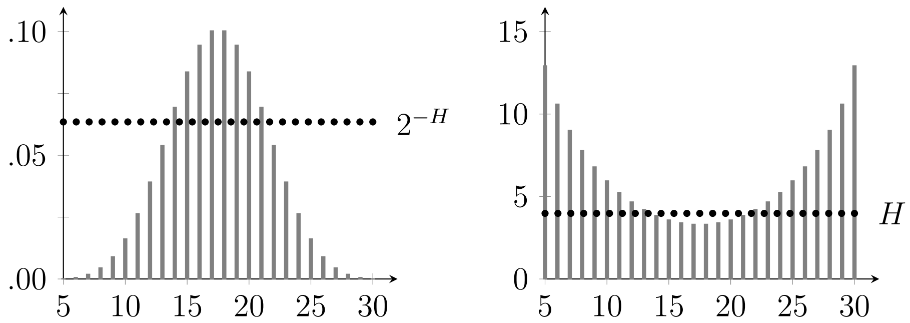


Figure 4.3: Left: The probability of rolling $\Sigma = 5, 6, 7, \ldots, 30$ with five dice, showing the typical probability 2^{-H} as a dotted line. Right: The same, plotted on an inverse logarithmic scale.

which ranges over the set of surprisal values. The expected value of this surprisal variable is $H = H(X)$, the entropy of X.

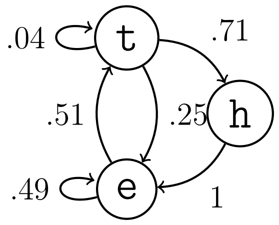
Using these concepts, we can then reformulate the definition of the typical set: It is the set of events $X = x$ whose surprisal value is close to the average surprisal. For a fixed $\varepsilon > 0$, the ε-typical set of a random variable X thus contains the x for which

$$\left| \log \frac{1}{p(x)} - H \right| \leq \varepsilon.$$

The ε-typical set associated with an experiment does not necessarily include the most probable outcome of the experiment. However, as mentioned by Shannon (1948, Th. 3) and explained in more detail by Cover and Thomas (1991, ch. 3.3), the set of outcomes that are less surprising than H usually differs very little from the typical set, since any two high-probability sets must have a high-probability overlap. Specifically, if the values of the random variable X are long sequences from an ergodic random process (such as strings of letters or words; cf. Fig. 4.4), then including the most probable outcomes will only change the size and total probability of the set slightly. In many important respects, it thus makes little difference whether we define the typical set by the bound $-\log p(x) \leq H + \varepsilon$ or the symmetric condition $H - \varepsilon \leq -\log p(x) \leq H + \varepsilon$.

4.4.2 Decoding by Jointly Typical Sets

The conditions discussed above define the typical set for a single random variable. In the context of information transmission, however, we are rarely interested in a single, isolated variable, but much more often in the relationship between two variables X and Y that model a cause and an effect, or a transmitted message and a received signal.



x	$-\frac{1}{N}\log p(x)$
`ttheeeteeeeeetheeeeeeeteth`	0.71
`ethetheetheettheeethetheth`	0.74
`etetheetheeeteeeethethethe`	0.49
`theetheethethetheeeethethe`	0.71
`thetheetheetetheteeeeethee`	0.81

Figure 4.4: Sequences of length $N = 25$ from an ergodic source. The entropy rate of the source is $H = .80$ bits per symbol.

In order to study such relationships, it is useful to consider the typical sets associated with the joint stochastic variable $X \times Y$. This set consists of pairs (x, y) that are typical with respect to the joint probability distribution on $X \times Y$, and it is called the jointly typical set.

In an information transmission context, a jointly typical pair (x, y) consists of a high-probability message x and a noisy signal y which has a high probability given the message. The set of all such pairs collectively describe the properties of the communication channel modeled by $X \times Y$. A model of written English, for instance, might include (*forty*, *fourty*) as a typical pair, since *forty* is a common word, and *fourty* a common misspelling of it. Other channels will have other characteristic types of confusability (cf. Fig. 4.5).

This way of looking at the issue suggests that a table of the typical pairs could be used for decoding: When you receive a message y, you just skim through the table, find a pair of the form (x, y), and then return x as your decoded message. Although this hypothetical algorithm is usually not feasible in its naive form, it captures many of the conceptual ideas about channel coding, and its complexity

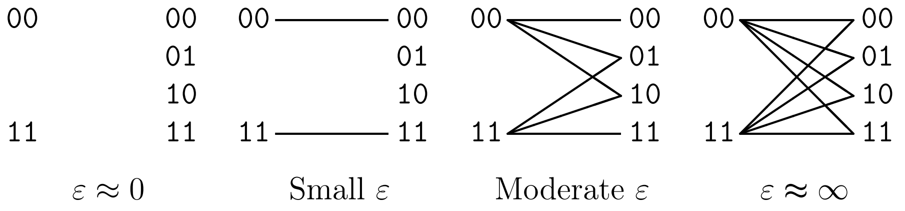


Figure 4.5: One possible structure of a jointly typical set for increasing tolerance thresholds and thus decreasing lower bounds on the joint probability $p(x, y)$.

problems are largely irrelevant to the point I want to make here.

4.4.3 Two Types of Decoding Error

As is apparent from the description of the hypothetical brute-force algorithm for decoding by jointly typical sets, two distinct problems can arise for a decoder:

1. There is no x for which (x, y) is typical
2. There is more than one x for which (x, y) is typical

As an illustration of these two types of error, consider a source which can only send two messages, $X = x_1$ and $X = x_2$, which are both points in the plane. Suppose further that the received message Y is equal to the original message plus some Gaussian noise (cf. Fig. 4.6).

Such a channel is characterized by a joint probability density over the pairs (x, y) in the plane. The typical pairs for this distribution is a union of two discs, one centered on x_1, and one centered on x_2. How large the discs are depends on the value of ε as well as the variance of the noise.

In this situation, errors of the first kind correspond to everything outside the two discs. A message y_1 found in that region is not explained very well by either of the two hypotheses $X = x_1$ and $X = x_2$, even if one of them is much more probable than the other *a posteriori*.

A message y_1 found in this region may produce a highly skewed posteriori distribution over x_1 and x_2, but it will still have an extremely low probability under either hypothesis. The shape of the posterior distribution will therefore be based on the difference between two very tiny numbers, as if we were comparing the weight of two grains of dust using a kitchen scale.

Errors of the second kind correspond to the overlap between the two discs. When a received message y_2 is located inside this region, both hypotheses have about the same posterior probability and thus explain y_2 about equally well.

This means that there is a high risk of attributing the received signal to the wrong message. Even if one of the two messages x_1 and x_2 is more probable than the other *a posteriori*, the differences are slight, and we cannot choose between them with any substantial amount of confidence.

As the picture suggests, the risk of an error of the first kind can be reduced by making ε larger, that is, by letting the discs cover more of the plane. However, this will also expand the overlap between the two discs and thus increase the probability of an error of the second kind. There is thus a trade-off between accuracy and coverage when selecting ε.

I should note that these two types of decoding error are distinct from the two kinds of selection error which are often discussed in the statistical literature (Neyman and Pearson, 1928, p. 177). Those errors obtain when a selected answer differs from the correct answer. Such selection errors are logically independent of

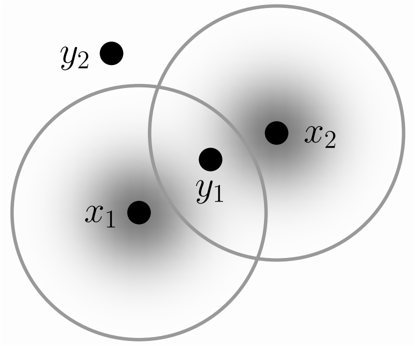


Figure 4.6: Decoding in a language with two possible input messages, x_1 and x_2. Any point in the plane represents an output message. Received signals like y_1 and y_2 produce a decoding error.

the decoding errors discussed here. However, a decoding error can be interpreted an indication that we may be about to make a selection error.

4.4.4 Degrees of Error

For a fixed ε, the space of possible signals is equipped with two regions that define the two kinds of error. One kind excludes the other. However, for our present purposes, it will often be useful to be less categorical about the extent to which a given messages leads to excess surprise or excess ambiguity. I will therefore now briefly discuss one possible way of computing a graded measure of the extent to which a specific observation constitutes a decoding error.

The concept of an "excessively improbable" observation y can be quantified in terms of its marginal probability $p(y) = E\,[\,p(y\,|\,X)p(X)\,]$. Since Y has a marginal distribution, it has a marginal entropy. We can therefore measure the excess surprisal about the observation $Y = y$ by

$$\log\frac{1}{p(y)} \;-\; E\left[\log\frac{1}{p(Y)}\right].$$

Knowing the size of this gap, perhaps in conjunction other information about the random variable $-\log p(Y)$, may help us determine whether a specific observation $Y = y$ is an outlier, or whether it is a "reasonably probable" sample from the distribution of surprisal values (Shannon, 1948, §7, p. 23). This statistic thus provides a graded measure of the first kind of decoding error.

The "excess uncertainty" caused by the observation $Y = y$ can be quantified in a similar manner. For this, we first need to measure the conditional uncertainty about the intended message x given the received signal y, or the "equivocation," as Shannon called it (Shannon, 1948, §12, p. 33). For each received message y, this posterior uncertainty is $H(X\,|\,Y = y)$. Since Y is a random variable that can take on different values, $H(X\,|\,Y = y)$ is also a random variable with a specific

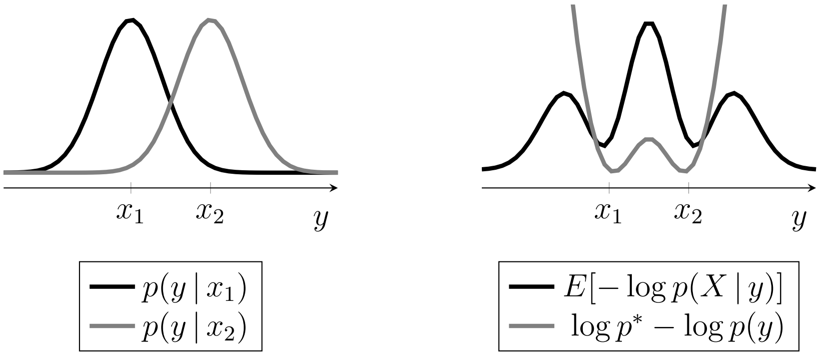


Figure 4.7: A sketch of a Gaussian channel model. Left, the conditional distributions associated with the two messages x_1 and x_2. Right, the conditional uncertainty about X given $Y = y$ plotted as a function of y, and the surprisal induced by $-\log p(y)$, relative to a baseline of $-\log p^* = -\max_y \log p(y)$. (The adjustment according to this base level is done only for visualization purposes; it has no effect on how $\log p(y)$ compares to $E\left[\log p(Y)\right]$.)

distribution. We can therefore quantify the excess entropy as

$$E\left[\log\frac{1}{p(X \mid Y = y)}\right] \; - \; E\left[\log\frac{1}{p(X \mid Y)}\right].$$

When this gap is many times larger than the standard deviation of the distribution $E\left[-\log p(X \mid Y)\right]$, the observation $Y = y$ is more likely than usual to be decoded wrongly.

To make these concepts more concrete, consider the case of a one-dimensional Gaussian channel with two possible messages (Fig. 4.7).

Under this model, the most equivocal observations are the ones right between the two messages, and these observations are also relatively unsurprising. Observations that almost coincide with the two messages are by contrast neither surprising nor equivocal.

The equivocation eventually drops to almost 0 as y reaches very extreme values, because one of the two hypotheses then becomes exponentially more probable than the other. At these extreme values, however, the marginal probability is very low, so the marginal surprisal diverges to infinity at the speed of a second-degree polynomial in y.

I shall later in this chapter provide more concrete examples that show how one can apply considerations about these distributions of $-\log p(Y)$ and $H(X \mid Y)$ to specific samples. In the following section, I will consider a single example of a channel model in order to spell out in detail what decoding might look like in a linguistic context.

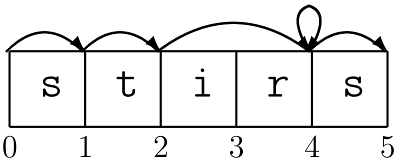


Figure 4.8: Consuming the word *stirs* by walking through positions 0, 1, 2, 4, 4, and 5.

4.5 An Application to Locally Noisy Strings

While the previous section considered some rather abstract statistical concepts, I would now like to turn to a more concrete example which is highly relevant to the general topic of this chapter. Nominally, the example is a statistical model of misspellings, but it applies to any kind of communication channel that can distort a sequence through local changes, such as reversing the order of two adjacent symbols.

Such models have been studied extensively in computer science (Levenshtein, 1966; Damerau, 1964), and many textbooks contain discussions of related ideas. The theory that underlies such models has roots in classical ideas from computer science, most notably, dynamic programming (Bellman, 1952; Viterbi, 1967).

4.5.1 A Generative Story for Misspellings

There are many complicated reasons why people misspell words, but in order to get a better mathematical grip on the situation, I will present a very schematic model of the process here. The model takes the form of a cartoonish generative story which depicts people as a certain kind of machines that experience random glitches during the writing process. These glitches cause them to incorrectly externalize the contents of their internal memory, and someone reading their output thus needs to reconstruct their original intentions based on the corrupted output signal.

More specifically, let us assume that a writer chooses a word x with a probability close to the frequency of the word in written English. Next, the writer consumes the word, letter for letter, in a left-to-right manner (cf. Fig. 4.8). While traversing the input string in this way, the writer also writes an output string y which more or less faithfully reflects the input. How exactly the writer traverses the word x and how he or she chooses which letters to write is decided stochastically as summarized in Table 4.1. Note that this channel model explicitly includes local permutations such as *percieve* instead of *perceive* (here modeled as an operation called "Reverse").

As the table shows, the model contains several hyperparameters that be can set to arbitrary values, depending on what we think best reflects actual behavior. In the examples discussed below, I have used the values $\alpha = .96$ and $\beta = \gamma =$

$\delta = \eta = .01$, meaning that an average of 24 out of 25 letters are expected to be faithfully reproduced.

State	Action	Probability	Output	Next state
Start	Choose x	$\Pr(x)$	–	0
$i < \lvert x \rvert$	Echo	α	x_i	$i+1$
$i < \lvert x \rvert$	Change	$\beta/25$	$c \neq x_i$	$i+1$
$i < \lvert x \rvert$	Insert	$\gamma/26$	c	i
$i < \lvert x \rvert$	Delete	δ	–	$i+1$
$i < \lvert x \rvert$	Reverse	η	$x_{i+1}x_i$	$i+2$
$i = \lvert x \rvert$	Echo	α	a_i	$i+1$
$i = \lvert x \rvert$	Change	$\beta/25Z$	$c \neq x_i$	$i+1$
$i = \lvert x \rvert$	Insert	$\gamma/26Z$	c	i
$i = \lvert x \rvert$	Delete	δ	–	$i+1$
$i = \lvert x \rvert + 1$	Insert	$\gamma/26$	c	i
$i = \lvert x \rvert + 1$	Halt	$1-\gamma$	–	–

Table 4.1: Generative model for a spelling channel. The rows labeled "Change" and "Insert" should be read as abbreviated forms of several distinct entries, one per possible output letter. Z is the normalizing constant $(\beta+\gamma)/(\beta+\gamma+\eta)$.

Thanks to classic computer science trick, the transmission likelihoods $p(y \mid x)$ for this channel can be computed quite easily. The way a writer consumes a string x and outputs a string y can be identified with a path through a matrix. Each step of the path corresponds to a particular editing operation like deletion or reversal (cf. Fig. 4.9). The problem of finding the likelihood $p(y \mid x)$ is thus equivalent to the problem of summing up the probabilities of all paths through this matrix. The probability of any specific path can be found by multiplying the probability of the individual editing operations, since these operations independent.

The crucial fact that allows us to sum up the probability of these paths is that the transmission errors only have local effects. This means that the only matrices for which we need to count all the possible paths are the ones of size 2×2. For all larger sizes (e.g., 3×4), we can ease the computation by using the partial results we have already computed (e.g., the sums for the matrices of size 3×3). Thus, by working upwards from size 2×2 and keeping careful score of our partial results, we can add up the probability of every path through the matrix without

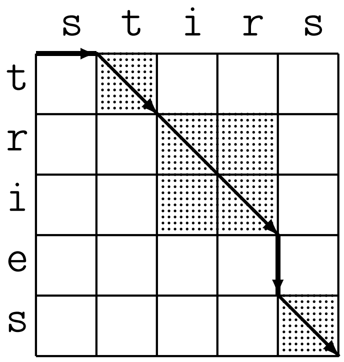


Figure 4.9: Misspelling *stirs* as *tries* due to a deletion, a reversal, and a spurious insertion.

actually looking at each one individually.

This style of a bottom-up recursion is known as dynamic programming (Bellman, 1952). I will tacitly use it in all the examples discussed below.

4.5.2 Estimating Joint Surprisal Values

Analyzing the joint entropy of the channel $X \times Y$ is not completely trivial for the model used here, and the details are not terribly important for my present purposes. It will, however, be very illustrative to see a few representative examples, so I will now provide a rough back-of-the-envelope calculation of some descriptive statistics associated with the spelling channel.

Let us first assume that an input word x is given and then have a look at what our uncertainty about the output string y is. We can think roughly of the consumption of a single input letter as a process requiring two choices to be made: Deciding whether to make a mistake, and if so, deciding which mistake to make. The first choice involves two options with probabilities 0.96 and 0.04, respectively, so it contains

$$0.96 \log \frac{1}{0.96} \; + \; 0.04 \log \frac{1}{0.04} \; = \; 0.17 \text{ bits of uncertainty.}$$

The second choice is a random selection from a set of 25 letters, so it amounts to $\log_2 25 = 4.64$ bits of uncertainty. However, since this choice only comes up in the four percent of the cases, the grand total is

$$H(Y \,|\, X) \; = \; 0.17 + 0.04 \cdot 4.64 \; = \; 0.36 \text{ bits of uncertainty per letter.}$$

An input word with N letters is thus associated with about $2^{0.36N}$ typical output strings, according to this simplified calculation.

This accounts for the channel entropy $H(Y \,|\, X)$. The source entropy $H(X)$, on the other hand, only involves choosing a word from the dictionary. Using the frequencies in the Brown corpus (Francis and Kucera, 1967) as estimates of the

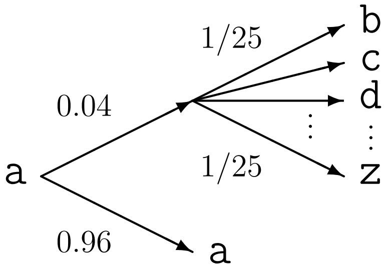


Figure 4.10: The simplified channel model in which the uncertainty about the output is reduced to two consecutive choices, first whether to make an error, and then which error to make.

word probabilities, we arrive at an entropy of about $H(X) = 10.54$ bits. Words like *know*, *while*, *last*, *us*, *might*, *great*, and *old* have surprisal values close to this average. The most frequent word, *the*, has a surprisal value of 3.84.

To show more concretely what these numbers mean, consider the word *great*. This word consists of five letters, and as we have seen, each letter contributes about 0.36 bits of surprisal. We can thus expect an average conditional surprisal about the output, $H(Y \mid X = \textit{great})$, in the ballpark of

$$5 \cdot 0.36 \;=\; 1.80 \text{ bits.}$$

In order to find the prior surprisal associated with the input, we can look up the probability of the word *great*. This turns out to be about $6.6 \cdot 10^{-4}$, corresponding to a surprisal value of 10.56 bits. Adding this prior surprisal to the average conditional surprisal, we find an average joint surprisal for pairs of the form (*great*, y). On average, this should be about

$$10.56 + 1.80 \;=\; 12.36 \text{ bits.}$$

x	y	$-\log p(x,y)$
great	*great*	10.84
great	*graet*	17.32
great	*grate*	24.00
great	*grxqz*	30.42

Table 4.2: Examples of joint surprisals.

In fact, we can pick apart this average even further: Suppose for instance that the transmission does not introduce any errors, so that the input-output pair is $(x, y) = (\textit{great}, \textit{great})$. Then the joint surprisal is

$$10.56 - \log 0.96^5 \;=\; 10.85 \text{ bits,}$$

or in fact slightly less, since the output *great* can either be produced by a faithful reproduction of x (with probability 0.96^5), or an unlikely combination of errors that cancel each other out. If a single error is introduced, on the other hand, the joint surprisal for the pair (x, y) will be about

$$10.56 - \log 0.96^4 - \log 0.01 \;=\; 17.44 \text{ bits,}$$

depending a bit on what the error is. Table 4.2 gives some direct computations that corroborate these estimates.

4.5.3 Cumulative Hypothesis Revision

All of the preceding discussion assume that we selected an entire word from our dictionary and then transmitted it through the spelling channel in a single transmission. How does the situation change if we suppose that the letters of the word are transmitted one by one, and that the receiver is allowed to iteratively revise his or her predictions in the light of the incoming data?

The probability of observing two events, $p(x_1, x_2)$, is the same as the probability of observing the first event and then the second, $p(x_1)\, p(x_2 \,|\, x_1)$. Since the logarithm turns products into sums, this means that if you observe a string of events, the sum of your individual surprisals equals the bulk surprisal you would experience from observing the whole sequence at once. For instance,

$$\log \frac{1}{p(x_1, x_2, x_3)} \;=\; \log \frac{1}{p(x_3 \,|\, x_1, x_2)} + \log \frac{1}{p(x_2 \,|\, x_1)} + \log \frac{1}{p(x_1)}.$$

As an illustration of what this means in the context of the spelling channel, suppose that you have forecasted, rather arbitrarily, that you are about to receive the message $x =$ *fall*. In fact, however, the actual output that you will see is $y =$ *flat*, so at some point along the way, your expectations will be violated. As you add up the surprisal values associated with each individual letter, you will gradually accumulate a total of 25.05 bits, the surprisal of the pair (*fall*, *flat*).

The surprisal need not be evenly distributed, though. When you have only seen the letter *f*, for instance, the hypothesis $x =$ *fall* is still perfectly consistent with the data, and your surprisal is moderate. It is only when the next letter turns out to be an *l* that your surprisal jumps upwards. More examples are given Table 4.3.

As can be seen from the last row of the table, the hypothesis *flat* terminates at a moderate 14.07 bits, which is not too far from the roughly 12 bits that would expected for a word of this length. For an ε between roughly 2 and 10, this hypothesis would thus be accepted, and the others rejected.

Notice also that the table contains occasional cases of reanalysis. In particular, when the *a* in *flat* is revealed, the hypothesis $x =$ *fall* hardly changes surprisal level. This is because the segment *fla* is consistent with a reversal hypothesis under which the *l* and the *a* were written in the reverse order. Observing the *a* thus in a sense explains away the unexpected *l* that would otherwise have to be explained as a spurious insertion.

x	–	*f*	*fl*	*fla*	*flat*
for	6.73	6.77	13.38	19.03	25.05
large	11.44	18.04	18.11	18.17	24.78
fall	12.73	12.77	19.39	19.39	25.05
flat	13.85	13.89	13.95	14.01	14.07

Table 4.3: Cumulative letter-for-letter surprisals at the string *flat* from the perspective of various hypotheses.

4.5.4 Decoding Error in the Spelling Channel

These observations bring me back to the main topic of this section, the two types of decoding error in a noisy channel. As explained in section 4.4.3, decoding by typical sets can either break when you have too many or too few candidates that can explain a received signal.

In the context of the spelling channel, an error of type I behaves largely as we would expect it to: Unrepairable nonsense words like *srxzk* do not look like any known words, and any hypothesis about the underlying message will thus keep accumulating surprisal as more letters are revealed (cf. Fig. 4.11). The most probable hypothesis for the string *srxzk*, the word *size*, eventually accumulates 31.78 bits of surprisal, well above the level expected for a word of this length. The string *srxzk* could thus lead to a decoding error of type I, since a decoder might simply give up trying to reconstruct the original word given this unexpectedly high surprisal score.

This situation should be compared to that of the non-word *flide* (cf. Fig. 4.12). As can be seen from the graph, there is a whole clutter of hypotheses which are roughly equally far away from this word. The hypothesis *slide* is the best candidate, but the difference up to the second best, *flies*, is only 0.70 bits. For many choices of ε, the two hypotheses *slide* and *flies* would thus be simultaneously rejected or simultaneously accepted. Both of those decisions would result in a decoding error.

From the present perspective, it is also interesting that the graph for the various hypotheses cross each other so frequently: After the fragment *fli* is presented, the two hypotheses *slide* and *fled* are practically equiprobable. After the fragment *flid* is presented, the hypothesis *flies* is temporarily more probable than the eventual winner, *slide.* Again, this flip-flopping between roughly equiprobable hypotheses means that a decoder with a not too cautious choice of ε would be in a high danger of committing an error of type II.

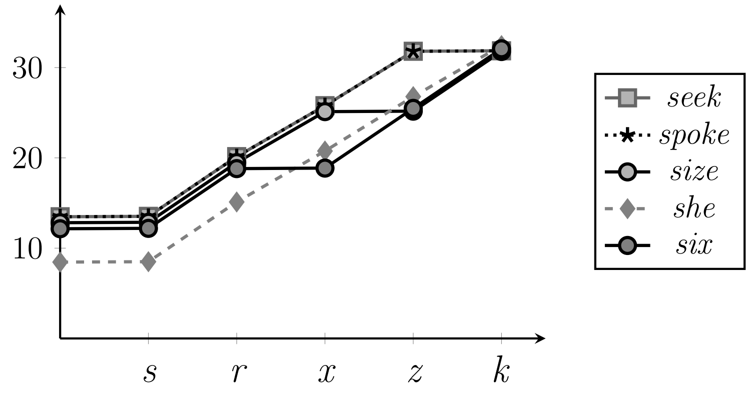


Figure 4.11: Letter-for-letter surprisal values for various decodings of *srxzk*.

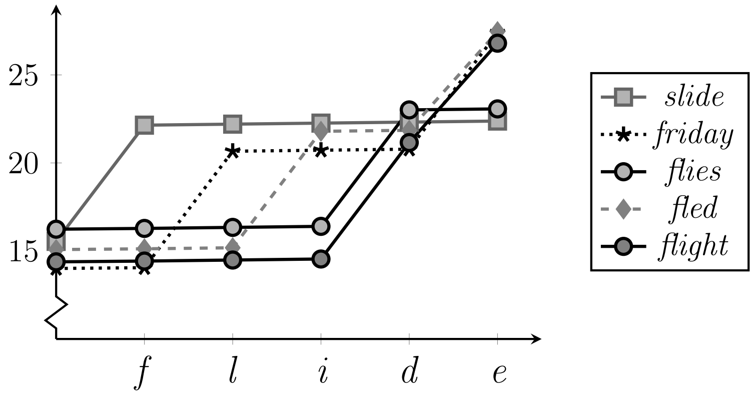


Figure 4.12: Letter-for-letter surprisal values for various decodings of *flide*.

This oscillation between competing hypotheses also means that the online decoding comes with its own "garden path" phenomenon: Hypotheses with high prior probability attract early attention which may in retrospect turn out to be unwarranted. This is, of course, particularly true if the output word contains errors or is otherwise abnormal so that ordinary expectations do not apply. Regardless of the cause, however, it should be emphasized that this phenomenon occurs here in a situation that does not involve any kind of "syntax" except the possibility of swapping two neighboring letters.

We have thus seen two kinds of statistical error related to two kinds of manipulated strings. Figuratively speaking, one abruptly pushes the receiver away, while the other keeps pushing the receiver back and forth between a portfolio of hypotheses that all seem to fit somewhat, but not quite. The first is associated with errors of type I, and the other with errors of type II.

x	$p(x)$	x	$p(x)$
hunter shoots lion	.4444	*lion shoots lion*	.0002
hunter fears lion	.2189	*lion fears lion*	.0098
hunter shoots hunter	.0045	*lion shoots hunter*	.0002
hunter fears hunter	.0022	*lion fears hunter*	.3198

Table 4.4: A toy language with eight sentences.

4.6 Linguistic Plausibility

In the previous section, I discussed what the statistics of decoding looks like in the case of a sample spelling channel that maps strings of letters to strings of letters. This example was chosen with an analogy between strings of letters and strings of words in mind, not because I believe that words are read one letter at a time. In this section, I want to discuss extent to which we might apply similar techniques to the problem of decoding sequences of words.

This question can be approaches from two sides: We can either adapt our computational models to the existing experiments, or we can adapt our experiments to the exiting computational models. In this section, I consider both of these options in turn.

4.6.1 A Finite Language

As a first example, consider the toy language shown in Table 4.4. This language consists of eight different sentences, whose prior probabilities I selected so as to approximately match how the actual frequencies of these phrases or others expressing roughly similar ideas. However, the exact numbers should of course not be taken all too seriously.

This language defines a stochastic source, X. We can feed this source into a channel $p(Y \mid X)$ which can make two kinds of errors:

- Each sentence has a probability .005 of being scrambled by a randomly chosen permutation (from the set of six permutations).
- Each word in the sentence has a probability of .005 of being replaced with a randomly chosen alternative (from the lexicon of four words).

This channel model defines the conditional probabilities $p(Y \mid X)$. In combination with the prior probabilities $p(X)$, this gives rise to a distribution over the 64

y	$p(y)$	$-\log p(y)$	z	$H(X \mid y)$	z
hunter shoots lion	.4370	1.19	−0.48	0.03	−0.09
hunter fears lion	.2173	2.20	0.09	0.10	0.04
lion shoots hunter	.0022	8.85	3.88	1.17	1.91
lion shoots lion	.0025	8.67	3.79	0.60	0.91
lion fears lion	.0123	6.35	2.46	0.98	1.58

Table 4.5: Examples of output surprisals and posterior entropies.

possible output sequences y. We have, for instance, output probabilities like

$$p(Y = \textit{lion lion lion}) = .0002.$$

Since Y thus has a specific, known distribution, so does the corresponding surprisal variable $-\log p(Y)$. The same holds for the posterior entropy variable $H(X \mid Y)$ considered as a statistic of Y. For a specific observation $Y = y$, these variables have the values $-\log p(Y = y)$ and $H(X \mid Y = y)$. Some examples are shown in Table 4.5.

Averaging over the entire set of 64 possible sequences of words, we can find the means and standard deviations of these two random variables:

$$\begin{aligned} E\left[-\log p(Y)\right] &= 2.03 \pm 1.75 \\ E\left[H(X \mid Y)\right] &= 0.08 \pm 0.57 \end{aligned}$$

The second number is here the usual unbiased estimate of the standard deviation. Using the corresponding means and standard deviations, we can map any value v to a score $(v - \mu)/\sigma$ which indicates how far it is from the mean, μ, measured in terms of standard deviations, σ. These z-scores are also provided in Table 4.5.

The channel model described here is of course not a realistic model of ordinary communication in spoken or written English. However, let us imagine that we were psychologists of a certain type of human being that did in fact use this eight-sentence language and then consider the predictions that we could potentially derive from this table.

The first thing to notice in this respect is that sentences like *lion shoots hunter* and *lion shoots lion* have surprisal values that are several standard deviations above their means. These sentences are thus not only improbable but in a certain sense more improbable than they are supposed to be. They are the type of observations that would provoke an N400 response.

The second thing to notice is that the sentence *lion shoots hunter* has a quite high posterior entropy, close to two standard deviations above the expected value. This is due to the fact that it can plausibly be explained by several different hypotheses, including

$$\begin{aligned} p(X = \textit{lion fears hunter} \mid Y = \textit{lion shoots hunter}) &= .7255 \\ p(X = \textit{hunter shoots lion} \mid Y = \textit{lion shoots hunter}) &= .1730 \\ p(X = \textit{lion shoots hunter} \mid Y = \textit{lion shoots hunter}) &= .0903 \end{aligned}$$

Based on these figures, we would expect the sentence *lion shoots hunter* to elicit a P600 response in our hypothetical eight-sentence English-speaker. Since it is also quite improbable, it might elicit an N400 too. Indeed, with the parameter settings used here, every output string that produces a high posterior entropy also has an above-average surprisal value, although this does not hold generally.

4.6.2 Short- and Long-Distance Prediction

In the previous example, I simply postulated a language and assigned its sentences probabilities in a somewhat plausible but rather arbitrary manner. I now want to consider a different approach that can accommodate more realistic statistics but still handle long-distance dependencies in the sentence.

Short-distance dependencies are most naturally and effectively modeled by a Markov model such as a trigram model. Such models consist of a table of continuation probabilities that provide a probability distribution over the next word in a sentence given the two last words, as in

$$p(\textit{time} \mid \textit{is still}) = .0035.$$

These models capture perfectly the short-distance relations in a sentence but are incapable of representing the constraints that a responsible for the predictability of sentences like

(4.28) *I have a brother but I don't have any* ________.

One crude but simple way of capturing such dependencies is to simply let all the preceding words in the sentence act as independent and unordered cues and thus inform how the blank is filled in. For instance, a predictor given the sentence above would know that the sentence up to the blank had included the words

(4.29) *a*, *any*, *brother*, *but*, *don't*, *have*, *have*, *I*, *I*.

However, no information about their order of appearance would be given.

In order to implement these two models, I collected the relevant statistics from the Brown corpus. I then constructed a language model that would make predictions about the next word in a sentence by combining the predictive distributions returned by the two models with equal weight, $p = \frac{1}{2}p_T + \frac{1}{2}p_C$. This language model defines a stochastic source X which can be fed into an information channel.

For the purposes of this subsection, the most interesting kind of error that such a channel can introduce is a confusion of two nonadjecent words, as in *eat your wine and drink your meal.* Table 4.6 contains a series of examples of possibly corrupted sentences and potential candidates for the input sentence that produced them. The per-letter surprisal value of each sentence is given in the second column.

The table is divided up into pairs of lines. In each pair, the first line shows an artificially constructed sentence specifically designed so as to share important properties with examples discussed previously in this chapter. The second sentence in each pair is a decoding: It is the most probable string of words that can be obtained by performing one transposition of two (possibly nonadjecent) words in the sentence. The second sentence was thus not selected by hand, but located automatically using the language model.

As the table shows, even this simple language model captures several interesting types of linguistic knowledge and provides a correct decoding. It includes, for instance, the seemingly deep semantic knowledge required to guess that it is *the prisoner* who is a fugitive from *the police* and not the other way around. It does so without actually having any semantic representation or even a notion of hierarchical syntax.

Not all of these examples are equally convincing though. The output sentence Y = *The mouse was chasing the cat* is in spite of its simplicity not decoded correctly. A more careful analysis of the corpus statistics that went into making that choice could perhaps reveal why this example failed, but I do not currently have a good explanation.

I should mention again that these decodings are based only on the source model, that is, on the prior probabilities of the input sentences. Depending on the probability of introducing errors, not all of these candidates may in fact be interpreted as corrupted by the decoder. For instance, the two sentences Y = *I'm afraid so* and X = *I'm so afraid* have almost identical surprisal levels, and even at a relatively high level of noise, the contrast between them would not be large enough for the decoder to favor the latter.

4.6.3 Experimental Evidence

The theory that I have presented in this chapter has been formulated as a generalization about certain patterns that I noticed in the existing experimental record. To properly test this theory, we would have to test it against experimental materials explicitly designed to distinguish it from other accounts, such the conventional grammatical theories.

For the specific computational formulation of the theory that I have presented here, such experimental results do not yet exist. However, a unpublished study

s	$-\frac{1}{\lvert s\rvert}\log p(s)$
The police was trying to escape the prisoner.	19.00
The prisoner was trying to escape the police.	18.44
I'm afraid so.	19.71
I'm so afraid.	19.71
The patient looked at his doctor.	18.84
The doctor looked at his patient.	18.68
The doctor looked at his patient.	18.68
The at looked doctor his patient.	17.90
The house was parked in front of the car.	10.84
The car was parked in front of the house.	10.73
The water went straight through the boat.	18.14
The boat went straight through the water.	17.87
He walked straight to the room of the middle.	11.09
He walked straight to the middle of the room.	10.45
I heard a flash and saw thunder.	19.17
I saw a flash and heard thunder.	18.97
The hat put on his man.	18.15
The man put on his hat.	11.38
The mouse was chasing the cat.	20.36
The the was chasing mouse cat.	19.20

Table 4.6: Potential decodings of an output string. Each pair of rows shows one possibly corrupted output sentence and the most *a priori* probable input sentence in the neighborhood of that sentence. The surprisal values are based on the prior probabilities, since there is no quantitative channel model in this example.

that I only became aware of after finishing the first version of this chapter did in fact test the notion that the P600 can be interpreted as an "error-monitoring process" (Stearns, 2012, p. 5), a theory that obviously has a large overlap with what I have been proposing here.

In this study, participants read sentences of the following type (p. 25):

(4.30) *The dentist congratulated me on not having a single cavalry.*

(4.31) *The monk sat quietly and fell into a deep medication.*

(4.32) *The teacher interrupted his sentence to make a grammatical collection.*

These sentences are anomalous. However, they are anomalous in a specific way that can be corrected with a moderate mental effort, substituting *cavity*, *meditation*, and *correction* for for *cavalry*, *medication*, and *collection*.

Given the right channel parameters, such a highly attractive alternative decoding may trigger a conflict; on the one hand, the subjects have a potential decoding with high likelihood but low prior probability (the original sentence), but on the other hand, they have a potential decoding with low likelihood but high prior probability (the corrected sentence). We should therefore expect these sentences to produce a P600 effect in addition to, perhaps, and N400 effect.

Compared to the control sentences, the waveforms resulting from the anomalous sentences in fact showed both an N400 and a P600 effect (Stearns, 2012, pp. 12–14). This result can be read slightly differently depending on whether we explain the P600 as an effect of error-correction, error-handling, conflict resolution, suppression, or any of the other mental processes that might be involved in the handling of sentences with multiple plausible decodings. However, under all of these interpretations, we are bound to read the "syntactic" problem with sentences like *the swimmer urged to lose weight* in a different light.

4.7 Physiological Plausibility

Throughout this chapter, I have approached the problem of explaining the N400 and the P600 from a functional perspective; that is, I have tried to formulate a hypothesis about when they occur without providing any explanation for why they might occur.

In this section, I will briefly review some physiological aspects of the situation and speculate about some of the possible bodily processes that might underlie these two electrical responses.

4.7.1 The Physiology of the N400

It is not currently known with certainty which brain structures generate the N400 component, and there are a number of practical obstacles to combining EEG readings with brain imaging data. However, two regions consistently show up whenever brain imagining studies use materials known to elicit N400 effects. One is the left-hand side of the inferior frontal cortex, just behind the left eyebrow, and another is a frontal part of the temporal lobe, behind the left ear (Friederici et al., 2003; Maess et al., 2006). These regions are possible generator sites for the N400.

These findings are consistent with the conventional picture of the anterior temporal lobe as the site of semantic memory, remembering what words mean. A number of researchers have therefore suggested that the N400 can be explained on a physiological level as the product of a conflict between low-level processing in the temporal region and slightly higher-level processing in the left inferior frontal region (Lau et al., 2008; Baggio and Hagoort, 2011).

According to this story, the normal course of events is that we look up the individual words in a sentence in a mental dictionary located in the left anterior temporal lobe, and we then use these dictionary entries to create some kind of context representation, which is held in the inferior frontal lobe. This context representation can then communicate with the dictionary and inform expectations about what the next word might be. The N400 can then be attributed to a conflict between the contents of these two registers, either in the form of a mismatch between a word and an "integrated semantic representation" (Lau et al., 2008, p. 923), or between a sentence-level meaning representation and a dictionary entry (Baggio and Hagoort, 2011, p. 18).

These theories capture important aspects of the process, and they are probably right in explaining the N400 effect as the trace of a dissonance between upstream and downstream processes. On the other hand, however, they may also have taken their own semantic representations a bit too seriously: The expectations formed during reading or listening may be both more finely textured and more coarsely structured than conventional semantic representations suggest.

For instance, word pairs like *rain–snow* produce a significantly less negative waveform than word pairs like *snow–tulip* (Bentin et al., 1985). This can hardly be ascribed to the construction of some intricate representation of "sentence meaning," although it clearly has something to do with expectation and prediction. Similarly, you will exhibit an N400 if you watch me shaving with a rolling pin or ironing a loaf of bread (Sitnikova et al., 2008). This response is probably explained much better in terms of a general capacity for prediction (Amoruso et al., 2013) than in terms of, say, the problems of unifying two dictionary entries with inconsistent features like +ANIMATE and −ANIMATE. This too suggests that we should perhaps look for a more general scheme of explanation.

One possibility for such a scheme is the hierarchical information cascade described by Friston (2005). According to his story, information transitions from upstream to downstream layers comes in the form of corrections of top-down predictions: "Forward connections simply provide feedback by conveying prediction error to higher levels" (Friston, 2005, p. 825).

Within such a model, the presence of an N400 effect would indicate that the upstream layer was sending more corrections than usual to a downstream layer. The position of these layers in the hierarchy would determining the timing of the negative wave.

4.7.2 The Physiology of the P600

Already in the 1960s, Sutton et al. (1965) found that when their subjects were uncertain whether they would hear a sound or see a flash or light, the corresponding EEG reading showed a more positive-going wave than when they knew in advance. However, unlike the P600 effect I have discussed in this chapter, the effect they observed peaked already around 300 milliseconds after the onset of the stimulus.

However, in a recent paper, Sassenhagen et al. (2014) have argued that this "P3" effect in fact is closely related to the P600. They substantiated this claim by showing that the peak of the P600, unlike the N400 but like the P3, falls earlier or later depending on how much time a subject takes to make a decision (in the case of their experiment, a grammatical judgment).

If this conclusion is sound, then it provides a strong argument for seeing the P600 as a response to ambiguity and uncertainty, especially situations that force us to choose one of two options when accuracy is a concern.

It would also allow us to draw on the physiological explanation of the P3 proposed by Nieuwenhuis et al. (2005). Based on a number of different sources of evidence, they suggested that this ERP component is generated ultimately by the locus coeruleus, a small structure in the brain stem which is responsible for producing much of the norandrenaline consumed by the brain. This connection between the P3, the P600, and the norandrenaline system is also consistent with the notion that norandrenaline "signals *unexpected* uncertainty, as when unsignaled context switches produce strongly unexpected observations" Angela and Dayan (2005, p. 661; emphasis in original).

One possible hypothesis we might formulate about the P600 is thus that it is caused by a domain-general process of decision-making governed by the norandrenaline system of the locus coeruleus. The fact that occurs later than the P3 would then have to be explained by the fact that the experimental materials traditionally used to elicit a P3 (such as noises and flashes of light) are processed faster than meaningful words.

This is a speculative account, but it does provide a potential physiological basis for the statistical explanation I have spelled out in this chapter.

4.8 Conclusion

With the discovery of the P600, it seemed for a while as if the brain sciences had found a physiological counterpart of the distinction between semantics and syntax. Over the course of the last decade, this interpretation of the early results has been considerably complicated by further experiments which have produced P600 effects in a large variety of conditions, many of which are very difficult to explain in terms of syntax as such.

These discoveries have led a number of researchers to stipulate "parallel stream" explanations of the data: For instance, Bornkessel-Schlesewsky and Schlesewsky (2008, p. 67) propose that the P600 occurs when cues from "animacy, voice and other language-specific information types" clash with the "plausibility" of the assignment of logical roles that these cues suggest. Kuperberg (2007, p. 44) argues that it is due to conflicts between "a semantic memory-based analysis and possibly a semantically-driven combinatorial thematic analysis." Kolk et al. (2003, pp. 29–32) suggest that the cause is a conflict between an interpretation supported by a "conceptual bias" and another one supported by a "syntactic bias".

These hypotheses are all very interesting and have a lot of empirical merit. However, I think they are formulated in too linguistic terms and on a too algorithmic level, and hence that they miss the larger conceptual picture. The working philosophy I have been applying in this chapter is that we should take a principled perspective on the inference task our subjects are tyring to solve (Anderson, 1991). Instead of immediately committing to a decoding algorithm in all its gory detail, we should choose a model of the communication situation and analyze its statistical properties before we postulate any specific heuristics or neurological implementations. This way, we can let the machinery of statistics decide what counts as a "cue," a "surprise," or a "conflict" instead of inventing a new version of those concepts for every new experiment. This approach will promote models that are more modular, more transparent in their assumptions, and easier to simulate on individual words or sentences.

To illustrate what this approach entails, this chapter has presented a very conventional noisy-channel model of spelling mistakes and looked at some of its properties. The decoding problem for this channel was simple enough to be solved by exhaustive enumeration, and this liberated me to focus on the statistical aspects of the situations. As the example showed, the exhaustive-search decoder for this simple model could run into two distinct kinds of statistical error: Words like *srxkz* produced an excess of surprisal and had no plausible reconstructions, while words like *flide* had several competing reconstructions, all of which had a surprisal level within the expected range. The decoding process could thus derail either because there were too few plausible decoding hypotheses, or because there were too many.

My claim in this chapter is that these two kinds of decoding error — too few and too many plausible decodings — map onto the N400 and P600, respectively.

The P600 is thus not produced by a conflict between two "streams" that handle specific types of information (such as animacy vs. word order). Rather, it is caused by more general issues that arise whenever we try to make sense of ambiguous stimuli in an uncertain world. And while it is tempting to describe these issues in a highly task-specific vocabulary, I would argue that the process of decoding noisy signals provides a fruitful framework for understanding N400 and P600 effects.

More broadly, this statistical perspective on reading and listening could have some rather profound consequences for linguistic theory. One of the foundational assumptions of the Chomskyan program for linguistics is that sentences can be "ungrammatical," a property that can allegedly be separated cleanly from the property of being nonsensical. From the statistical perspective, by contrast, the notion of "making sense" would not be assigned to a dark corner, but rather be the centerpiece of the whole theory. Noisy-channel coding is, at its core, a formalization of the process of guesswork and sense-making.

These speculations are, however, part of a larger discussion about the foundations of linguistics that cannot be settled by anything in this chapter. I have proposed a specific explanation of the N400 and P600 effects which made no reference to grammatical and language-specific concepts, but instead explained these phenomena in statistical terms. This defuses one of the arguments in favor of linguistic nativism and related philosophical ideas, but it is still an open question how far the style of analysis can be pushed.

Chapter 5

A Quantitative Measure of Relevance Based on Kelly Gambling Theory

This chapter proposes a quantitative measure relevance which can quantify the difference between useful and useless facts. This measure evaluates sources of information according to how they affect the expected logarithmic utility of an agent. A number of reasons are given why this is often preferable to a naive value-of-information approach, and some properties and interpretations of the concept are presented, including a result about the relation between relevant information and Shannon information. Lastly, a number of illustrative examples of relevance measurements are discussed, including random number generation and job market signaling.

Defining a good concept of relevance is a key problem in all disciplines that theorize about information, including information retrieval (Cooper, 1971), epistemology (Floridi, 2008), and the pragmatics of natural languages (Sperber and Wilson, 1986).

Shannon information theory (Shannon, 1948) provides an interesting quantification of the notion of information, but no tools for distinguishing useless from useful facts. The microeconomic concept of value-of-information can formalize this concept in terms of expected gains (Avriel and Williams, 1970), but this notion is not easily combined with information theory, and is largely unable to exploit its tools and insights.

In this chapter, I propose a framework that integrates information theory more naturally with utility theory and thus tackles these problems. Specifically, I draw on John Kelly's application of information theory to gambling situations (Kelly, 1956). Kelly showed that when we take logarithmic capital growth as our measure of real utility, information theory can integrate seamlessly with the classical calculus of expectations. My approach here is to turn this idea on its head and base a novel notion of information on the concept of utility.

The resulting measure coincides with Shannon information when all information can be converted into strategy improvements. However, when the environment provides sources of both useful and useless information, the concept explains and quantifies the difference, and thus suggests a novel notion of value-of-information.

5.1 Doubling Rates and Kelly Gambling

The ideas in this chapter are based on some observations about the relationship about logarithmic information measures and good gambling strategies (Kelly, 1956). I will thus start by giving a largely self-contained discussion of some background concepts.

My presentation loosely follows that of Cover and Thomas (1991, ch. 6), and readers who already know this material should recognize everything up until section 5.1.4, in which I sketch the generalization which drives the rest of the argument. Throughout the paper, I assume a small amount of prior familiarity with the basic concepts of information theory, such as entropy.

5.1.1 Growth Rates and Degenerate Gambling

In many gambling situations, people evaluate a strategy in terms of its effect on the **growth rate** of their capital, that is,

$$R = \frac{\text{Posterior capital}}{\text{Prior capital}}.$$

As an example, consider a horse race with n horses with winning probabilities $p_1, p_2, \ldots, p_n$, and suppose that a bookmaker pays the odds $o_1, o_2, \ldots, o_n$ for these horses. So for example, if you bet everything on horse 1, you get your money back o_1 times if it wins and 0 times if it loses. If you split your capital into the n piles $b_1, b_2, \ldots, b_n$ and bet those on horses $1, 2, \ldots n$, your payoff is $b_i o_i$ when horse i wins.

If we normalize the initial bankroll to 1, the **expected growth rate** associated with such a betting scheme b is thus

$$E[R] = \sum_i p_i \left(b_i o_i\right)$$

This is a linear function of the betting scheme, so it always has a maximum in one of the corner points of the simplex of capital distributions. In other words, you can achieve the maximal expected growth rate by betting your whole capital on a horse with a maximal growth factor $p_i o_i$. I will call this betting scheme **degenerate gambling**, in analogy with degenerate probability distributions.

There are some reasons to disprefer degenerate gambling, however. Since it will suggest that you bet all of your capital on a single horse, it will usually also entail a substantial probability of losing all of your money. This has some highly unfortunate consequences in situations involving repeated investments and reinvestments: After k runs of the game, your initial capital of 1 will have grown or shrunk by a factor of

$$R_1 \cdot R_2 \cdot R_3 \cdots R_k.$$

If each of these factors have a positive probability of vanishing, the whole product will tend to 0 with probability 1. In other words, if you keep exposing yourself to the risk of bankruptcy, it will eventually happen with probability 1.

5.1.2 Doubling Rates

The growth of an initial stock of capital is described by a long product whose factors are random variables. This suggests that the quality of a betting scheme should be measured by a statistic which interacts just as nicely with products as expectations interact with sums.

A candidate for such a statistic is the logarithm of the growth rate,

$$W \;=\; \log R.$$

When the logarithm is base two, this quantity is called the **doubling rate** of the capital, in analogy with the half-life of a radioactive material. W measures how many times your capital is expected to double in a single game, and $1/W$ measures the average waiting time before your capital has doubled once.

A betting scheme which maximizes W is in many ways preferable to one that maximizes R. Since $2^W = R$, the capital after k horse races is

$$R_1 \cdot R_2 \cdot R_3 \cdots R_k \;=\; 2^{W_1+W_2+W_3+\cdots+W_k}.$$

The exponent in this expression is a sum of random variables. If the runs of the game are independent and identically distributed, and the gambling scheme is fixed, then the string of doubling rates will also be independent and identically distributed.

The weak law of large numbers thus applies, and we can conclude that the sum is very close to its expected value with high probability. Thus,

$$W_1 + W_2 + \cdots + W_k \;\approx\; kE[W],$$

where $E[W]$ is the **expected doubling rate**

$$E[W] \;=\; \sum_i p_i \log(b_i o_i).$$

This statistic can be both positive and negative, depending on whether the game is favorable or unfavorable. In the long run, the evolution of a stock of capital in a horse race can thus be considered as an exponential function $2^{kE[W]}$, and if you invest the capital according to the scheme which maximizes $E[W]$, you will achieve the fastest exponential growth that the game allows for (cf. Fig. 5.1).

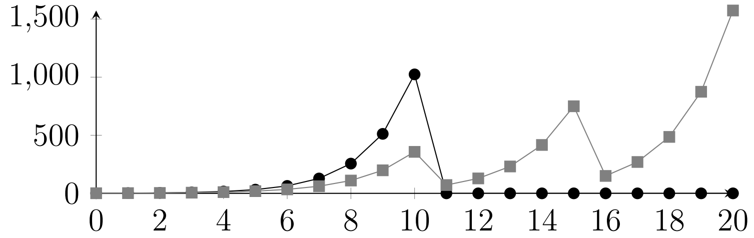


Figure 5.1: Accumulated capital after 16 rounds of betting on the outcome of a bent coin flip with bias $\theta = 9/10$ and even odds. The black dots shows a sample performance of the strategy which maximizes $E[R]$, that is, betting the entire capital on the most likely event; this leads to a short life of explosive growth and then bankruptcy. The gray squares show a sample performance of the strategy which maximizes $E[W]$, as explained in the next section; this strategy leads to a trend of moderate but consistent exponential growth.

5.1.3 Proportional Gambling in the Horse Race

What can we say about the strategy that optimizes the doubling rate? To answer this question, let us rewrite the odds o_i of the horse race in the form $o_i = c/r_i$, where c is a constant chosen so that $\sum_i r_i = 1$. For instance, the odds $o = 2, 4, 4$ would be transformed into $r = 1/2, 1/4, 1/4$. (Horses with odds 0 will never be part of an optimal solution and can be discarded from the analysis.)

Having changed representation in this way, we can rewrite the expected doubling rate as

$$\begin{aligned}
E[W] &= \sum_i p_i \log(b_i o_i) \\
&= \sum_i p_i \log\left(b_i \times \frac{c}{r_i}\right) \\
&= \sum_i p_i \log\left(\frac{1}{r_i}\right) - \sum_i p_i \log\left(\frac{1}{b_i}\right) + \log c \\
&= \sum_i p_i \log\left(\frac{p_i}{r_i}\right) - \sum_i p_i \log\left(\frac{p_i}{b_i}\right) + \log c.
\end{aligned}$$

Considering the shape of the last two expressions, it becomes apparent that the bookmaker and the gambler are in a completely symmetric situation when $c = 1$: Both are trying to express an approximation of the winning probabilities of the horses by means of the two distributions they control, namely, the bets and the odds. Whoever has the better approximation will, in the long run, make money at the expense of the other. If $c < 1$, the game is unfairly skewed in favor of the house so that the bookmaker can make money even when the gambler has a better approximation of the underlying distribution.

A compact representation of these statements is that

$$E[W] \ = \ D(p \,||\, r) + D(p \,||\, b) + \log c,$$

where $D(s \,||\, t)$ is the **Kullback-Leibler divergence** from s to t (Kullback and Leibler, 1951). This divergence is a measure of the error a probability distribution t will make in an environment in which the actual probabilities are given by s. Its minimum is obtained at $D(s \,||\, s) = 0$, and $D(s \,||\, t) > 0$ for all distributions $t \neq s$.

As either of these representations show, the optimal betting scheme for a horse race is the one that matches the underlying probabilities: If a horse wins with probability $1/3$, you should bet $1/3$ of you capital on that horse. This betting scheme is called **proportional betting**.

Somewhat surprisingly, the odds of the race are thus immaterial to the choice of betting scheme. The optimal capital distribution is determined solely by the winning probabilities.

5.1.4 The Limits of the Horse Race Model

As it turns out, the horse race model has a very simple structure: With respect to long-term growth, the optimal strategy is simply to translate your subjective probability distribution into a distribution of capital. As a consequence, updates of your subjective probabilities translate directly into updates of your capital distribution. This gives rise to a tight connection between the collection of information and increases in doubling rates.

However, this correspondence relies on a number of assumptions that are particular to the horse race model, such as the assumption that the situation involves only one random variable, and that the gambler can distribute his wealth onto the horses without any restrictions. In more complex and more realistic situations, an agent's representation of the environment may contain more variables and fewer feasible strategies. In such a situation, not all messages will afford the same opportunities for capital growth.

In fact, even a random variable which strongly affects an agent's payoff might not contain any useful information. If you don't own an umbrella, it might not be worth anything for you to know whether it rains or not, even if it changes your utility drastically. Thus, the announcement of a fact about the world does not always translate into a possibility for strategy improvement.

The suggestion I want to make in this chapter is that we take the notions of utility and strategy as primitives and derive a notion of relevance from those. This contrasts with Shannon information theory, which defines information independently of the agents using that information. It also contrasts with arithmetic value-of-information in using a logarithmic target statistic, rather than the nominal size of the capital.

Both of these assumptions lead to a number of unique features which are illustrated by several examples in section 5.3. However, in the next section, I will give a few more formal definitions and discuss some of the properties of the concept of relevant information.

5.2 Relevant Information and Relevance Rates

Relevant information is a notion of information defined in terms of utility. The notion of utility itself only makes sense in the context of agents faced with choices, so I first need to define a notion of a decision problem.

Definition 5.1. A **decision problem** $D = (S, \Omega, p, u)$ consists of

- a strategy set S;
- a sample space Ω;
- a probability measure $p : \Omega \to \mathbb{R}$;
- a utility function $u : S \times \Omega \to \mathbb{R}$.

When u is bounded and non-negative, we further define the **(expected) doubling rate** of the strategy s as

$$W(s) \; = \; \int p(x) \log u(s, x) \, dx$$

$W^* = \sup_s W(s)$ is the **optimal (expected) doubling rate** of the decision problem.

By convention, we set $W = -\infty$ if $R = 0$ with positive probability. A risk of bankruptcy will thus outweigh potential gains of any finite magnitude.

Definition 5.2. Let a decision problem $D = (S, \Omega, p, u)$ be given as above, and let Y be a random variable. Then the **posterior decision problem given the event** $Y = y$ is $D' = (S, \Omega, p', u)$, where $p'(x) = p(x \,|\, y)$. The **amount of**

relevant information in $Y = y$ is the increase in optimal doubling rate that the announcement of $Y = y$ leads to,

$$K(y) \;=\; \sup_{s\in S} W'(s) \;-\; \sup_{s\in S} W(s),$$

where W' is the doubling rate in D'. Further,

$$K(Y) \;=\; E_y\left[K(Y=y)\right] \;=\; \int p(y)\, K(Y=y)\, dy$$

is the **expected amount of relevant information** contained in Y.

Proposition 5.1. Expected relevant information is non-negative.

Proof. With respect to the marginal distribution of Y, we have the expectations

$$\begin{aligned} E_y\left[\sup_{s\in S} W'(s)\right] \;&=\; \int p(y)\left(\sup_{s\in S}\int p(x\,|\,y)\,\log u(s,x)\,dx\right)\,dy \\ &\geq\; \sup_{s\in S}\left(\int p(y)p(x\,|\,y)\,\log u(s,x)\,dx\,dy\right) \\ &=\; \sup_{s\in S}\left(\int p(x)\,\log u(s,x)\,dx\right) \\ &=\; \sup_{s\in S} W(s) \\ &=\; E_y\left[\sup_{s\in S} W(s)\right], \end{aligned}$$

where the last equality follows from the fact that W does not depend on Y. The expected posterior doubling rate is thus higher than the expected prior, and so

$$E_y\left[K(Y)\right] = E_y\left[\sup_{s\in S} W'(s) - \sup_{s\in S} W(s)\right] = E_y\left[\sup_{s\in S} W'(s)\right] \;-\; E_y\left[\sup_{s\in S} W(s)\right]$$

is non-negative. □

This proposition closely mirrors the well-known fact that Shannon information content is non-negative on average. So although bad news may occasionally represent a setback, information cannot hurt you on average. Notice also that the proof can be read as saying that an irrationally risk-averse agent can secure an unchanged average doubling rate by ignoring all incoming information.

Proposition 5.2. $1 - 2^{-K(Y)}$ is the greatest taxation rate that an agent can accept, without expected loss, in exchange for learning the value of Y.

Proof. Let $D = (S, \Omega, p, u)$ be the original decision problem, and let its prior and posterior doubling rates be W and W', respectively. Accepting a taxation rate of f in exchange for the ability to observe the value of Y will modify this problem so that the utility function in the posterior decision problem is downscaled by a factor of $1 - f$.

Let $K'(Y)$ denote the expected amount of relevant information contained in Y in this modified problem. We then have

$$\begin{aligned} K'(Y) &= \int p(y) \left(\int p(x|y) \, \log\left((1-f)u(s,x)\right) \, dx \right) dy \\ &= \int p(y) \left(\int p(x|y) \, \log u(s,x) \, dx \right) dy \; + \; \log(1-f) \\ &= K(Y) \; + \; \log(1-f). \end{aligned}$$

The agent can thus expect an on-average loss in the modified problem if and only if $K(Y) + \log(1-f) < 0$. Solving this inequality for f gives the desired result. □

To distinguish relevant information from Shannon information in the usual sense, we further define a concept of "raw" information:

Definition 5.3. Let p be a probability measure on Ω, and let X be the random variable whose values are the sample points $\omega \in \Omega$. For any random variable Y, the expected amount of **raw information** contained in Y is then

$$G(Y) \;=\; I(X;Y) \;=\; H(X) \;-\; H(X \,|\, Y).$$

Here $I(X;Y)$ is the mutual information between X and Y, and $H(X)$ and $H(X \,|\, Y)$ are the unconditional and conditional entropies of X, respectively (Cover and Thomas, 1991; MacKay, 2003).

Raw information is thus defined by comparing a variable Y to the unique state variable $X(\omega) = \omega$, that is, the maximally specific random variable. As a consequence, raw information is not a measure of dependence between two random variables in particular, but rather a measure of global decrease in uncertainty. Any source of uncertainty in your environment is thus a potential source of raw information, but not necessarily of relevant information.

These two measures of information suggest a natural measure of relevance:

Definition 5.4. Let $D = (S, \Omega, p, u)$ be a decision problem, and Y a random variable on Ω. Then the **relevance rate** of Y is $K(Y)/G(Y)$.

The relevance rate of a random variable can be both larger than and smaller than 1. However, the following theorem shows that the two coincide when an agent can bet with fair odds on the outcome of any random event whatsoever:

Proposition 5.3. Let $D = (S, \Omega, p, u)$ be a decision problem in which the strategy space is the set of probability distributions on Ω, and Y is a random variable on Ω. Suppose further that the utility function u has the form $u(s, x) = s(x)v(x)$ for some non-negative, real-valued function v. Then $K(Y) = G(Y)$.

Proof. This observation is due to Kelly (Kelly, 1956). We prove it by noting that a utility function of the form $u(s, x) = s(x)v(x)$ leads to a doubling rate of the form

$$\begin{aligned} W(s) &= \int p(x) \log s(x)v(x)\, dx \\ &= \int p(x) \log \left(\frac{p(x)}{o(x)}\right) dx - \int p(x) \log \left(\frac{p(x)}{s(x)}\right) dx \\ &= D(p \,||\, o) - D(p \,||\, s), \end{aligned}$$

where $o = 1/v$ can be interpreted as the odds, and s as the bets.

As a consequence of Jensen's inequality, the unique minimum of $D(p \,||\, s)$ is $s = p$. The doubling rate is thus maximized by proportional betting ($s = p$), regardless of the probability environment and the odds. The optimal doubling rate under the distribution p is thus

$$\begin{aligned} \sup_{s \in S} W(s) &= D(p \,||\, o) - D(p \,||\, p) \\ &= D(p \,||\, o) \\ &= \int p(x) \log \left(\frac{p(x)}{o(x)}\right) dx. \end{aligned}$$

Similarly, the optimal doubling rate given the condition $Y = y$ is

$$\sup_{s \in S} W'(s) = \int p(x|y) \log \left(\frac{p(x|y)}{o(x)}\right) dx$$

Taking expectations with respect to Y and subtracting the prior doubling rate from the posterior, we find

$$\begin{aligned} K(Y) &= E_y \left[\sup_{s \in S} W'(s)\right] - E_y \left[\sup_{s \in S} W(s)\right] \\ &= E_y \left[\int p(x|y) \log p(x|y)\, dx\right] - E_y \left[\int p(x) \log p(x)\, dx\right] \\ &= -H(X|Y) - (-H(X)) \\ &= G(X). \end{aligned}$$

This establishes the desired equality. □

5.3 Examples of Relevance Measurements

Having now introduced the notion of relevance, I would to like present a series of examples that illustrate how and where the concept can be used.

In the following five subsections, I will thus sketch five different gambling situations and analyze their informational properties. The first three cases are intended as toy examples illustrating the mechanics of relevant information, while the last two hint in the direction of more realistic applications.

5.3.1 A Horse Race with Pocket

Consider a horse race with winning probabilities $p = (2/3, 1/3)$ and odds $o = (2, 2)$. If you place a proportion b of your capital on the first horse in this race, your doubling rate will be

$$E\left[W(b)\right] \;=\; \frac{2}{3}\log\left(2b\right) + \frac{1}{3}\log\left(2(1-b)\right).$$

As discussed in section 5.1, this doubling rate is maximized by $b^* = 2/3$, and the optimal doubling rate is then

$$E\left[W\left(b^*\right)\right] \;=\; \frac{2}{3}\log\left(\frac{2\cdot 2}{3}\right) + \frac{1}{3}\log\left(\frac{2\cdot 1}{3}\right) \;=\; 0.08 \text{ bits.}$$

However, this assumes that you invest your whole capital in the horse race. What does the situation look like if you could keep some of your money in your pocket rather than risking it in the game?

In such a situation, a gambling strategy would be defined by two parameters, the fraction f of capital invested in the game, and the fraction b of that capital betted on the first horse. Allowing f to take other values than 1 would change the payoff structure of the game as informally summarized in Table 5.1. In effect, the modified race would have a third horse which always paid one cent of winnings for one cent of bets.

In this modified game, the expected doubling rate associated with a strategy (f, b) is

$$E\left[W(f,b)\right] \;=\; \frac{2}{3}\log\left(1 - f + 2bf\right) + \frac{1}{3}\log\left(1 - f + 2(1-b)f\right).$$

By differentiating, we find that this doubling rate is optimal when

$$b \;=\; \frac{3f+1}{6f}.$$

It thus turns out that the doubling rate function has an optimal ridge running through the unit square, as shown in Figure 5.2. This ridge contains the old

Winner	1	2
Returns	2	0
Returns	0	2

Winner	0	1	2
Returns	$1-f$	$2f$	0
Returns	$1-f$	0	$2f$

Table 5.1: Payoff structures for two different gambling situations: Once in which each situation comes with a single and unique source of income, and one in which there is a universal “pocket” source with a fixed rate of return.

optimum $(f, b) = (1, \nicefrac{2}{3})$, but also other equally good strategies like $(\nicefrac{1}{2}, \nicefrac{5}{6})$ and $(\nicefrac{1}{3}, 1)$. Since the odds of the game favors the gambler, investments less than $f = \nicefrac{1}{3}$ are strictly irrational.

Note that if we somehow learned that the first horse would win the game, the expression for the expected doubling rate would reduce to

$$E\left[W(f, b)\right] \;=\; \log\left(1 + (2b - 1)f\right).$$

This expression is optimized by $f = b = 1$, corresponding to the fact that one should invest as much money as possible in a rigged game. With this betting scheme, the doubling rate would be $\log 2 = 1$, consistent with the fact that this game would deterministically double your capital.

Note that in this horse race with a pocket, it turned out that the relevance analysis did not actually lead to a new result: Since an optimal solution existed which did not make use of the pocket, information about the winning horse had the same relevance rate in either game. However, we are now in a better position to vary more parameters, and the following examples will present some scenarios in which relevant and raw information diverge.

5.3.2 Guessing with Optional Investment

Suppose you can invest any fraction of your capital in a lottery defined as follows: If you can guess the four binary digits of my credit card code in a single try, you get your investment back 16-fold; otherwise, you lose it. Note that unlike the horse race, this game does not allow for arbitrary capital distributions.

Let us assume that you invest a fraction f of your capital in this lottery and keep a fraction of $1 - f$ in your pocket. Since your chance of guessing correctly is $1/2^4$, your expected doubling rate will be

$$E\left[W(f)\right] \;=\; \left(\frac{1}{2^4}\right)\log(1 - f + 16f) + \left(1 - \frac{1}{2^4}\right)\log(1 - f).$$

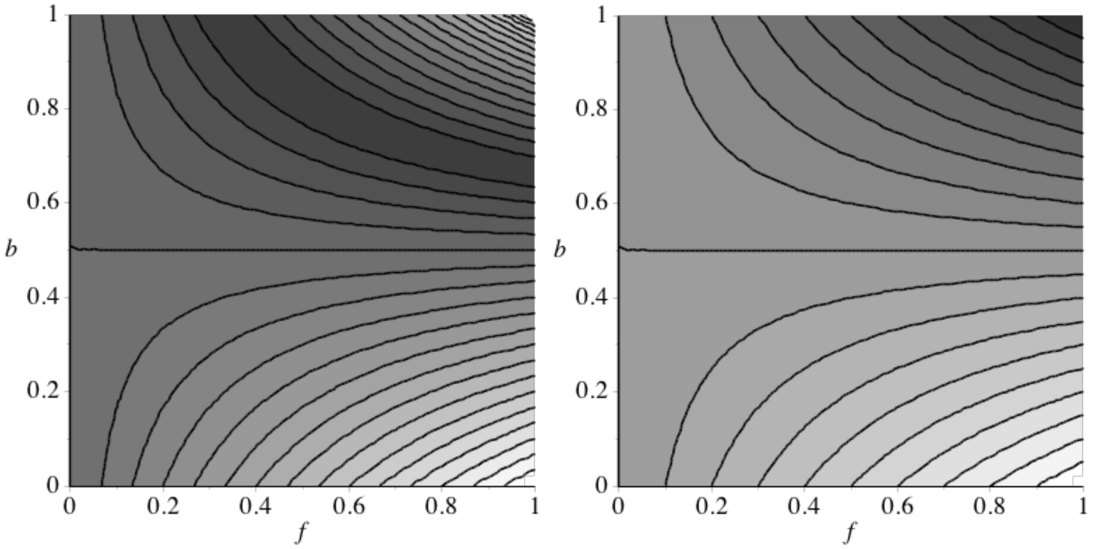


Figure 5.2: Prior and posterior doubling rates for various combinations of f and b. The left graph shows the contours of W when the first horse wins with probability $p_1 = {}^2\!/_3$, and the right graph shows the same for $p_1 = 1$. Darker hues correspond to higher values, but for visibility, the two graphs use different color scales.

This is a decreasing function in f, and your optimal strategy is $f^* = 0$, i.e., not betting anything.

However, suppose that you have an inside source that can supply you with some of the digits of the credit card code. For each digit you receive, your chance of guessing the code in a single attempt obviously increases; more specifically, the new expected doubling rate will be

$$E\left[W_i(f)\right] \;=\; \left(\frac{2^i}{2^4}\right)\log(1+15f) + \left(1-\frac{2^i}{2^4}\right)\log(1-f)$$

after you have received i of the four digits. As illustrated in Figure 5.3, this function attains its maximum on the unit interval at $f^* = {}^0\!/_{15}, {}^1\!/_{15}, {}^3\!/_{15}, {}^7\!/_{15}, {}^{15}\!/_{15}$ after $i = 0, 1, 2, 3, 4$ bits of the code has been revealed.

The optimal expected doubling rates in these cases are

$$W_i^* \;=\; 0.00,\ 0.04,\ 0.26,\ 1.05,\ 4.00, \qquad \text{for } i = 0, 1, 2, 3, 4.$$

It thus turns out that the four digits you receive are not equally relevant to you. The first contains only 0.04 bits of relevant information although the revealed digit contained one bit of raw information. The second contains 0.22 bits of relevant information per bit of raw information, the third 0.79, and the fourth 2.95.

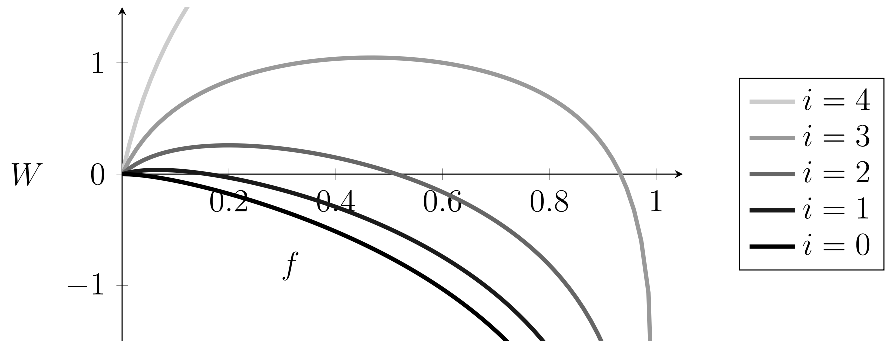


Figure 5.3: Doubling rates in the guessing game. The graphs shows the doubling rate as a function of investment level, assuming i messages have been received.

The raw and the relevant information add up to the same number, 4 bits of information, but do so at different paces. This difference in accumulation speed is only present because you are not forced to invest all your money in the lottery: If you were, all four bits would supply you with exactly one bit of relevant information, giving them a relevance rate of 1.

5.3.3 Guessing with Irrelevant Side-Information

Continuing the code-guessing scenario as above, suppose now that you receive your side-information about my credit card code from an unreliable source which may abort the communication at any time. In this case, you have uncertainty about two independent variables, my actual credit card code (C), and the number digits you will receive (L).

Since receiving a digit removes uncertainty not only about C but also about L, you will, paradoxically, receive more than one bit of raw information per transmitted digit under these assumptions. The amount of relevant information you receive, however, will remain the same as in the previous scenario, since the information about L has no bearing on your guessing strategy.

To make this more concrete, suppose that just before transmitting each character, your source flips a coin to decide whether to continue or abort. This means that L takes the five values 0, 1, 2, 3, and 4 with probabilities $1/2$, $1/4$, $1/8$, $1/16$, and $1/16$, respectively (cf. Fig. 5.4). Excluding one of those possible outcomes at a time, beginning from the left, gives the conditional entropies shown in Table 5.2.

By subtracting these entropy levels from each other, we find that the four digits you receive contain $9/8$, $5/4$, $3/2$, and 2 bits of raw information, respectively. However, the optimality of a strategy in this game depends only on your chance of guessing the code in a single try. The amount of relevant information contained in the messages consequently remains as in the previous example.

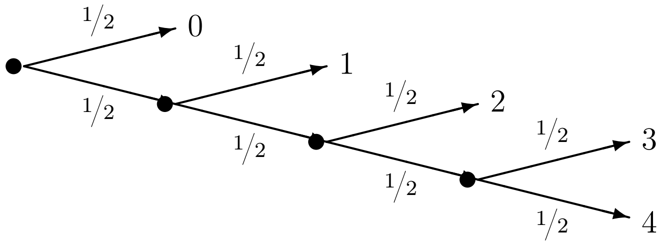


Figure 5.4: A probability distribution on L, the number of digits you receive from the undependable source in example 5.3.3.

i	0	1	2	3	4
$H(C)$	4	3	2	1	0
$H(L)$	$15/8$	$7/4$	$3/2$	1	0
Sum	$47/8$	$19/4$	$7/2$	2	0

Table 5.2: The decreasing uncertainty about the code (C) and transmission length (L) after i digits of my credit card code have been revealed.

This example illustrates how the addition of a variable to a model can change the information-theoretic analysis of a situation without necessarily the decision-theoretic. By my judgment, the notion of relevant information captures many intuitively important aspects of this scenario.

5.3.4 Randomization

Suppose the two of us put down 1 cent for a game of Rock-Paper-Scissors, and that the winner gets both coins. If you play the three moves with probabilities $p = (p_1, p_2, p_3)$, and I play them with probabilities $q = (q_1, q_2, q_3)$, then your expected payoff is

$$u_1(p, q) \;=\; q_1(p_1 + 2p_2) + q_2(p_2 + 2p_3) + q_3(p_3 + 2p_1).$$

My expected payoff is $u_2 = 2 - u_1$.

This function depends linearly on the probabilities I assign to the three moves, and it consequently has a maximum and a minimum in a corner point. Whatever your strategy is, one of my three pure strategies $(1, 0, 0)$, $(0, 1, 0)$, or $(0, 0, 1)$ is thus a best response. Because of the zero-sum nature of the game, this best response will also minimize your expected payoff.

From your perspective, the consequence is that if you have chosen some randomized strategy $p = (p_1, p_2, p_3)$, and I have chosen a deterministic response q^* which minimizes $u_1(p, q)$, then your expected payoff is

$$u_1(p, q^*) \;=\; \min\{p_1 + 2p_2,\, p_2 + 2p_3,\, p_3 + 2p_1\},$$

and your doubling rate is the logarithm of this minimum. These quantities are optimal when you use the uniform strategy $p = (1/3, 1/3, 1/3)$.

This describes the game-theoretical aspects of this situation from an abstract, normative perspective. However, as Claude Shannon noted in the early 1950s (Zucker and Meyer, 2000), real people are in fact curiously bad at making random choices, and they invariably slip computable structure into the "random" sequences they produce. This means that a computer (or a statistician) equipped with a simple inference algorithm often outperforms humans vastly in randomization games such as Matching Pennies or Rock-Paper-Scissors.

The purpose of this example is to present a model of this limitation, and to measure how much it would change the situation if people had access to additional randomization resources like random number tables, coins, or quantum-random hardware. For clarity, I will analyze the most extreme case of this situation, namely that of a purely deterministic device that plays Rock-Paper-Scissors using a finite number of calls to a randomization oracle.

Suppose therefore that you have to play Rock-Paper-Scissors by submitting a publicly accessible program for a Turing machine. Since the program is completely deterministic, your strategy is going to be completely predictable, and your opponent can adapt perfectly to your strategy. This leads to a doubling rate of $\log \min\{0, 1, 2\} = -\infty$.

However, suppose now that your Turing machine has a module which can request a fixed number of fair coin flips per game. You can then encode these coin flips into the strategy in order to make it less predictable. The optimal way to do this is to feed the coin flips into an arithmetic decoder (MacKay, 2003; Cover and Thomas, 1991) which translates them into an approximately uniform distribution on $\{R, P, S\}$. The more coin flips you have, the flatter this distribution can get.

This situation is depicted in Table 5.3, which shows the arithmetic payoffs that you can achieve with i calls to the coin flipping module. As the table shows, the first coin flip will contain infinitely much relevant information, since it increases your arithmetic payoff from 0 to 1/2. The second contains

$$\log 3/4 - \log 1/2 \;=\; 0.59 \text{ bits of relevant information.}$$

The third, fourth, and fifth contain 0.22, 0.10, and 0.05 bits, respectively. As you add more calls to the randomization module in your program, the marginal benefit of adding another one decreases quite rapidly.

Readers familiar with Kelly gambling should note the difference between a horse race and the present model. In the horse race, a gambler chooses bets but does not control the probabilities. In the present example, the reverse is true.

i	0	1	2	3	4	5	6	7	$\cdots$	∞
p_1	$^1/_1$	$^1/_2$	$^2/_4$	$^3/_8$	$^6/_{16}$	$^{11}/_{32}$	$^{22}/_{64}$	$^{43}/_{128}$	$\cdots$	$^1/_3$
p_2	–	$^1/_2$	$^1/_4$	$^3/_8$	$^5/_{16}$	$^{11}/_{32}$	$^{21}/_{64}$	$^{43}/_{128}$	$\cdots$	$^1/_3$
p_3	–	–	$^1/_4$	$^2/_8$	$^5/_{16}$	$^{10}/_{32}$	$^{21}/_{64}$	$^{42}/_{128}$	$\cdots$	$^1/_3$
u_1	0	$^1/_2$	$^3/_4$	$^7/_8$	$^5/_{16}$	$^{31}/_{32}$	$^{63}/_{64}$	$^{127}/_{128}$	$\cdots$	1

Table 5.3: Payoffs for increasingly randomized rock-paper-scissors strategies.

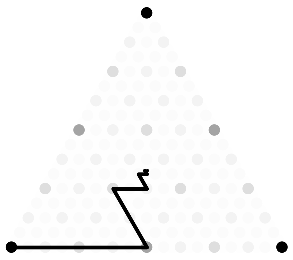


Figure 5.5: A graphical representation of the process of approximating a uniform distribution using an increasing number of coin flips.

Notice also that just like the horse race model implicitly measures deviations from the optimal strategy in terms of the Kullback-Leibler divergence, this model too defines a notion of distances from the uniform distributions on $\{R, P, S\}$. In this sense, the notion of relevant information as it appears here is a relatively natural generalization of entropy as a measure of how far away we are from an optimal solution (e.g., as in the context of ideal codeword lengths).

5.3.5 Non-cooperative Pragmatics

Following loosely the ideas by Spence (1973) and Glazer and Rubinstein (2006), suppose you regularly hire new staff from a pool of people that have taken two qualifying exams. Suppose further that the grades on these two exams, X and Y, are distributed uniformly on the set $\{1, 2, 3, \ldots, 10\}$. We can define the productivity of a hired person as units of profit per unit of salary, and we may assume that this profit rate depends on the two qualifying grades as

$$R = \frac{X + Y}{10}.$$

Hiring a person will thus in general affect your doubling rate $W = \log R$ either negatively or positively, depending on whether that person is qualified above or below a threshold of $X + Y = 10$ (cf. Fig. 5.6a and Table 5.4).

10	0.14	0.26	0.38	0.49	0.58	0.68	0.77	0.85	0.93	1.00
9	0.00	0.14	0.26	0.38	0.49	0.58	0.68	0.77	0.85	0.93
8	−0.15	0.00	0.14	0.26	0.38	0.49	0.58	0.68	0.77	0.85
7	−0.32	−0.15	0.00	0.14	0.26	0.38	0.49	0.58	0.68	0.77
6	−0.51	−0.32	−0.15	0.00	0.14	0.26	0.38	0.49	0.58	0.68
5	−0.74	−0.51	−0.32	−0.15	0.00	0.14	0.26	0.38	0.49	0.58
4	−1.00	−0.74	−0.51	−0.32	−0.15	0.00	0.14	0.26	0.38	0.49
3	−1.32	−1.00	−0.74	−0.51	−0.32	−0.15	0.00	0.14	0.26	0.38
2	−1.74	−1.32	−1.00	−0.74	−0.51	−0.32	−0.15	0.00	0.14	0.26
1	−2.32	−1.74	−1.32	−1.00	−0.74	−0.51	−0.32	−0.15	0.00	0.14
	1	2	3	4	5	6	7	8	9	10

Table 5.4: The doubling rates associated with different combinations of grades.

However, under the distribution assumed here, it is in fact rational to hire a person in the absence of any information about that person's skill level. This holds because your average doubling rate across the whole pool of applicants, $E\,[W]$, is slightly larger than 0:

$$\begin{aligned} E\left[\log\frac{X+Y}{10}\right] &= \sum_{i=1}^{10}\sum_{j=1}^{10}\Pr(X=i,Y=j)\times\left(\log\frac{i+j}{10}\right) \\ &= 15.23 \text{ millibits.} \end{aligned}$$

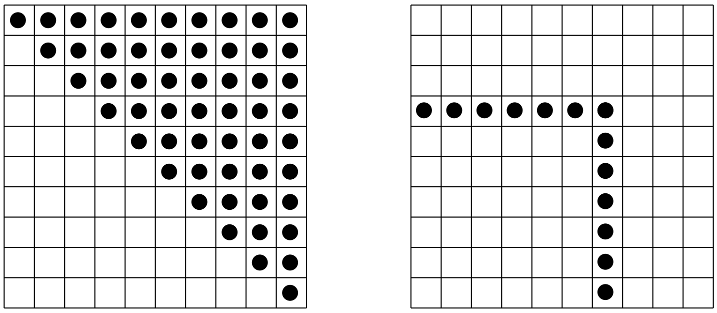
Hiring a randomly plucked person will thus give you an expected productivity of $E\left[2^W\right] = 2^{E[W]} = 2^{0.01523} = 1.01$ units of profit per unit of salary.

However, suppose you take a person into an interview, and that person shows you one of his or her grades. Assuming that you were shown the largest of the two grades, how much relevant information does this piece of data give you? At which grade level should you hire the applicant?

To answer this question, let $M = \max\{X, Y\}$ be the grade you were shown. The doubling rate you can expect from hiring an applicant with $M = m$ is then

$$E\,[W \mid M = m] \;=\; \sum_{i=1}^{10}\sum_{j=1}^{10}\Pr(X=i,\,Y=j \mid M=m)\log\frac{i+j}{10}.$$

By fixing m and summing up over all pairs (i, j) for which $i, j \leq m$ and $i = m$ or $j = m$ (cf. Fig. 5.6b), this doubling rate turns out to be negative for $m < 7$ and positive for $m \geq 7$. In other words, hiring a person whose largest grade is smaller



(a) (b)

Figure 5.6: (a) The applicants with positive productivity. (b) The applicants whose largest grade is $M = 7$ (see also Table 5.4).

than 7 will, on average, lead to a loss. The optimal decision in that case is thus to keep the salary in your pocket, retaining an expected doubling rate of 0.

So, observing $m < 7$ leads to a doubling rate of 0. On average, the doubling rate resulting from learning the value of M will thus be

$$\sum_{m=7}^{10} \Pr(M = m) \times E\left[W \mid M = m\right].$$

The probabilities in this sum are of the form

$$\Pr\left(M = m\right) \;=\; \frac{2m-1}{100},$$

and the expected doubling rates are 0.08, 0.27, 0.44, 0.59 for m = 7, 8, 9, 10. A bit of computation then gives the posterior doubling rate $E\left[W'\right] = 0.24$ bits. Since the prior doubling rate was $E\left[W\right] = 0.01523 \approx 0.02$, an announcement of the value of M will on average give you

$$K(M) \;=\; 0.24 - 0.02 \;=\; 0.22 \text{ bits of relevant information.}$$

It follows that if you can observe the applicant's maximal grade before hiring him or her, your expected capital will, on average, grow by a factor of

$$R \;=\; 2^{E[W']} \;=\; 2^{0.24} \;=\; 1.18.$$

Further, you should thus be willing to trade up to $1 - 2^{-0.22} = 14.1\%$ of your future profits in return for this piece of information.

Finally, we will compute the amount of raw information contained in M. Observing $M = m$ narrows down the space of possible values for $X \times Y$ so that it has $2m-1$ possible values instead of 100. Since these values are equally probable, the amount of information contained in the message $M = m$ is

$$H(X \times Y) \; - \; H(X \times Y \,|\, M = m) \; = \; \log 100 \; - \; \log(2m-1).$$

To compute the average value of this quantity, we again note that $M = m$ has point probabilities $\frac{2m-1}{100}$. On average, the information gain resulting from learning the value of M is thus

$$G(M) \; = \; \sum_{m=1}^{10} \left(\frac{2m-1}{100} \right) \times (\log 100 - \log(2m-1)).$$

Computing this sum, we find that $M = \max\{X, Y\}$ contains 3.05 bits of raw information. However, as we have seen, learning its value only buys you an increase of $0.24 - 0.02 = 0.22$ bits in doubling rate on average. M thus has a relevance rate of

$$\frac{K(M)}{G(M)} \; = \; \frac{0.22}{3.05} = 0.07$$

bits of relevant information per bit of raw information.

5.4 Conclusion

In this chapter, I have proposed a logarithmic of value-of-information measure as a quantitative elaboration of the concept of relevance.

This leads to an agent-oriented measure of relevance, as opposed to a system-oriented one (Borlund, 2003): It takes relevance to be a relation between events and agents rather than events and events. Because of this connection to agents and their utilities, the approach taken here forms a natural bond with ideas from decision theory, Bayesian statistics, and Shannon information theory. It thus represents a fairly conservative extension of the calculus of reasoning which is already canonical in the behavioral sciences.

This concept of relevance may shed some new light on the ways in which dynamics of information can interact with problems of resource allocation. The examples I have given can, I believe, only be fully understood if we see the microeconomic and the information-theoretic aspects of the situation as two sides of a single coin. The notion of relevant information might be one out of several paths into such a style of analysis.

Chapter 6

Multi-Agent Probability Models of Social Cognition

Computational models of social cognition tend to fall into one of two separate classes, those based primarily on logical methods, and those based primarily on probability theory. In this chapter, I suggest that there is something to be gained by integrating the insights from these two traditions, borrowing the versatile multi-agent uncertainty representations from the logical toolbox and the fine-grained inference methods from probability theory. I explain how to construct and use models that include both of these sets of features and then apply these tools to two case studies that have previously been modeled by conventional Bayesian models.

6.1 Introduction

Recent years have seen a surge of interest in the computational modeling of social interaction. Although many finer distinctions can be made, two large-scale trends are discernible in this literature.

One such trend is the use of tools from probability theory and applied statistics. For instance, Baker et al. (2011) constructed a hidden Markov model that could mimic how human subjects guessed the preferences of a little cartoon figure moving around on a stylized map; Goodman et al. (2006) used a similar Bayesian model to explain why children tend to rationalize other people's behavior in terms of preferences rather than beliefs up to a certain age; and a number of such models have been used to explain linguistic behaviors such as pragmatic inferences from asserted content to intended meaning (e.g., Franke and Jäger, 2014).

Another major approach is the use of various logical systems, often borrowed from artificial intelligence or computer science. For instance, Arkoudas and Bringsjord (2009) used an event-calculus model to explain how children reason about other people's beliefs; Stenning and Van Lambalgen (2008, Ch. 9.4) explained why this reasoning often goes astray by modeling it as a kind of closed-world reasoning that sometimes fails to override rules of thumb; and Braüner (2014) proposed a related model based on hybrid logic, suggesting that it would provide more accurate predictions under certain circumstances.

These models all differ in many interesting and important details, and some of them may produce more accurate predictions than others. The purpose of this chapter, however, is not to decide these questions of empirical merit; it is rather to devise a shared language that can bridge the gap between the two types of computational model. Having such a shared language may allow us to draw on insights from the two traditions that are often present in isolation, but rarely combined.

Specifically, some benefits of the probabilistic models are that they can

- deal more easily with quantitative or noisy data;
- model, in a very natural fashion, the process of learning from evidence.

On the other hand, the advantages of the logical models are that they

- allow flexible and explicit representations of belief states in multi-agent situations;
- provide systematic methods for constructing the models and knowledge structures on the basis of protocols of events.

Although all of these model features are very desirable, they are rarely if ever attained simultaneously. For instance, Goodman et al. (2006) use a special-purpose "belief variable" rather than a generalizable representation scheme, and they note that they have no explanation of how the child might come to construct the right probability distribution in the first place (p. 1387). On the other hand, while Stenning and Van Lambalgen (2008) pinpoint the possible computational reasons that lead to the conflation of one's own and others' beliefs, they need to invoke physiological rather than logical reasons in order to explain why only some children are able to make these distinctions (cf. pp. 254–255).

In this chapter, I will argue that we can have it both ways. By using a collection of methods developed under the banner of dynamic logic, we can systematically construct probability distributions based on a series of events, and multi-agent knowledge representations based on facts about who sees what. These models can contain probabilistic uncertainty on all levels, and they can therefore form the basis of probabilistic learning in the usual ways.

In the following two sections, I will first present some general background on these ideas, and then give a few concrete examples of how they can be applied to specific modeling tasks. Throughout, my presentation works on the assumption that our goal is to construct a probabilistic multi-agent model, and that we assume as much or little logic as is necessary to reach that goal.

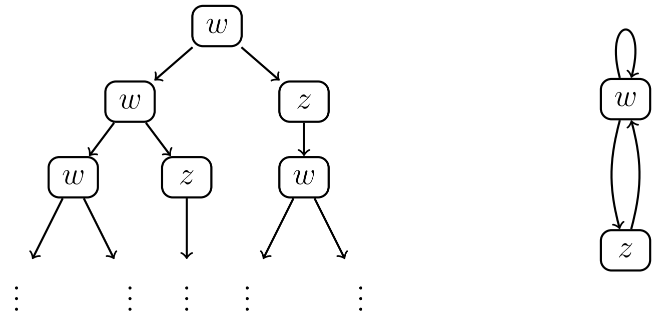


Figure 6.1: Left, an infinitely deep tree that might be used to represent arbitrarily deep beliefs about beliefs. Right, a structure that represents the same relations in a finite but loopy network.

6.2 Multi-Agent Uncertainty Representation

In order to model social cognition in a probabilistic framework, we need to choose a formalism that is capable of representing beliefs about beliefs. An obvious candidate for this job would be a probability distribution over probability distributions.

This encoding will work as long as we only want to model a single agent's reasoning about the factual knowledge of another agent. However, if we want to model beliefs about beliefs about beliefs to an arbitrary depth, this approach will fail. The only way we can encode all these layers of reasoning in set of nested distributions is by using an infinitely deep tree. We thus need a compression strategy.

The classical solution to this problem is to use a cyclic graph instead of a tree (Kripke, 1963). Such a structure will permit arbitrary depths of reasoning, since a nested process like reasoning about someone else's reasoning will then correspond to walking around on this graph rather than descending down through the tree. For instance, the act of randomly selecting a probability distribution and then sampling from that distribution will correspond to taking two consecutive steps on this graph. At the cost of introducing a mild form of circularity, we can thus fold an infinite tree up into a finite graph (cf. Fig. 6.1).

The nodes in such a representation correspond to specific states of the world, or "possible worlds" in the terminology of modal logic. The possible worlds are the finest grain of description in the model, and they settle all matters of fact. For the purposes of probabilistic multi-agent reasoning, a possible world also needs to prescribe a choice of probability distribution for each agent in the model (van Benthem et al., 2009).

An agent's uncertainty about some question φ is thus always represented in terms of uncertainty about what world we are in. This also goes for uncertainty about the beliefs of other agents. In any particular world, the agents will have a probability distribution over the entire logical space, and this distribution will allow them to assign probabilities not only to factual claims but also to propositions about the distributions of other agents. At a world ω, an agent A can for instance assign a probability to the claim that agent B considers the proposition φ is more likely than not. Similarly, agent A can compute her own mean $E_A[X]$ of some random variable X, but she can also make probabilistic statements about the plausible range of the, for her, unknown quantity $E_B[X]$.

Such nested probability evaluations are the core of the calculus of multi-agent probability. In the following subsections, I will make this informal discussion more precise, often deviating from standard presentations of the topic in order to increase clarity or make the connection to conventional probability theory more obvious.

Although the topic is multi-agent probability theory, I will begin by introducing a series of concepts that, formally speaking, only refer to one single agent. However, due to the unusual way these concepts are defined, they are relatively easy to duplicate and combine in a multi-agent situation. How this works will be clear later in the chapter.

6.2.1 Kripke Frames

As mentioned above, the crucial idea behind the concept of multi-agent probability is to use a family of different distributions to represent an agent's beliefs. We capture this idea in the following definition:

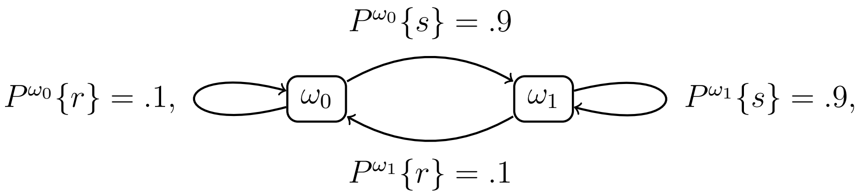
Definition 6.1. A **single-agent Kripke frame** is function from some sample space Ω to the set of probability measures on Ω.

A Kripke frame thus sticks a probability distribution to each sample point $\omega \in \Omega$. This distribution will be interpreted as the beliefs of an agent in that particular state of the world. I will denote this specific distribution by P^ω and the frame itself (i.e., $\omega \mapsto P^\omega$) by P.

When Ω is finite, a Kripke frame on Ω can be exhaustively described by a table of probability distributions, as in the following two-world example:

ω	$P^\omega\{\omega_0\}$	$P^\omega\{\omega_1\}$
ω_0	0.1	0.9
ω_1	0.1	0.9

As discussed above, we can also visualize the frame as a graph:



This specific frame represents the uncertainty that will result if A flips a highly bent coin without looking at the outcome. There are then two possible worlds, ω_0 and ω_1, but A can not distinguish between them.

Definition 6.2. A single-agent Kripke frame P is **coherent** if

$$P^{\omega_0}\left\{\omega_1 \in \Omega :\ P^{\omega_1} = P^{\omega_0}\right\} \;=\; 1 \qquad \text{for all } \omega_0 \in \Omega.$$

Coherence means that the agent cannot hold two different beliefs in two indistinguishable situations (Aumann, 1976). All Kripke frames considered in this chapter are coherent.

6.2.2 Announcements

In conventional probability theory, an agent's beliefs are represented by a probability measure $P(\,\cdot\,)$ on a space Ω. The observation of a fact $\varphi \subseteq \Omega$ can then be modeled by replacing $P(\,\cdot\,)$ by the conditional measure $P'(\,\cdot\,) = P(\,\cdot\,|\,\varphi)$.

However, this approach does not generalize very well to the multi-agent case, since it does not lend itself very easily to the modeling of differentiated information levels. We therefore need to distinguish between learning that a variable X has a specific value (such as $X = 0$) and learning the value of X (whatever it is). The following definition, which is a reformulation of a similar concept from van Benthem et al. (2009), formalizes the latter of these concepts.

Definition 6.3. Suppose a Kripke frame P is given. Then the **conditional Kripke frame** $P(\,\cdot\,|\,X)$ **given the announcement of** X is the function

$$\omega \;\mapsto\; P^{\omega}(\,\cdot\,|\,X = X(\omega)).$$

The conditional Kripke frame $P(\,\cdot\,|\,X)$ represents an agent's knowledge after she learns the value of X. We can thus read "$P(\varphi\,|\,X)$" as a shorthand for "the conditional probability that the agent assigns to φ given the value of X (whatever it is)." Note that X can be any variable that has a well-defined value in each possible world. It can, for instance, be the indicator function of a proposition, or a function of certain other random variables that have already been defined.

Another way of thinking about such an announcement is that it cuts up the sample space Ω into slices of the form $\{X = x\}$, one for each value x (cf. Fig. 6.2). The effect of the announcement of X is thus to improve the agent's powers of

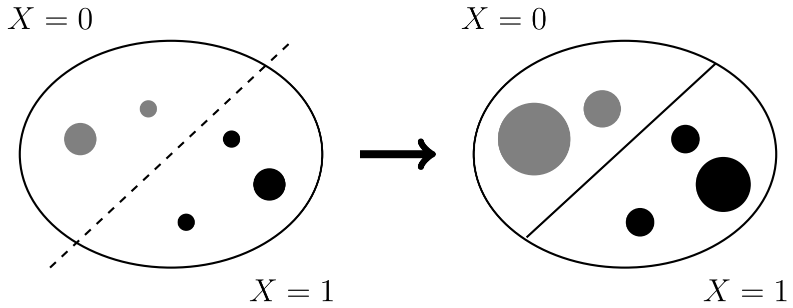


Figure 6.2: The effect of an announcement on a Kripke frame. Before the announcement of X, all possible worlds are equipped with the same probability distribution. After the announcement, the probability distributions in $\{X = 0\}$ are replaced by their conditional counterparts given $\{X = 0\}$, and the distributions in $\{X = 1\}$ are replaced by their conditional counterparts given $\{X = 1\}$.

discrimination, since two worlds ω_0 and ω_1 that might previously have been indistinguishable become distinguishable if $X(\omega_0) \neq X(\omega_1)$.

For instance, if $X(\omega_0) = 1$ and $X(\omega_1) = 0$, then the announcement of X will have the following effect on the coin flipping example considered above:

ω	$P^\omega\{r\}$	$P^\omega\{s\}$
ω_0	0.1	0.9
ω_1	0.1	0.9

Before

ω	$P^\omega\{r\}$	$P^\omega\{s\}$
ω_0	1	0
ω_0	0	1

After

Knowing the value of X may lead the agent to split up Ω into thinner slices. The thinnest possible slices are achieved by learning the variable of the state variable $X(\omega) = \omega$, and a constant variable like $X(\omega) = 1$ corresponds to a single, thick slice.

The following fact might be worth noting for its own sake:

Proposition 6.1. If the Kripke frame P is coherent, then so is the conditional Kripke frame given the announcement of X.

Proof. By the definition of conditional probability, $\{\omega_1 : X(\omega_1) \neq X(\omega_0)\}$ has probability 0 given $X = X(\omega_0)$. Hence, P^{ω_1} and P^{ω_0} will be updated by the same condition and thus remain identical with probability 1. □

6.2.3 Uncertain Announcements

The definition in the previous section defines what it means for an agent to make an observation. However, it does not allow very easily for the modeling of uncertain observations. For instance, if I roll a die and leave the room, I do not know whether you look at the die while I'm gone. This uncertainty about your possible states of information cannot be represented using the announcements described in the previous subsection.

It is conventional in dynamic logic to model such private and secret announcements by means of special-purpose epistemic actions (Baltag et al., 1998). However, the coherence property that I assume in this chapter has the consequence that we can achieve the same effects merely by the use of convenient dummy variables. Hence, we do not actually need to introduce any additional definitions, only to point out a clever way of using the existing apparatus of conditional Kripke frames.

As a simple example of the general idea, suppose that R is a random variable which takes one of the two values $\{\pm 1\}$, and suppose that we may or may not want to announce the value of R to an agent A. In order to express this uncertain announcement, we flip a coin $K \in \{0, 1\}$ and then announce the product KR to our agent. This product has three different values, -1, 0, and $+1$, and the announcement of KR consequently splits the sample space up into three different cells:

	$R = +1$	$R = -1$
$K = 0$	0	0
$K = 1$	$+1$	-1

Thus, when $KR = \pm 1$, the agent can infer the value of R; but when $KR = 0$, nothing is revealed about R. This trick thus achieves the goal of producing a conditional announcement.

Suppose now that R takes instead on the two values $\{0, 1\}$. This means that we cannot use the same trick as above. However, it should be clear that the only feature of the function $R \mapsto KR$ we really needed was the fact that when $K = 0$, it worked as a "black hole" that threw all values of R into the same, uninformative bin. Rather than relying on an arithmetical accident to produce this effect, we can also enforce explicitly directly by constructing a random variable that takes a Null value whenever $K = 0$:

$$C(K, R) \;=\; \begin{cases} R & (\text{if } K = 1) \\ \text{Null} & (\text{if } K = 0) \end{cases}$$

The Null value can be anything, as long as it is not one of the existing values of R. The effect of announcing the value of C will then be to divide Ω up into one large chunk where $K = 0$, and a number of smaller slices where $K = 1$, while R ranges over all its values. Again, this achieves the effect of defining an announcement which is only made when a certain condition is met.

Note again that K can either be a newly minted variable or a conditionally deterministic function of the existing variables, such as the indicator of a particular proposition. We may even, by pasting together several conditional announcements of this kind, create the effect of a random variable that toggles between several different possible announcements on different regions of Ω. This can be useful for modeling, for instance, the effect of an agent that looks in a random compass direction and thus learns a random fact about the environment. For the purposes of this chapter, however, this kind of stochastic meta-uncertainty will not be necessary.

6.2.4 Multi-Agent Probability

The concepts introduced in the preceding subsections were deliberately defined so as to allow easy generalization to the multi-agent case. All we need to do in order to make the transition is to make more copies:

Definition 6.4. A **multi-agent Kripke frame** on a sample space Ω is a family of single-agent Kripke frames on Ω, one for each agent in some index set N.

I will use the notation P_A to designate the single-agent Kripke frame representing agent A's knowledge state. The multi-agent Kripke frame can thus be compactly represented as $\{P_A\}_{A \in N}$. On a mathematical level, $\{P_A\}_{A \in N}$ thus consists of a family of families of probability measures. This is a complex object, but the complexity is inevitable if we want to model situations with multiple agents that may know multiple things.

In many applications, an underlying probability distribution P is given in advance, often in the form a Bayesian generative model, and the individual knowledge representations P_A are then derived from P by means of announcements. All relevant aspect's of an agent's state of knowledge are thus expressed by listing the random variables that have been announced to that agent. All models in this chapter are of this kind.

Announcing the value of a random variable X to an agent A is just another way of saying that we define a new Kripke frame $P_A(\,\cdot\,) = P(\,\cdot \mid X)$. If all A's information is summarized in the variable X, then this family of probability measures describes A's knowledge after the announcement. By equipping a cast of characters with such Kripke frames, we can construct models of social situations that allow reasoning about nested beliefs of arbitrarily high order.

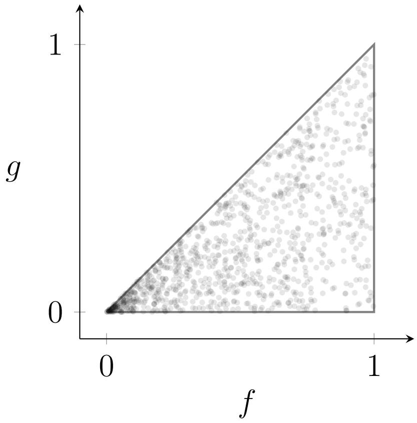


Figure 6.3: A thousand samples from the joint distribution defined by the generative model described in the main text.

Suppose for instance we define a probability measure P on $[0,1] \times [0,1]$ in terms of the following generative model (cf. Fig. 6.3):

- Draw F according to a uniform distribution on $[0,1]$.
- Draw G according to a uniform distribution on $[0,F]$.

This little story defines a sample space Ω and a probability distribution P. Once this probability distribution has been defined, we can go on to define the knowledge structures we are interested in by making the relevant announcements:

- Announce F to agent S; that is, define $P_S(\,\cdot\,) = P_S(\,\cdot \,|\, F) = P(\,\cdot \,|\, F)$.
- Announce G to agent T; that is, define $P_T(\,\cdot\,) = P_T(\,\cdot \,|\, F) = P(\,\cdot \,|\, G)$.

This defines two Kripke frames, or families of probability measures. Since agent S knows the value of F, he can distinguish worlds that have different f-coordinates, but is unable to distinguish worlds with the same f-coordinate and different g-coordinates. His conditional distribution at a world $\omega = (f,g)$ is thus the version of P that he obtains by conditioning on the horizontal slice containing ω. Agent T is in a similar situation, but with one conditional distribution for each of the vertical slices defined by a shared g-coordinate.

We thus have a generative model which defines a probability measure P and a set of announcements that define two knowledge states P_S and P_T. Having these two components in place, we can now compute both conventional probabilities and higher-order probabilities by integrating over the relevant regions of the sample space.

For instance, in the possible world $\omega = (F, G) = (f, g)$, we have

$$\begin{aligned}
P_S\left\{F < \frac{1}{2}\right\} &= \begin{cases} 1 & \text{if } f < 1/2 \\ 0 & \text{if } f \geq 1/2 \end{cases} \\
P_S\left\{G < \frac{1}{2}\right\} &= \min\left\{1, \frac{1}{2f}\right\} \\
P_T\left\{G < \frac{1}{2}\right\} &= \begin{cases} 1 & \text{if } g < 1/2 \\ 0 & \text{if } g \geq 1/2 \end{cases} \\
P_T\left\{F < \frac{1}{2}\right\} &= \max\left\{0, 1 + \frac{\ln 2}{\ln g}\right\}
\end{aligned}$$

Using the formula for $P_T\{F < 1/2\}$, agent S can also derive the probability

$$P_S\left\{P_T\left\{F < \frac{1}{2}\right\} > \frac{1}{2}\right\} = \frac{1}{4} + \frac{1}{4}\ln 4.$$

From the point of view of agent S, the probability $P_T\{F < 1/2\}$ is not a fixed number, but a random variable whose value depends on which world we are in. By the same token, the proposition $P_T\{F < 1/2\} > 1/2$ is true or false depending on which world we are in. However, since S has a probability distribution over these worlds, he can compute the probability of $P_T\{F < 1/2\} > 1/2$ by measuring the set of ω for which it is true.

This is the sense in which multi-agent probability models allow reasoning about reasoning. In the following two sections, I will discuss two examples that demonstrate how it can be used as a tool for modeling social cognition.

6.3 Example 1: Preference Attribution

As a first example, I want to present multi-agent probability model that captures the essential aspects of the task discussed by Baker et al. (2011). This is a relatively simple inference problem and will therefore be useful as a starting point.

6.3.1 The Situation

In the original experiment by Baker et al. (2011), subjects watched a little video clip showing tiny cartoon character moving around on a stylized map.

The map showed two parking lots where food trucks can park. Between these parking lots, a large building obscured the character's view, so that in order to see what food truck was parked at the other end of the map, he would have to walk around the structure.

After observing the little video clip of the character walking around the map, presented as a student walking around a university campus, the human subject

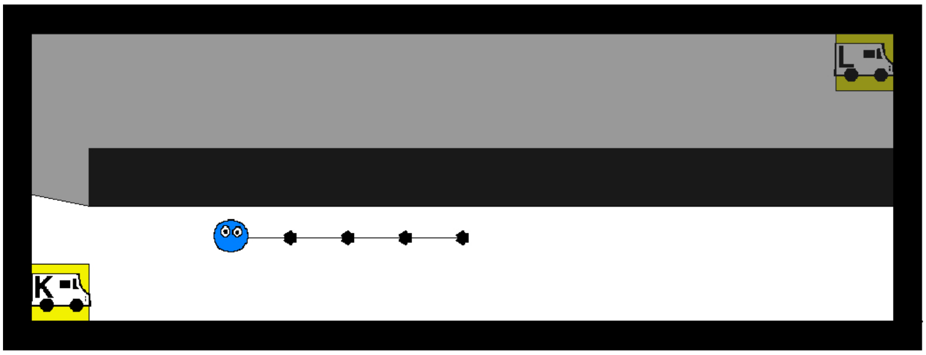


Figure 6.4: A frame from one of the videos that Baker et al. (2011, Fig. 2) used to investigate how people attribute preferences to artificial agents.

was asked to guess the beliefs and food preferences of the little video game character. Such guesses are possible because not all attributions of belief and desire to the "student" explain his observable behavior equally well.

Baker et al. described informally how this might work if the character first sees a food truck labeled "K," then checks to see whether the other parking lot contains an "L" or an "M" truck, and then eventually opts for the "K" truck:

> Because the student can see K, they know that the spot behind the building either holds L, M, or is empty. By frame 10, the student has passed K, indicating that they either want L or M (or both), and believe that their desired truck is likely to be behind the building (or else they would have gone straight to K under the principle of rational action). After frame 10, the agent discovers that L is behind the building and turns back to K. Obviously, the agent prefers K to L, but more subtly, it also seems likely that the agent wants M more than either K or L, despite M being absent from the scene! (Baker et al., 2011, p. 2471)

The model used by Baker et al. included a dynamic programming algorithm that could infer where the character was going based on the path traversed so far. However, it is clear from their description that the inherently social aspects of the situation can be described exhaustively by simply specifying where the agent looked, and what he subsequently choose to do. We can therefore abstract away from the spatial aspects of the task without too much loss of essential detail.

6.3.2 A Multi-Agent Model

Under the simplifying assumptions introduced above, the situation described by Baker et al. (2011) could be generated by a generative model of the following kind:

- The parking lots F_0 and F_1 are assigned values from $\{k, l, m\}$ at random.
- The character's preferences for foods k, l, and m are ordered at random by distributing the numbers 1, 2, and 3 onto the variables U_k, U_l, and U_m.

The values k, l, or m here represent the three types of food. Since there are nine different combinations, each combination has probability $P(F_0, F_1) = {}^1\!/_9$.

The number U_f represents the "utility" of the food f to the character C. There are six possible permutations of three numbers, so each joint assignment of utilities has probability $P(U_k, U_l, U_m) = {}^1\!/_6$.

The sample space Ω thus consists of vectors of the form $(f_0, f_1, u_k, u_l, u_m)$, and there are $9 \times 6 = 54$ possible worlds in the model. In each of these worlds, we can further compute the value of the following variables, which are conditionally deterministic given a possible world:

- $L = \mathbb{I}\,(U_{F_0} < 3)$, the event of F_0 not being C's most preferred food.
- $G = \mathbb{I}\left(P_C(U_{F_1} > U_{F_0}) > \frac{1}{2}\right)$, the event of C choosing the food F_1.

We can then describe the epistemic events of the story in terms of the following announcements:

- F_0 and F_1 are announced to the subject S.
- U_k, U_l, and U_m are announced to C.
- F_0 is announced to the character C.
- If $L = 1$, then F_1 is announced to C
- L and G are announced to S.

In informal terms, these announcements state that C chooses the best food he can get, and that he only checks the parking lot F_1 if F_0 doesn't already have his favorite food.

The announcements also cut up the 54 possible worlds in a number of information cells. They do so, however, in different ways for C and S.

6.3.3 Features of the Kripke Frames

Agent C knows his own preferences as well as the value of F_0, but he will not always be able to distinguish possible worlds that differ only on F_1. This situation can be illustrated if we imagine the sample space as a large table with the vectors (U_k, U_l, U_m) in the rows and the vectors (F_0, F_1) in the columns; then the row with $(U_k, U_l, U_m) = (1, 3, 2)$ will be split in the following way:

(k,k)	(k,l)	(k,m)	(l,k) (l,l) (l,m)	(m,k)	(m,l)	(m,m)

The fact that the worlds in the middle are indistinguishable reflect the fact that the value of F_1 is unknown to C in this row, since $U_{F_0} = U_l = 3$. Conditional on the announcements that the agent received, C's conditional Kripke frame will thus contain three information cells of size 3 ($U_{F_0} = 3$) and 45 information cells of size 1 ($U_{F_0} < 3$).

Agent S know both F_0 and F_1, but she will not always be able to distinguish worlds that differ on U_k, U_l, and U_m. However, since she knows the values of L and G, she can distinguish different preference attributions indirectly, although at a more coarse-grained level than C.

In order to see this more clearly, we need to go through the four possible values of (L, G) one by one. To make this discussion more concrete, I will assume that $(F_0, F_1) = (k, m)$, but the other cases are symmetrical.

We now have the following divisions:

- The case $(L, G) = (0, 0)$ represents the event of C choosing truck 0 without looking at F_1. Since $F_0 = k$, this implies that $U_k = 3$, but otherwise, nothing is revealed about the relative preferences of U_l and U_m.

- The case $(L, G) = (0, 1)$ represents the event of C choosing F_1 without looking at it first. By construction, this case has probability 0.

- The case $(L, G) = (1, 0)$ represents the event of C looking at F_1, but then choosing F_0. From $L = 1$, S can infer that $U_k < 3$, and from $G = 1$, she can infer that $U_k > U_m$. It follows that the only option is $(U_k, U_l, U_m) = (2, 3, 1)$.

- The case $(L, G) = (1, 1)$ represents the event of C looking at F_1 and then choosing F_1 over F_0. This implies that $U_k < 3$ and $U_k < U_m$, but leaves three options open.

To summarize these inferences, we can turn our imaginary table on its side, depicting the column $(F_0, F_1) = (k, m)$ as follows:

$(1,2,3)$ $(1,3,2)$ $(2,1,3)$	$(2,3,1)$	$(3,1,2)$ $(3,2,1)$

The other columns in S's conditional Kripke frame are also divided into three cells of sizes 1, 2, and 3, but the contents are different.

6.3.4 Quantitative Results

Using these observations, we can now compute conditional probabilities on behalf of the subject S, such as

$$\begin{aligned}
P_S\left(U_m = 1 \,|\, F_0 = k, F_1 = m, L = G = 1\right) &= 0\\
P_S\left(U_m = 2 \,|\, F_0 = k, F_1 = m, L = G = 1\right) &= {}^1\!/_3\\
P_S\left(U_m = 3 \,|\, F_0 = k, F_1 = m, L = G = 1\right) &= {}^2\!/_3\\
P_S\left(U_m = 1 \,|\, F_0 = k, F_1 = m, L = 1, G = 0\right) &= 1\\
P_S\left(U_m = 2 \,|\, F_0 = k, F_1 = m, L = 1, G = 0\right) &= 0\\
P_S\left(U_m = 3 \,|\, F_0 = k, F_1 = m, L = 1, G = 0\right) &= 0
\end{aligned}$$

We can also compute higher-order probabilities, such as S's beliefs about C's beliefs, both before and after the conditional announcement of F_1 (which is here named LF_1):

$$\begin{aligned}
P_S\left(P_C\left(F_1 = m \,|\, F_0\right) \,|\, F_0 = k, F_1 = m, L = G = 1\right) &= {}^1\!/_3\\
P_S\left(P_C\left(F_1 = m \,|\, F_0, LF_1\right) \,|\, F_0 = k, F_1 = m, L = G = 1\right) &= 1\\
P_S\left(P_C\left(F_1 = m \,|\, F_0, LF_1\right) \,|\, F_0 = k, F_1 = m, L = G = 0\right) &= {}^1\!/_3
\end{aligned}$$

These higher-order probability assignments thus replicate in a systematic fashion the computations that Baker et al. (2011) performed using methods that were specific to the situation at hand.

6.4 Example 2: The Sally-Anne Task

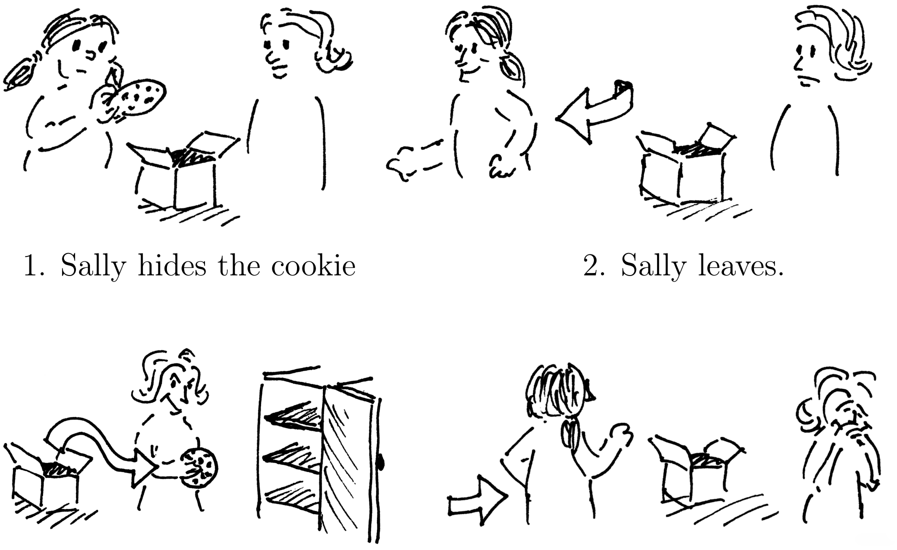
The Sally-Anne task is a type of false belief task which is frequently used to test children's cognitive development. The test involves telling a child a story about two characters and then asking a child about the beliefs of the character with an information deficit.

In this section, I will design a multi-agent probability model that replicates certain aspects of children's behavior on this task, building on a similar probabilistic model proposed by Goodman et al. (2006).

6.4.1 The Situation

The blueprint for this task was developed by Baron-Cohen et al. (1985), who used it to investigate the effects of autism. They described their experimental procedure as follows (cf. Fig. 6.6):

> There were two doll protagonists, Sally and Anne. ... Sally first placed a marble into her basket. Then she left the scene, and the marble was transferred by Anne and hidden in her box. Then, when



1. Sally hides the cookie

2. Sally leaves.

3. Anne moves the cookie

4. Sally comes back.

Figure 6.5: An outline of the Sally-Anne story.

> Sally returned, the experimenter asked the critical Belief Question: "Where will Sally look for her marble?". If the children point to the previous location of the marble, then they pass the Belief Question by appreciating the doll's now false belief. If however, they point to the marble's current location, then they fail the question by not taking into account the doll's belief. (Baron-Cohen et al., 1985, p. 41)

A now very rich literature on tasks of this sort has consistently documented that children up until age of about three or four tend to give systematically wrong answers to this "Belief Question" (Wellman et al., 2001). Rather than recognizing that Sally has a false belief, the children report that she will look in the box, seemingly conflating Sally's beliefs with reality as presented in the story. Similarly, when children are asked why Sally looked in the box rather than the basket, younger children tend to answer that she wasn't actually looking for the marble, while the older children understand that she did so because of a false belief.

In order to model these findings, Goodman et al. (2006) constructed a Bayesian model which encompassed two different methods of reasoning, a "mind-reading" submodel with no separate representation of beliefs, and a "perspective-taking" submodel which kept belief and reality apart. The responses given by young or autistic children then corresponded to a "mind-reading" theory, while the responses of older children corresponded to a "perspective-taking" theory.

However, the perspective-taking model makes more accurate predictions. A Bayesian learner considering both options will thus eventually favor the perspective-taking model in spite of its higher complexity. The transition from one kind of answer to the other can thus be explained, if this Bayesian model is correct, as a matter of experience rather than, say, the execution of a fixed developmental script.

The model used by Goodman et al. (2006) represented Sally's two possible beliefs by switching a binary "belief variable" on or off. This suffices for this very particular situation, but it obviously generalizes very poorly to other multi-agent scenarios. In the following, I will therefore explain how to express these matters of social cognition in a more principled fashion.

6.4.2 A One-Shot Sally-Anne Trial

My goal here is to explain how a child might come to learn how to replace a "mind-reading" model by a "perspective-taking" model. This will require the use of several rounds of observation, with information from the previous rounds feeding into the next. However, for presentational purposes, I will start by discussing how to model a single trial and then only consider experiments with multiple trials later.

One way of capturing the facts presented in Sally-Anne narrative is given by the following generative model:

- A coin M with parameter m is flipped to decide where Anne hides the marble.

- A coin K with parameter k is flipped to decide whether Sally will spuriously learn this fact in spite of having been absent while Anne hid the marble.

- A coin W with parameter w is flipped to decide whether Sally wants to find her marble when she comes back.

- A coin R with parameter $\frac{1}{2}$ is flipped to decide what Sally will do if she is not looking for the marble.

These four coin flips define a probability distribution P on a sample space Ω with 16 possible worlds. The intended reading is here that m is a number close to 0, since the marble is unlikely to spontaneously change position. k, however, might be very high initially, since we are modeling a young child with a tendency to use the simpler "mind-reading" model.

Now that we have introduced all the randomness we need, we can define a couple of useful auxiliary variables:

- $\hat{M} = \mathbb{I}\left(P_S(M = 1) > \frac{1}{2}\right)$, Sally's best guess as to where the marble is.

- $L = W\hat{M} + (1 - W)R$, the observable behavior that Sally performs.

The relevant Kripke frames are derived from the distribution P be means of the following announcements:

- M and L are announced to Anne.
- K, W, and R are announced to Sally.
- If $K = 1$, then M is announced to Sally.

Sally can thus distinguish most worlds, confusing only the four pairs that have $K = 1$ and two different values of M. Her conditional Kripke frame consequently splits the sample space up into 12 information cells, Four of which consist of 2 possible worlds, and the rest contain 1.

6.4.3 Inference in a Single Trial

Anne's conditional Kripke frame partitions the sample space according to M and L. As will become apparent below, this means that she splits up Ω into four cells of sizes 6, 2, 4, and 4, respectively.

At the risk of being overly explicit, I have included a complete table of this probability distribution in Table 6.1. The parameter values used in this table were

$$m = \frac{1}{11}, \quad k = \frac{10}{11}, \quad w = \frac{5}{6}.$$

This table can also be used to check which possible worlds Anne can distinguish: Since she knows only M and L, she will be able to tell two rows apart if they have different values on at least one of these two variables. The probabilities P_A are computed by renormalizing each such information cell.

We can also specify these knowledge states explicitly. This amounts to listing the information cells that correspond to Anne's and Sally's conditional Kripke frames. In terms of the row numbers in the table, these information cells are

$$\begin{array}{ll} A: & \{1,9\}, \{2,10\}, \{3,11\}, \{4,12\}, \{5\}, \{6\}, \{7\}, \{8\}, \{13\}, \{14\}, \{14\}, \{16\} \\ S: & \{1,3,4,5,7\}, \{2,6\}, \{9,11,12,13\}, \{10,14,15,16\} \end{array}$$

Using this table, we can compute higher-order probabilities such as

$$P_A\left(P_S(M = 1) > \frac{1}{2} \;\middle|\; M = 1, L = 1\right) \;=\; .092 + .447 + .447 \;=\; .986$$

This number thus represents the probability that Anne assigns, after observing that Sally looked in the right place, to the hypothesis that she did so because she already knew that it had moved. Due to the high value of k, this explanation is overwhelmingly likely at this point.

ω	M	K	W	R	P	P_S	P_A	$\hat{M}$	L
1	0	0	0	0	.007	.909	.008	0	0
2	0	0	0	1	.007	.909	.092	0	1
3	0	0	1	0	.034	.909	.041	0	0
4	0	0	1	1	.034	.909	.041	0	0
5	0	1	0	0	.069	1	.083	0	0
6	0	1	0	1	.069	1	.908	0	1
7	0	1	1	0	.344	1	.413	0	0
8	0	1	1	1	.344	1	.413	0	0
9	1	0	0	0	.001	.091	.071	0	0
10	1	0	0	1	.001	.091	.013	0	1
11	1	0	1	0	.003	.091	.214	0	0
12	1	0	1	1	.003	.091	.214	0	0
13	1	1	0	0	.007	1	.500	1	0
14	1	1	0	1	.007	1	.092	1	1
15	1	1	1	0	.034	1	.447	1	1
16	1	1	1	1	.034	1	.447	1	1

Table 6.1: The probability of each possible world according to P, P_S, and P_A.

6.4.4 Multiple Sally-Anne Trials

According to the model presented above, a single run of a Sally-Anne trial corresponds to the simultaneous sampling of a vector (M, K, W, R) from four coin flipping distributions with parameters $(m, k, w, {}^1\!/_2) = ({}^1\!/_{11}, {}^{10}\!/_{11}, {}^5\!/_6, {}^1\!/_2)$. In this subsection, I will assume that the parameters m, k, and w are themselves random variables. This will allow us to model the process of learning by means of a hierarchical Bayesian model.

This requires a couple of notational complications. First, we need to create a family of variables

$$(M_i, K_i, W_i, R_i), \qquad i = 1, 2, \ldots, n,$$

each sampled from to the same distribution with the shared parameter vector $(m, k, w, {}^1\!/_2)$. Second, it will also improve clarity if we create n copies of Sally and Anne, $S_1, S_2, \ldots, S_n$ and $A_1, A_2, \ldots, A_n$, since this will underscore the fact that we imagine them having different amounts of information at different times.

The hierarchical model can thus be defined by the following generative story:

- $m \sim \text{Beta}(10, 1)$
- $k \sim \text{Beta}(1, 10)$
- $w \sim \text{Beta}(1, 5)$
- For $i = 1, 2, \ldots, n$:
  - $M_i \sim \text{Bernoulli}(m)$
  - $K_i \sim \text{Bernoulli}(k)$
  - $W_i \sim \text{Bernoulli}(w)$
  - $R_i \sim \text{Bernoulli}({}^1\!/_2)$

The shorthand $X \sim \text{Beta}(\beta_0, \beta_1)$ here means that the variable X is sampled from a Beta distribution with parameters β_0 and β_1, that is, the probability distribution with densities proportional to $p(x) = x^{\beta_0}(1-x)^{\beta_1}$ on the unit interval. Similarly, $X \sim \text{Bernoulli}(p)$ means that X is a Bernoulli distribution, that is, a coin flip, with $P(X = 1) = p$ and $P(X = 0) = 1 - p$. The hyperparameters used in this model were selected so as to agree with those used by Goodman et al. (2006).

The announcements we need for this model are as follows:

- For $i = 1, 2, \ldots, n$:
  - M_i and L_i are announced to A_i.
  - K_i, W_i, and R_i are announced to S_i.
  - If $K_i = 1$, then M_i is announced to S_i.

- For $j = 1, 2, \ldots, i-1$:
  - M_j and L_j are announced to A_j and S_j.

I thus assume that both agents are informed of where the marble actually was in the previous trial. This means that Sally generally has a good, empirically informed idea about the value of m, and Anne can pool her observations of Sally's past behavior to make inferences about the likely values of k and w.

6.4.5 Learning Across Trials

Based on this generative model, we can now ask start asking questions about Anne's posterior distribution over k and w, the two parameters that determine Sally's behavior. The exact shape of this posterior distribution at time $n+1$ will of course depend on the observations Anne made in the first n rounds. I have therefore generated a synthetic data set that might reflect the typical content of a child's experience.

The data set contains 200 rounds and was generated randomly according to the generative story above, with the hyperparameters set to the values

$$m = \frac{1}{11}, \quad w = \frac{10}{11}, \quad k = \frac{1}{20}.$$

The values of m and w were thus typical of the distributions they were drawn from. The value of k, however was deliberately set much below expectation in order to simulate a realistic environment without mind-readers as opposed to an environment conforming to Anne's prior expectations. In the actual sample used below, $M = 1$ occurred 19 times, $W = 1$ occurred 176 times, and $K = 1$ occurred 11 times.

The changes in Anne's posterior distribution over w and k are determined by two factors:

- the prior probability that Anne assigned to various choices of w and k before observing the data;
- the likelihood of the data given any such a combination of values.

The prior probability distribution is a product of two Beta distributions. According to P and therefore also P_A, it has a peak around the point $(5/6, 10/11)$. The likelihood, however, has two be extracted from the data by computing how probable Sally's observable behavior is given a specific setting of w and k.

Fortunately, it is relatively easy to compute these conditional probabilities once the parameters w and k are fixed. For any pair of values for these parameters, there are only 16 distinct cases to consider in each round, and each of these 16 cases has a specific probability.

The likelihood of the ith observation L_i is thus the sum of the probabilities, one probability for each case consistent with Sally's behavior. The likelihood of these entire data set is the product of the likelihoods for the individual rounds. The plot in Figure 6.6 shows what this function of w and k looks given the 200 first observations.

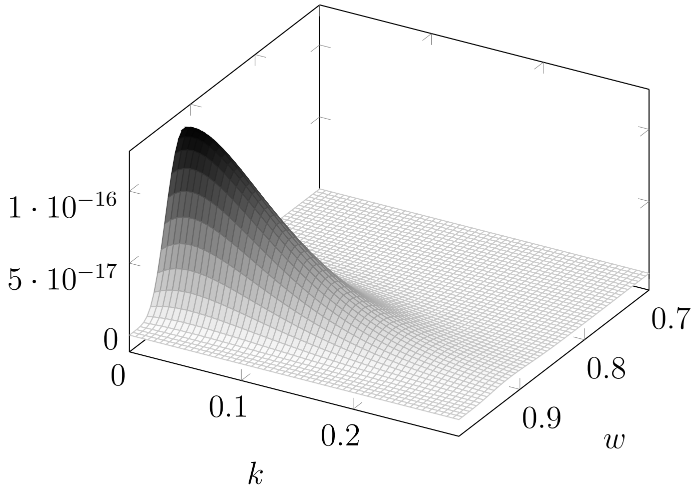


Figure 6.6: The likelihood of the entire 200-item data set as a function of the hyperparameters w and k.

As the plot shows, the data induces a quite distinct peak in the likelihood function close to the actual parameter vector,

$$(w, k) \;=\; \left(\frac{10}{11}, \frac{1}{20}\right).$$

For a sufficiently flat prior, this data thus provides sufficient evidence for Anne to learn the approximate value of these two hyperparameters w and k based on Sally's observable behavior.

This process of learning can also be followed as it progresses through time. Figure 6.7 shows how three tell-tale tendencies, $E_{A_i}[w]$, $E_{A_i}[k]$, and $E_{S_i}[m]$, change their values after the observation of $i = 0, 5, 10, 15, \ldots, 200$ data points. As this plot shows, Anne's estimate of k gradually descents from its initial value of $k = {}^{10}\!/_{11}$ to something closer to its true value of $k = {}^{1}\!/_{20}$.

Based on this string of observations, Anne thus gradually transitions from a model that predicts Sally to be a mind-reader to a model that does not. The tipping point lies around $i = 100$, which is what we should expect given the assumptions: It takes 10 observations of $K = 0$ to overcome the initial bias in k, and the marble only moves ($M = 1$) about 1 in 10 times.

Notice also that the data does not give Anne any reason to revise her estimate of w: All Sally's observable behavior is explained away by assuming a low value of k, and no change in w is therefore needed (Pearl, 1988). This effect is what we

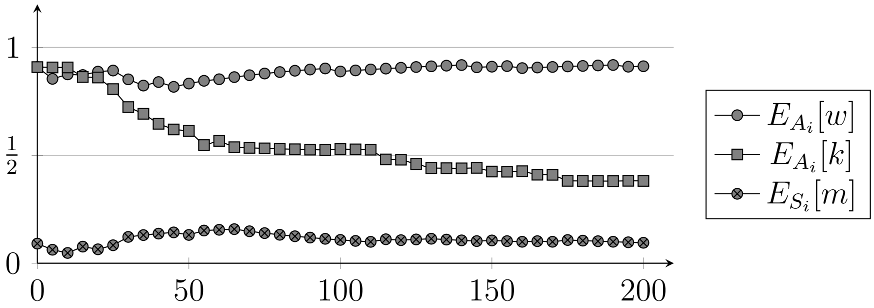


Figure 6.7: The development in the posterior means of w and k according to A's probability distribution, and in m according to S's probability distribution.

should expect, since a "no preference" explanation of the observation $(M, L) = (1, 0)$ predicts that Sally will act randomly, while a "no knowledge" explanation makes the more precise prediction that she looks deterministically at $L = 0$. Since the more precise prediction happens to be true, it is eventually favored by the data.

6.5 Conclusion

Multi-agent probability theory provides a flexible and very systematic way of constructing of models of social cognition. This chapter has discussed two case studies that demonstrate in detail how to formulate and reason with such models. In both examples, we saw how one agent could use the observable behavior of another agent to make inferences about the possible preferences and beliefs of that agent. These examples could quite easily be modified or extended so as to accommodate more agents or more deeply nested knowledge.

An obvious next step in the refinement of these models would be to include a more explicit and general story about how the agents choose their actions. In the examples above, I simple defined random variables that postulated certain behaviors under explicitly listed conditions.

A more principled solution would be to define a decision rule that systematically maps utility functions and a probability distributions over states onto probability distributions over actions. Such a rule could select actions leading to maximum utility, minimax risk, minimum regret, or any other well-defined success criterion from the literature on decision theory. In more realistic models with a larger space of possible actions, such principles are unavoidable.

I have deliberately refrained from evaluating the empirical adequacy of any of the models I discussed in this chapter. However, in this conclusion, it would be appropriate to at least mention the question of what these models actually reflect, if anything. Probability theory is an incredibly flexible formalism, and we can model practically any human behavior if we are promiscuous enough about our assumptions.

On the negative side, this should be cause for some concern. If all the predictive power really resides in the selection of the appropriate model, then what cognitive science should provide is a principled story about how human beings construct their models, rather than a story about how they draw their conclusions once the right model is around.

In a probabilistic framework, this issue will often be a matter of explaining where the sample space came from: When do I interpret my observations as hard knowledge, and when do I interpret them as noisy evidence? Before any inference can get off the ground, we need to have source of hard information; and considered over time, the rate at which we receive such information should be kept at about the same level as the rate at which we add additional dimensions to our model of the world.

On the positive side, however, we already know what the best response to this threat is: The cure against a proliferation of overfitted models is to be consistent about formulating general principles that can be used to construct classes of models rather than inventing a new hack for every new problem. In this sense, the marriage of probability theory and dynamic logic may, in spite of the increase in expressivity it brings about, turn out to decrease rather than increase the number of models in the world.

Chapter 7

On the Consistency of Approximate Multi-Agent Probability Theory

Bayesian models have proven to accurately predict many aspects of human cognition, but they generally lack the resources to describe higher-order reasoning about other people's knowledge. Recently, a number of suggestions have been made as to how these social aspects of cognition might be codified in computational reasoning systems. This chapter examines one particularly ambitious attempt by Stuhlmüller and Goodman (2013), which was implemented in the stochastic programming language Church. This chapter translates their proposal into a more conventional probabilistic language, comparing it to an alternative system which models subjective probabilities as random variables. Having spelled out their ideas in these more familiar and intuitive terms, I argue that the approximate reasoning methods used in their system have certain statistical consistency problems.

7.1 Introduction

Probability theory is a useful and largely reasonable model of many aspects of human reasoning and decision-making. However, by its nature, it is geared towards reasoning about a static, inanimate environment, and it is thus not very well equipped to handle reasoning about mutually adapting agents in inherently social situations such as game-playing, war, bargaining, and conversation (Blackwell and Girshick, 1954; Luce and Raiffa, 1957).

Consider for instance the following example:

Example 7.1. Ann and Bob both draw five cards from a standard deck of 52 cards with 4 aces.

1. What is the probability that Bob draws exactly one ace?

2. Given that Ann has two aces on her hand, what probability will she assign to Bob having exactly one ace? Is this number less than 25%?

3. Given that Bob has exactly three aces on his hand, what probability will he assign to the event that Ann assigns less than 25% probability to him having exactly one ace?

Although these questions are somewhat cumbersome to formulate, they have the look and feel of math problems with well-defined answers. For instance, we could reason that if Ann has two aces on her hand, she knows that there are only two aces left, and this will influence her private beliefs about Bob's hand. Her uncertainty can therefore be expressed in the form of a conditional probability distribution, and Bob can reason perfectly rigorously about what the relevant condition might be, and what the corresponding distribution might look like.

The goal of multi-agent probability theory, whatever the details of style and implementation, is to formalize inferences such as these. Such models have often played an important background role in game theory (Aumann, 1976; Harsanyi, 1967) and artificial intelligence (Fagin and Halpern, 1994).

In a recent article, Stuhlmüller and Goodman (2013) have proposed a different approach to this problem. The system they propose is implemented in the stochastic programming language Church (Goodman et al., 2008) and models probability distributions over probability distributions in terms of sampling schemes that sample other sampling schemes.

The purpose of this chapter is to evaluate to what extent this approach is equivalent to a more conventional probabilistic model that defines multi-agent probability theory in terms of probability distributions rather than in terms of individual samples. The conclusion I reach is that there are some rather serious consistency issues separating the two.

However, before I can present these ideas, I will first need to briefly discuss the representation of distributions and probabilities in Church. I will then, for reference, introduce another formalism that will allow us to reason about multi-agent probabilities directly. Having presented these two alternatives, I will proceed to point out the differences between them.

7.2 Probabilistic Reasoning in Church

Church is a stochastic dialect of Lisp which allows the user to describe a probability distribution in the form of a random program. For instance, if we want to define a generative model that first selects a number θ randomly from the unit interval, then flips a coin with bias θ, we might write

```
(define theta (uniform 0 1))
(define x (flip theta))
```

This program defines a joint probability distribution over the pairs (θ, x) in the space $\Omega = [0,1] \times \{0,1\}$.

Once this distribution has been defined in Church, a single sample point can be obtained executing the program once (i.e., feeding it to the equivalent of Lisp's `eval` function). Larger samples are often more efficiently obtained by using one of the several sampling schemes implemented in Church under the name `query`. These functions use techniques such as Markov Chain Monte Carlo sampling to simulate the effect of running a program repeatedly, sometimes discarding samples that fail to meet some condition.

Church thus allows one to sample from any probability distribution that one can build out of the primitive randomization devices in the language (such as coin flips and uniform distributions). Since the deterministic fragment of the language is Turing-complete, this in principle allows the user to simulate any computable probability distribution by constructing a random program that lies arbitrarily close to it.

7.2.1 Probability and Frequency

Church is not a tool for probability theory as such. The language has no internal representation of probabilities, expectations, or distributions, and it does not permit any direct reasoning about such quantities inside the stochastic programs.

This means that we cannot evaluate the probability of a given event directly in Church. If we want to compute the probability of some event $R \subseteq \Omega$, the only generally available option is to use one of the variants of the `query` function to obtain a sample

$$x \;=\; (x_1, x_2, \ldots, x_n) \;\in\; \Omega^n,$$

and then approximate the probability of R by its empirical frequency in this sample. This corresponds to approximating the underlying probability distribution P by the empirical frequency distribution

$$\tilde{P}(R) \;=\; \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}_R(x_i),$$

where $\mathbb{I}_R$ is the indicator function of the event R. By plugging in $\tilde{P}$ in place of P in the relevant places, we can use this approximation to compute functions of P, such as probabilities and expectations (Wasserman, 2006, Ch. 2).

Church comes with fidelity guarantees in the sense that a value returned by `query` indeed is a sample from the distribution P. However, it does not come with consistency guarantees in the sense that any statistic $T(P)$ is consistently estimated by $T(\tilde{P})$. As I will argue later in this chapter, this does not pose any problems for the estimation of probabilities in a single-agent context, but this picture changes radically when we instead consider multi-agent probability theory, even with a single extra layer of reasoning.

7.2.2 Multi-Agent Reasoning

In order to build a system of multi-agent reasoning using the resources in Church, Stuhlmüller and Goodman propose using several different query functions nested into each other as a representation of multi-agent reasoning. This way, an agent's actions can depend on a hypothetical execution of another agent's actions.

Consider for instance a two-player game between two agents, Ann and Bob, which choose their actions simultaneously. Ann might be interested in choosing a move which is good given what Bob might do, so she needs to perform some kind of mental simulation of what he might choose. In very rough pseudocode, the query function describing her actions could therefore be defined as

```
define query-Ann():
   Bob's actions = query-Bob()
   respond stochastically to Bob's actions
```

Similarly, Bob's choice of actions could be follow a probability distribution that depended on his hypothetical simulation of Ann's choice:

```
define query-Bob():
   Ann's actions = query-Ann()
   respond stochastically to Ann's actions
```

Calling either of these query functions would then unfold an infinitely deep recursion tree in which Ann's choice depends on Bob's, which in turn depends on Ann's, which depends on Bob's, and so forth. In order for this recursion to eventually halt, Stuhlmüller and Goodman add a maximum depth parameter to their examples and replace the mutually dependent query functions by an unconditional prior sampling function below this threshold.

While this system does represent a certain calculus of mutually dependent behavior, it would be a bit of a stretch to say that it encodes "reasoning about reasoning" as such. It rather represents a method for sampling a response from a distribution whose shape may depend on a sample of hypothetical stimuli. In many respects, this makes the model more behavioral than cognitive.

This also means that the system has no clean separation of its epistemic component (reasoning about reasoning), and its decision-theoretic component (mapping beliefs to actions). This makes it hard to evaluate the rationality of any specific distributions over actions, or to provide a priori reasons for favoring any specific family of stimulus-to-response mappings.

However, it is in fact possible to modify the system slightly so that its epistemic component is isolated for independent use. As indicated in the previous section, Church can represent probability distributions internally by approximating them with frequency distributions. This means that we can in fact use a nested query approach in the style of Stuhlmüller and Goodman to decide a question of higher-order probabilistic beliefs: In order to do so, we need to define a set of frequency distributions over frequency distributions, using those nested samples to estimate the probability that a given probabilistic statement obtains.

To make this more concrete, suppose that Bob has three aces on his hand. He then has a conditional distribution over the two possible location of the remaining ace, either in the deck or on Ann's hand. This distribution can be represented by a `query` function, and we can draw samples from that function to estimate probabilities related to that distribution.

In each of the two cases Bob considers, Ann will possess certain information about her own hand and thus have a conditional distribution of her own representing her beliefs about the number of aces Bob holds. This distribution too can be represented as a `query` function, and she can draw samples from this function to decide, say, whether the conditional probability that Bob holds three aces is less than 25%.

However, Bob can use his own `query` function to estimate the probability of the various conditions Ann might be in. He can thus estimate the probability that Ann's probability estimate is less than 25%. In this way, a set of nested `query` functions can thus be used to decide the truth of a higher-order proposition about probabilities.

This kind of epistemic reasoning is not explicitly the focus of Stuhlmüller and Goodman's approach. They leap directly from the observed behavior of others to one's own choice of response, passing quietly over the intermediate step of representing beliefs as probabilities. The estimation method described here is thus a compromise between the behavioral modeling favored by Stuhlmüller and Goodman and the more classical preoccupation with disinterested reasoning about the state of the world.

The kind of reasoning about frequencies described here is also somewhat cumbersome and unnatural to write out in Church. In the following section, I will therefore introduce a theoretical framework that will make it easier to think about such higher-order inferences, occasionally hinting at what the Church equivalents of these concepts are.

7.3 Multi-Agent Probability Theory

In order to make our reasoning about reasoning more systematic, we will need to introduce some theoretical concepts from the literature on multi-agent reasoning, particularly probabilistic epistemic logics (Fagin and Halpern, 1994; Baltag et al., 1998; van Benthem et al., 2009). A key insight from this literature is that we can think about the sample space Ω as a set of "possible worlds," or maximally informed states, and an increased knowledge level as an increased ability to distinguish different possible worlds (Kripke, 1963).

I will attempt to present these ideas in terms that bring them close in style to conventional probability theory. The presentation will be conceptually novel in that it explicitly represents probabilities as random variables. This internalizes subjective probabilities into the system, allowing us to compute and reason with them like any other random variables. This means that the modal fragment of the language reduces to a familiar and well-understood theoretical framework, contributing a significant conceptual clarity.

7.3.1 Definitions

In conventional probability theory for a single agent, inference is a matter of restricting and rescaling probability distributions.

Specifically, suppose the agent starts with a probability distribution P on a sample space Ω. After the event $S \subseteq \Omega$ happens, the agent then throws away the rest of the sample space and renormalizes:

$$P(R \,|\, S) \;=\; \frac{P(R \cap S)}{P(S)}.$$

The conditional probabilities $P(R \,|\, S)$ is thus the unconditional probability we get if we treat the condition $S \subseteq \Omega$ as a sample space in its own right.

In a multi-agent setting, we cannot represent conditions by subsets of the sample space, since this would conflate the internal and external perspectives on the model. If we really threw away the entire remainder of the sample space, $\Omega \setminus S$, we would have no way of representing mathematically that some agents might still assign a positive probability to $\Omega \setminus S$. To overcome this problem, we instead represent an agent's knowledge as a partition of the sample space into disjoint information cells. The coarseness of the partition then represents the granularity of information available to an agent. This will allow us to say that an agent learns *whether* something is the case without stating *that* it is the case.

Definition 7.1. A **conditional multi-agent probability space** is defined in terms of the following components:

1. a sample space Ω;
2. a probability distribution P on Ω;
3. a list of partitions of the sample space, $a, b, c, \dots$

The distribution P is interpreted as the prior probability distribution shared by all the agents. The partitions $a, b, c, \dots$ are interpreted both as a set of names for the agents and as a representation of their knowledge states.

By definition, a partition consists of mutually exclusive and collectively exhaustive classes, like tiles on a floor. A class can therefore be represented by any of its members. I will use the notation $[\omega]_a$ to denote the class of the partition a which contains the point $\omega \in \Omega$.

When a is interpreted as a knowledge state, its classes are called **information cells**. The information cell $[\omega]_a$ can be interpreted as the set of possible worlds that agent a considers possible when the actual world is in fact $\omega \in \Omega$.

Definition 7.2. Given an event $R \subseteq \Omega$, the **subjective probability** $P_a(R)$ is a random variable whose value at $\omega \in \Omega$ is

$$P_a(R)(\omega) \;=\; P(R \,|\, [\omega]_a).$$

We could similarly define conditional subjective probabilities as

$$P_a(R \,|\, S)(\omega) = P(R \,|\, S, [\omega]_a),$$

but these will not occur in this chapter.

For a fixed agent a and a fixed event R, the subjective probability $P_a(R)$ is thus a random variable. It can take different values in different regions of the sample space, but it has the same constant value inside each of a's information cells. In the cell $[\omega]_a$, this constant value represents the posterior probability a assigns to the event R given the information available to her.

This corresponds to the fact that when the actual world is ω, all that a knows is that we are somewhere in $[\omega]_a$. Since she cannot distinguish between the various possible worlds within this information cell, she cannot do better than condition on the entire event $[\omega]_a$. By contrast, an omniscient agent who could distinguish every possible world from every other would condition on the maximally informative event $\{\omega\}$, thus achieving a deterministic posterior distribution.

Note that any sample space comes with a trivial partition, $\{\Omega\}$. This partition represents the knowledge state of an agent who has no information at all. Such an agent will only know that vacuous statement that the actual world is a possible world, and the corresponding posterior distribution is thus P, the prior. This distribution can be identified with the perspective of an outside observer of the model.

The definitions above implicitly assume that the agents derive their posterior beliefs from a single shared prior, and that their beliefs are probabilistically coherent (de Finetti, 1937; Cox, 1946). These assumptions can be relaxed in various ways, as often discussed in the context of probabilistic logics (van Benthem et al., 2009). Such exotic variants are not relevant for my present purposes, however.

The definitions above also assume the existence of a sample space and a probability measure on that space without explaining how those objects might be constructed. In Bayesian statistics, it is common and very useful to specify the distribution over such a space in terms of a generative model that links together all the variables in the model in a non-circular network of conditional dependence constraints (Shannon, 1948; Good, 1980; Pearl, 1988). This idea is very important in practice and it is one of the most salient features of languages like Church, but it will not play any role in this chapter.

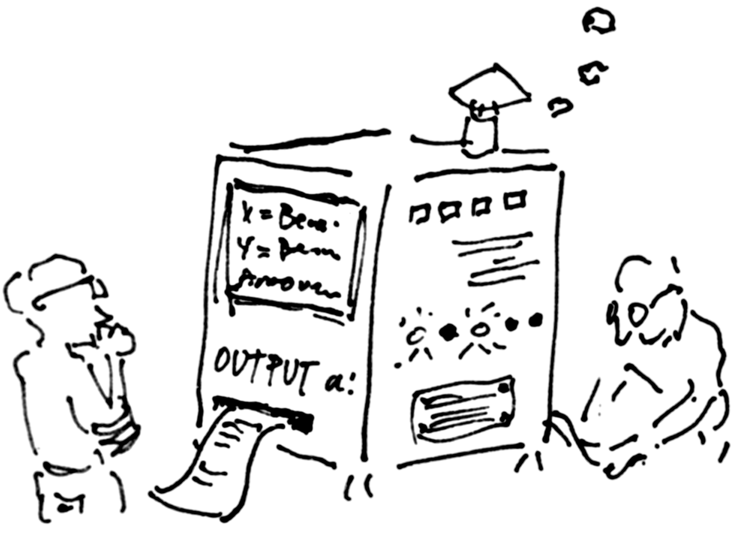


Figure 7.1: The calculus described in the text models a situation in which a machine randomly selects a possible world $\omega \in \Omega$ and then selectively prints different information to different people. However, the rules that govern how the machine works are known to everybody, and the agents can therefore reason rigorously about the knowledge states of each other.

7.3.2 Examples

Before returning to the consistency issues that are the topic of this paper, it will be useful to discuss a few simple examples. We start with a minimal toy example.

Example 7.2. An agent flips a coin and looks at the outcome.

This situation is described by the following model:

1. $\Omega = \{0, 1\}$;
2. $P\{0\} = P\{1\} = 1/2$;
3. $a = \{\{0\}, \{1\}\}$.

Since the agent looks at the coin, she is able to distinguish the two possible worlds in the sample space. The partition a which represents her state of knowledge consequently contains two information cells, $[0]_a = \{0\}$ and $[1]_a = \{1\}$.

Consider now the event $R = \{1\}$. The random variable $P_a\{1\}$ has different values in these two information cells. In the possible world $\omega = 1$, it has the value

$$P_a\{1\}(1) \;=\; P_a(\{1\} \,|\, [1]_a) \;=\; \frac{P_a(\{1\} \cap \{1\})}{P_a\{1\}} \;=\; 1.$$

In the possible world $\omega = 0$, on the other hand,

$$P_a(\{1\})(0) \;=\; P_a(\{1\}\,|\,[0]_a) \;=\; \frac{P_a(\{1\} \cap \{0\})}{P_a\{0\}} \;=\; 0.$$

These two probabilities codify the fact that when $\omega = 1$, the agent knows that $\omega = 1$, and when it isn't, she knows it isn't.

For an outside observer who doesn't know whether the coin came up heads or tails, we have

$$P(P_a\{1\} = 1) \;=\; P(\{1\}) \;=\; \frac{1}{2}.$$

There is thus 50% probability that a will know with certainty that $\omega = 1$.

This example can also be translated into a set of stochastic programs closer to Stuhlmüller and Goodman's idiom. In order to do so, we first need to define a program which represents the underlying prior:

```
define P():
   flip a coin to select w = 0, 1
   return w
```

We then need to define another program which models a's subjective beliefs. Since a possesses some information, this program will depend on which possible world we are in:

```
define P_a(w):
   until success:
      draw q according to P()
      if a can’t distinguish q and w:
         return q
```

If, as above, we wanted to estimate the probability that $P_a\{1\} = 1$, we could first sample a large number of worlds according to `P`; in each of those worlds, we could then estimate the probability that a sample from `P_a(w)` would be equal to `1`.

By construction, a large sample from `P` would contain roughly equally many `1`s and `0`s. A sample from `P_a(1)`, on the other hand, would consist solely of `1`s. A sample from `P_a(0)` would consist solely of `0`s.

Within the sampling error of a fair coin flip, an estimate of the higher-order probability

$$P(P_a\{1\} = 1) \;=\; \frac{1}{2}$$

would thus in this case be correctly estimated by the empirical frequency approximation, yielding

$$\tilde{P}(\tilde{P}_a\{1\} = 1) \;\approx\; \frac{1}{2}.$$

We now move on to a slightly more interesting example, formalizing the problem posed at the opening of this chapter.

Example 7.3. Assume that agents a and b both draw five cards from a standard deck and count the number of aces.

This situation is described by the following model:

1. The sample space is
$$\Omega = \{0,1,2,3,4\} \times \{0,1,2,3,4\}.$$
2. The prior probability distribution is given by
$$P\{(x,y)\} \; = \; h(x \,|\, 52, 4, 5)\, h(y \,|\, 52-5, 4-x, 5),$$
where $h(s \,|\, T, S, t)$ is the hypergeometric probability of finding s aces and $t-s$ non-aces in a sample of t cards from a population of T, of which S are aces:
$$h(s \,|\, T, S, t) \; = \; \frac{\begin{pmatrix} S \\ s \end{pmatrix} \begin{pmatrix} T-S \\ t-s \end{pmatrix}}{\begin{pmatrix} T \\ t \end{pmatrix}}$$
3. Assuming the agents only see their own hand, agent a's partition $a = \{[x,y]_a\}$ consists of five classes of the form
$$[x,y]_a \; = \; \{(x,0),(x,1),(x,2),(x,3),(x,4)\},$$
for $x = 0,1,2,3,4$. Agent b's partition consists of five classes of the form
$$[x,y]_b \; = \; \{(0,y),(1,y),(2,y),(3,y),(4,y)\}$$
for $y = 0,1,2,3,4$.

The prior probability distribution P is given in Table 7.1. Using this table, we can compute both conventional first-order probabilities and nested higher-order probabilities.

Suppose for instance that the actual world is
$$\omega = (x,y) = (3,1).$$

Then agent a knows that b has at most one ace on his hand, and this naturally changes her posterior distribution over the number of aces on b's hand. In formal terms, we can compute a's probability that $y = 1$ by conditioning on the information cell $[3,1]_a$:
$$P_a\{y=1\} \; = \; \frac{P(\{y=1\} \cap [3,1]_a)}{P([3,1]_a)}.$$

	0	1	2	3	4
0	.41345	.21202	.03180	.00155	.00002
1	.21202	.07951	.00776	.00018	0
2	.03180	.00776	.00037	0	0
3	.00155	.00018	0	0	0
4	.00002	0	0	0	0

Table 7.1: Prior probabilities, $P(x, y)$.

Inserting the relevant cells, this evaluates to

$$\frac{P\{(3,1)\}}{P\{(3,0),(3,1),\ldots,(3,4)\}} \;=\; 0.106.$$

Agent a will thus assign a probability of about 10.6% to the event that b should have the last remaining ace on this hand when she already holds three of them.

Note that this random variable would have other values in other information cells. For instance, if a had drawn 2 aces instead of 3, her probability that $y = 1$ would rise to about 19.4%, since there would then be one more ace in the deck for b to draw (cf. Table 7.2, left).

Consider now the random variable

$$P_b\{P_a(y = 1) < .25\}.$$

In the possible world $(x, y) = (3, 1)$, this variable has the value

$$P_b\{P_a(y = 1) < .25\}(3, 1),$$

which is equal to

$$\frac{P(\{P_a(y = 1) < .25\} \cap [3, 1]_b)}{P([3, 1]_b)}.$$

By the previous computation, we can see that the event $\{P_a(y = 1) < .25\}$ obtains whenever $x \geq 2$. The intersection of $\{P_a(y = 1) < .25\}$ with the information cell $[3, 1]_b$ thus consists of the three worlds $(2, 1)$, $(3, 1)$, and $(4, 1)$. We can therefore compute

$$P_b\{P_a(y = 1) < .15\} \;=\; \frac{P\{(2,1),(3,1),(4,1)\}}{P\{(0,1),(1,1),\ldots,(4,1)\}}.$$

This evaluates to about 0.265. When agent b in fact has exactly one ace, he will thus assign about 26.5% probability to the event that a assigns less than

	0	1	2	3	4
0	.322	.322	.322	.322	.322
1	.265	.265	.265	.265	.265
2	.194	.194	.194	.194	.194
3	.106	.106	.106	.106	.106
4	0	0	0	0	0

$P_a(y = 1)$

	0	1	2	3	4
0	.051	.027	.009	0	0
1	.051	.027	.009	0	0
2	.051	.027	.009	0	0
3	.051	.027	.009	0	0
4	.051	.027	.009	0	0

$P_b(P_a(y = 1) < .25)$

Table 7.2: Left, the value of the random variable $P_a(y = 1)$ in every possible world $\omega = (x, y)$, with x in the rows and y in the columns. The bars show a's knowledge partition, indicating that she can only distinguish possible worlds that differ on x. Right, the value of the variable $P_b(P_a(y = 1) < .25)$, with bars indicating b's knowledge partition.

25% probability to the event that he has exactly one ace. He thus considers it somewhat improbable, but definitely possible, that a assigns a low probability to his actual situation.

7.4 Aymptotic Accuracy

We now turn to the statistical consistency problem which is the main focus of this chapter. I wish to point out that the frequency approximation of probabilities, which we must rely on in order to internalize probabilities in Church, provides inconsistent estimates of second- or higher-order probabilities.

Suppose therefore we want to estimate some subjective probability

$$P_a(R)$$

where a is some completely uninformed agent, $a = \{\Omega\}$, and $R \subseteq \Omega$ is an arbitrary event. We can estimate this probability by taking n samples from the probability distribution P defined on the space Ω. This prior probability distribution is equal to a's posterior probability distribution in all worlds, since $a = \{\Omega\}$.

The Bernoulli theorem then tells us that the empirical frequency of the event, $\tilde{P}_a(R)$, is a uniformly good estimate of the probability $P_a(R)$. More precisely, the Chernoff-Hoeffding bound (Chernoff, 1952; Hoeffding, 1963) allows us to quantify this difference between the frequency and the probability after n samples by the inequality

$$P\left\{\left|\tilde{P}_a(R) - P_a(R)\right| > \varepsilon\right\} \;\leq\; 2\exp(-2n\varepsilon^2).$$

This bound holds for any precision level $\varepsilon > 0$ and any true value of $P_a(R)$. The probability of committing an estimation error larger than some fixed precision threshold ε converges exponentially fast to 0 as the number of samples increases. In technical terms, this means that frequencies are statistically consistent estimators of probabilities (Fisher, 1925).

However, suppose we are interested in approximating the nested probability

$$P_a(P_a(R) < 1/2).$$

Since $a = \{\Omega\}$, the random variable $P_a(R)$ has the same value in all possible worlds, and the inequality $P_a(R) < 1/2$ is either satisfied everywhere or nowhere. The higher-order probability $P_a(P_a(R) < 1/2)$ is thus either 1 or 0. No other values are possible.

We can estimate $P_a(P_a(R) < 1/2)$ by first computing an estimate of $P_a(R)$, and then deciding whether that constant random variable is larger than or smaller than 1/2. If we get the answer to this question wrong due to sampling noise, our estimate of $P_a(P_a(R) < 1/2)$ will thus deviate from the true value by a margin of $\varepsilon = 1$, and if we get it right, by $\varepsilon = 0$. Our estimate of $P_a(P_a(R) < 1/2)$ is therefore either perfectly correct or catastrophically wrong.

The problem is that we have no way of telling which of those cases we are in. If, for instance, $P_a(R) = 1/2$ in all worlds, we ought to decide that the proposition $P_a(R) < 1/2$ is everywhere false, so that

$$P_a(P_a(R) < 1/2) \;=\; 0.$$

However, since the estimate $\tilde{P}_a(R)$ will satisfy

$$\tilde{P}_a(R) \;\geq\; 1/2$$

with exactly 50% probability (given that n is even), our empirical approximation will in fact lead us to make the wrong decision half the time. For $P_a(R)$ close to 1/2, the chance of making a catastrophic error may thus be as high as 50% (cf. Fig. 7.2).

This leads us to the dismal conclusion that

$$\tilde{P}_a(\tilde{P}_a(R) < 1/2)$$

is a statistically inconsistent estimate of

$$P_a(P_a(R) < 1/2).$$

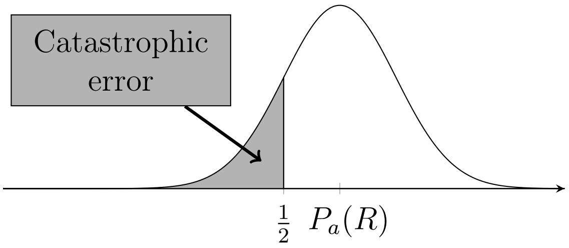


Figure 7.2: When the sample size is large, the empirical frequency $\tilde{P}_a(R)$ follows an approximately normal distribution around the actual probability $P_a(R)$. However, since estimates of higher-order statements about $P_a(R)$ may depend discontinuously on $P_a(R)$, even small estimation errors on the first-order level may lead to catastrophic estimation errors on the second-order level.

However large our sample is, there is always a value of $P_a(P_a(R) < 1/2)$ for which the empirical approximation is very likely to deviate significantly from the true value. Hence, even when our reasoning system returns a seemingly confident conclusion like

$$\tilde{P}_a(\tilde{P}_a(R) < 1/2) \ = \ 1 \qquad \text{(in all } \omega\text{)},$$

the actual result might in fact be the exact opposite.

It thus turns out that the estimation of second-order nested probabilities is a completely different problem from the estimation of first-order probabilities. The former has a straightforward and statistically consistent solution, and the latter does not. The underlying reason for this is that a thresholding operation like $\tilde{P}_a(R) < 1/2$ maps the probabilities $\tilde{P}_a(R)$ to truth values in a discontinuous way, and discontinuous functions are uncomputable by approximate means (Kushner, 1984, chs. 4–5). An approximate reasoning system like Church can thus not be unproblematically deployed to perform higher-order reasoning without occasionally delivering wrong results with high confidence.

Note also that this problem cannot be solved by using an adaptive sampling scheme that determines when to stop by looking at the results obtained so far (Wald, 1947). When $P_a(R) = 1/2$, any reasonable confidence interval around $\tilde{P}_a(R)$ will continue to contain 1/2 indefinitely and therefore not provide any firm conclusions about the proposition $P_a(R) < 1/2$. Since the statistician will not know whether this negative result is due to the probability being equal to 1/2 or merely to a lack of statistical power, the inequality will remain undecided indefinitely. Such an adaptive sampling scheme may thus remain undecided forever.

An example of a situation in which an estimation error of this kind may play a practical role is a coordination game in which two agents, a and b, try to meet at one of two a priori equally likely places. If a simulates b's unconditional behavior by flipping a coin, she is very likely to produce an estimate which skews slightly towards one of the two possible focal points, simply by the properties of the binomial distribution. If she makes her own choice based on a classical "hard" utility maximization rule, this would cause her to choose the slightly oversampled bar with probability 1, since this choice will have the highest payoff.

On the other hand, if she uses a "soft" maximization rule in the style of Stuhlmüller and Goodman (Stuhlmüller and Goodman, 2013, sec. 3.1), she would merely be more predisposed to choose the oversampled bar. However, even such a softmax decision scheme would effectively lead to an amplification of the sampling error, say, from 51% to 55%. If her choice was fed into a similar softmax decision rule for b, this would increase the bias even more, say, from 55% to 70%. As the layers of modal reasoning piled up, this trend would continue, pushing the probability of the two meeting places upwards. Thus, if a used a sufficient depth of modal reasoning, or if she set the "hardness" parameter for her decisions sufficiently high, the result would again be a decision distribution that would assign almost 100% probability to one of the two options.

However, this seemingly confident conclusion would be based on nothing but sampling error. If the entire chain of reasoning were repeated, the frequency in the initial sample would lean in the other direction with 50% probability, leading ultimately to an equally confident conclusion in the opposite direction. An agent using such a reasoning method would thus reach a high level in confidence that would not reflect the validity of the reasoning.

To illustrate this last point, we could imagine placing two agents of this sort in a situation that would require them to solve an unbiased coordination problem. Since their conclusions would ultimately go left or right with equal probability, they would fail to coordinate half the time. However, the hypothetical reasoning they performed internally would indicate that they should be close to 100% sure of achieving perfect coordination. Their reasoning would thus be inconsistent in the sense that it did not actually reflect the relevant aspects of the problem they were trying to solve.

7.5 Conclusion

The stochastic programming language Church is a useful conceptual tool for thinking about probabilistic reasoning, and it clarifies in many ways the logic underlying generative Bayesian models.

The recent proposal to use this language to perform reasoning about reasoning in a multi-agent setting is a welcome opportunity to merge the tradition of applied probability theory with the more speculative tradition in epistemic logic, which has typically focused on relatively small-scale, discrete problems. Both of these traditions, the statistical and the logical, have potentially huge contributions to make to this project once differences of terminology and style are overcome.

This chapter has had two goals: First, to clear away some of the potential confusion that might arise out of the complex Church syntax and show multi-agent probability theory can be embedded in classical probability theory by interpreting probabilities and expectations as random variables that are subject to uncertainty; and second, to use this insight to pinpoint a potentially catastrophic problem with the implementation of multi-agent probability theory in Church. As I have explained, the thresholding behavior which is an integral part of higher-order probability statements has the unfortunate consequence that otherwise statistically consistent estimation techniques can yield statistically inconsistent estimates of higher-order probabilities.

I thus see both grounds for excitement and for caution. The recent convergence of discrete logics with generative Bayesian models opens up many new possibilities for designing more realistic and practical models of human behavior, but without a proper theoretical understanding of these models, the risk is high that we will design systems that hide unpleasant surprises.

Chapter 8

An Epistemological Meditation on the Bias-Variance Trade-off

Statistical reasoning often requires a delicate balance between flexibility and caution. One way of formalizing this intuition is through the so-called bias-variance trade-off, which is a decomposition of prediction error into two sources, excessive data-sensitivity and excessive conservativity. While the conflicting aims of reducing bias and reducing variance is often presented as a justification for using estimators with arbitrary biases, the conventional form of the trade-off in fact depends crucially on the philosophical assumption that unknown parameters are constants rather than random variables. If we replace this objectivist philosophy of probability with a subjectivist one, the bias-variance trade-off appears as nothing more than a minimization problem with a unique set of solutions. All decisions about caution and flexibility are thus settled in advance by the choice of model. Whether this is a desirable feature of a reasoning calculus is not a question that I will address here; instead, I will try to clarify the conceptual ambiguity of the bias-variance trade-off in order to qualify any future discussions of its philosophical implications.

Should uncertainty always be expressed in terms of probability distribution? According to classical frequentist statistics, the answer is no: Parameter values are fixed but unknown constants rather than random variables, and the only legitimate move that the statistician can perform is to denounce certain parameter values as highly unlikely. The Bayesian answer, on the other hand, is yes: Probability theory is the universal calculus of uncertainty, and parameter values should thus be internalized in the model and treated like any other random variables (Jeffreys, 1939; Jaynes, 2003).

The Bayesian approach is attractive because of its uniformity, but it has proven controversial because of the element of arbitrary choice it entails (Keynes, 1921; Fisher, 1959). Before we can learn anything from data in a Bayesian framework, we need to choose priors on the parameter values, and there is often no compelling reason to choose one specific prior over all others. Maximum entropy arguments can narrow down the set of candidates, but even then, we usually need to select hyperparameters like the mean and variance of the priors without any statistical motivation. A Bayesian statistician thus always reaches a point where an arbitrary decision must be made about which priors or hyperpriors to begin with.

This element of arbitrary choice was historically been seen as problematic and unscientific by many critics. However, since the 1950s, it has become increasingly clear that arbitrary biases are not always a bad thing: For small sample sizes, unbiased estimators are often so data-sensitive that they are likely to be mislead by noise. Loading them with a slight preference for a certain region of the hypothesis space consequently increases rather than decreases their performance in many situations (Stein, 1956). One way of codifying this intuition is in terms of the bias-variance trade-off (Hastie et al., 2009).

8.1 The Bias-Variance Trade-off

In an estimation problem, the statistician attempts to guess the value of some hidden parameter θ based on a signal x which only gives noisy information about θ. Since the statistician can only use information available in x, any given estimation method can be represented of as a function $\hat{\theta}$ which maps a data set x onto a guess $\hat{\theta} = \hat{\theta}(x)$. The hope is that this guess tends to lie close to the actual value, that is, $\hat{\theta} \approx \theta$.

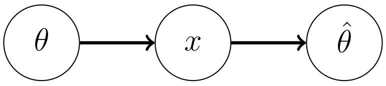


Figure 8.1: In an estimation problem, a data set x is drawn from a distribution with parameter θ, and the statistician then computes an estimate $\hat{\theta} = \hat{\theta}(x)$ of θ.

The bias-variance trade-off is a decomposition of the error committed by such an estimation method into two terms, the variance of the estimator and its squared bias:

$$\text{Mean squared error} \;=\; \text{Squared bias} + \text{Variance}$$

The mean squared error is a widely used loss function, and it is often used to evaluate the performance of a method for guessing a parameter value based on a data set. It is defined as the expected value of the squared distance between the guess $\hat{\theta}$ and the actual parameter value θ:

$$\mathsf{MSE}(\theta, \hat{\theta}) \;=\; E\left[(\theta - \hat{\theta})^2\right].$$

Alternatively, it could also be defined as the expected value of the squared error,

$$\mathsf{SE}(\theta, \hat{\theta}) \ = \ (\theta - \hat{\theta})^2.$$

The variance of the estimator $\hat{\theta}$ is here defined as usual,

$$\mathsf{VAR}(\hat{\theta}) \ = \ E\left[\left(\hat{\theta} - E[\hat{\theta}]\right)^2\right].$$

Variance is a measure of dispersion, so estimators that are very data-sensitive will tend to have a high variance.

The squared bias of the estimator $\hat{\theta}$ is the squared distance between the mean of the estimator and the true value of the target parameter,

$$\mathsf{Bias}^2(\theta, \hat{\theta}) \ = \ \left(\theta - E[\hat{\theta}]\right)^2.$$

Estimators that learn more slowly from data will tend to have a high squared bias, since they only take on extreme values if the data contains very strong patterns. The only situation in which low-variance estimators have low bias is when the mean of the estimator by sheer luck happens to coincide with the target parameter.

The point of decomposing the MSE into these two terms is that there is a conflict between minimizing bias and minimizing variance: Because conservative estimators tend to be more biased, driving down the variance will drive up the bias — unless you happen to load the dice towards just the right value.

8.2 Example: Beta Estimators

To illustrate these concepts, consider a bent coin with parameter $\theta = 1/5$. I toss it ten times and tell you k, the number of heads. You can then estimate θ by either of the following functions of the data:

$$\hat{\theta}_0(k) = \frac{k}{10}, \qquad \hat{\theta}_1(k) = \frac{k+1}{10+2}, \qquad \hat{\theta}_2(k) = \frac{k+2}{10+4}.$$

These three estimators are increasingly cautious about generalizing from data. By weighting their values with the binomial probabilities for $\theta = 1/5$, we find that their means are

$$E\left[\hat{\theta}_0\right] = \frac{1}{5}, \qquad E\left[\hat{\theta}_1\right] = \frac{1}{4}, \qquad E\left[\hat{\theta}_2\right] = \frac{2}{7}.$$

Computing the squared distances from these means to the true parameter value, we find the squared biases 0, 0.003, and 0.007. The variances, on the other hand, are 0.016, 0.011, and 0.008, respectively. The three estimators thus have increasing levels of bias and decreasing levels of variance.

As it happens, $\hat{\theta}_1$ is here the best estimator, with a MSE of $0.003 + 0.011 = 0.014$. However, this would not necessarily be the case for other values of θ, as shown in Figure 8.2.

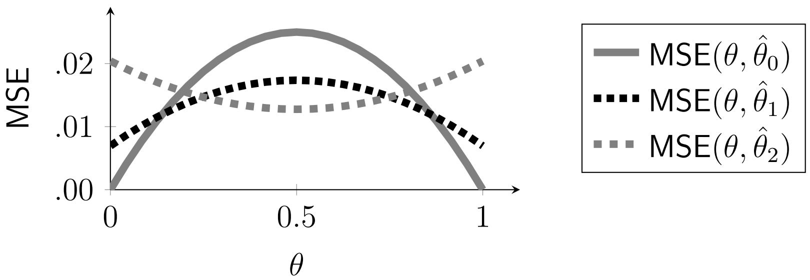


Figure 8.2: Mean squared error of three different estimators of the mean of a binomial distribution given a data set of size $n = 10$. The estimators $\hat{\theta}_1$ and $\hat{\theta}_2$ are biased towards $1/2$ and therefore have a higher error when θ is more extreme. They are more robust against noise, however, and therefore have a lower error when θ in fact does have value close to $1/2$.

8.3 The Frequentist Dilemma

In the example above, a more biased estimator outperformed a less biased one. This perhaps gives the impression that since we do not know the actual value of the parameter we are trying to estimate, we might as well choose an estimator on faith and hope that it has just the right level of caution (cf. Gigerenzer and Brighton, 2009).

This "leap of faith" interpretation of the bias-variance trade-off is largely valid from a frequentist perspective, but not from a Bayesian. To see this, we first need to inspect a conventional proof of the bias-variance decomposition in a bit more detail (I will here follow Hastie et al., 2009).

Consider therefore a situation in which we want to estimate a fixed parameter θ from a data set x, using an estimator $\hat{\theta} = \hat{\theta}(x)$, as in Figure 8.1. For any particular realization of x, this estimator will commit an error with a squared value of

$$\mathsf{SE}(\theta, \hat{\theta}) \;=\; (\theta - \hat{\theta})^2.$$

This error can also be thought of as a squared distance between θ and $\hat{\theta}$. We can thus, in a Pythagorean fashion, rewrite the expression in terms of an alternative path that goes from θ to $\hat{\theta}$ via a third point, $E[\hat{\theta}]$ (cf. Fig. 8.3). Inserting this waypoint and expanding the resulting expression, we get:

$$\begin{aligned}\mathsf{SE}(\theta, \hat{\theta}) &= \left(\theta - \hat{\theta}\right)^2 \\ &= \left(\theta - E[\hat{\theta}] + E[\hat{\theta}] - \hat{\theta}\right)^2 \\ &= \left(\theta - E[\hat{\theta}]\right)^2 + \left(E[\hat{\theta}] - \hat{\theta}\right)^2 - 2\left(\theta - E[\hat{\theta}]\right)\left(\hat{\theta} - E[\hat{\theta}]\right).\end{aligned}$$

Averaging these quantities over all data sets x, we get the expected values

$$\mathsf{MSE}(\theta, \hat{\theta}) \;=\; \mathsf{Bias}^2(\theta, \hat{\theta}) + \mathsf{VAR}(\hat{\theta}) - 2\,E\left[\left(\theta - E[\hat{\theta}]\right)\left(\hat{\theta} - E[\hat{\theta}]\right)\right]$$

We have thus decomposed the mean squared error into a bias and a variance minus a correction term which involves the covariance-like quantity

$$\mathsf{CR}(\theta, \hat{\theta}) \;=\; E\left[\left(\theta - E[\hat{\theta}]\right)\left(\hat{\theta} - E[\hat{\theta}]\right)\right].$$

This quantity can, to some extent, be seen as a measure of the amount of information that $\hat{\theta}$ carries about θ.

$$\begin{tikzcd}
 & E[\hat{\theta}] \arrow[dr] & \\
\theta \arrow[ur] \arrow[rr] & & \hat{\theta}
\end{tikzcd}$$

Figure 8.3: Two different routes connecting θ and $\hat{\theta}$.

However, we now invoke the frequentist assumption that θ is a fixed constant, and thus that the factor $(\theta - E[\hat{\theta}])$ is constant too. Because of the linearity of expectations, we can then pull it out of the expected value, leaving

$$\begin{aligned}
\mathsf{CR}(\theta, \hat{\theta}) &= (\theta - E[\hat{\theta}])\,E\left[\left(\hat{\theta} - E[\hat{\theta}]\right)\right] \\
&= (\theta - E[\hat{\theta}])\left(E[\hat{\theta}] - E[\hat{\theta}]\right) \\
&= 0.
\end{aligned}$$

When the value of θ is fixed, the value of $\hat{\theta}$ thus carries no information about θ. This is what we should expect, since the fluctuations in $\hat{\theta}$ are then only due to the randomness in x and not in θ. We thus have the clean decomposition of the mean squared error,

$$\mathsf{MSE}(\theta, \hat{\theta}) \;=\; \mathsf{Bias}^2(\theta, \hat{\theta}) + \mathsf{VAR}(\hat{\theta}).$$

Notice, however, that all expected values in this equation implicitly range over the conditional distribution of x given θ, not over a joint distribution of x and θ. The mean of $\hat{\theta}$, for instance, is found by averaging its values over all data sets weighted by their likelihood given the fixed value of θ. The model at hand is thus a hybrid model in which θ is a logical variable, while x and $\hat{\theta}$ are random variables. In a Bayesian model, they would all be random variables.

Notice further that the equation is false in such a Bayesian model, since the CR term does not necessarily vanish there. In such a model, we instead have the following inequality for all reasonable estimators:

$$\mathsf{MSE} \;\leq\; \mathsf{Bias}^2 + \mathsf{VAR},$$

with the gap accounted for by the amount of information $\hat{\theta}$ captures about θ (if this gap is negative, $E[\hat{\theta}] - \hat{\theta}$ is a better estimator than $\hat{\theta}$ and can be used instead). Imposing a prior distribution on θ and integrating out θ and x instead of fixing θ and integrating out x thus gives a substantially different analysis.

8.4 Two Examples with "Gaps"

As a toy example of how the Bayesian analysis of the bias-variance trade-off differs from the frequentist, consider the joint probability distribution in Table 8.1.

Informally, this distribution describes a kind of random walk in which θ starts at $\theta = 3$ and then takes a unit step with probability 1/2; then x is initiated at $x = \theta$ and also takes a unit step with probability 1/2.

x	1	2	3	4	5	Σ
$\theta = 2$	1/16	1/8	1/16	–	–	1/4
$\theta = 3$	–	1/8	1/4	1/8	–	1/2
$\theta = 4$	–	–	1/16	1/8	1/16	1/4
$\hat{\theta}_1$	2	2	3	4	4	
$\hat{\theta}_2$	2	3	3	3	4	
Σ	1/16	1/4	3/8	1/4	1/16	1

Table 8.1: A joint distribution between a parameter θ and a data point x, including the value of the two estimators $\hat{\theta}_1$ and $\hat{\theta}_2$ for each data point x.

	MSE	=	Bias^2	+	VAR	−	2 CR
$\hat{\theta}_1$:	3/8	=	1/2	+	5/8	−	2 (3/8)
$\hat{\theta}_2$:	3/8	=	1/2	+	1/8	−	2 (1/8)

Table 8.2: Decomposition of the mean squared error for the two estimators.

This model suggests two natural estimators, the more flexible $\hat{\theta}_1$ or the more conservative $\hat{\theta}_2$. As shown in Table 8.2, these estimators have the same bias but different variance. However, $\hat{\theta}_1$ changes more when θ does, and it consequently captures more information about θ. The net outcome is that the two estimators are equally good in terms of mean squared error.

As a second example, suppose θ is distributed according to a triangular prior on the interval $[-1, 1]$, and x is distributed according to a triangular likelihood on $[\theta - 1, \theta + 1]$, as depicted in Figure 8.4. In this situation, a natural choice of estimator is the maximum likelihood estimate $\hat{\theta}(x) = x$.

As it turns out, however, this estimator performs strictly worse than the more restrained estimator $\hat{\theta}(x) = x/2$ because of the difference in variance. This

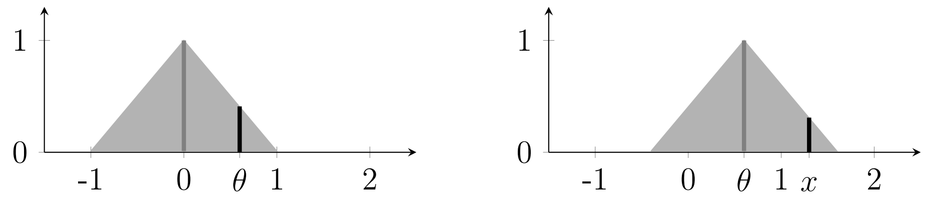


Figure 8.4: The parameter θ is first sampled from a triangular distribution with mean 0, and the data point x is then sampled from a triangular distribution with mean θ.

	MSE	$=$	Bias^2	$+$	VAR	$-$	$2\mathsf{CR}$
$\hat{\theta}(x) = x$:	$1/6$	$=$	$1/6$	$+$	$1/3$	$-$	$2\,(1/6)$
$\hat{\theta}(x) = x/2$:	$1/12$	$=$	$1/6$	$+$	$1/12$	$-$	$2\,(1/12)$
$\hat{\theta}(x) = 0$:	$1/6$	$=$	$1/6$	$+$	0	$-$	$2\,(0)$

Table 8.3: Decomposition of the mean squared error for three three estimators of the mean of a triangular distribution.

fact would be invisible if we had only looked for estimators that made the sum $\mathsf{Bias}^2(\theta, \hat{\theta}) + \mathsf{VAR}(\hat{\theta})$ small — in fact, among the estimators of the form $\hat{\theta}(x) = \alpha x$, the one that minimizes $\mathsf{Bias}^2(\theta, \hat{\theta}) + \mathsf{VAR}(\hat{\theta})$ is the "blind" estimator $\hat{\theta}(x) = 0$. Surprisingly, this estimator performs just as well as $\hat{\theta}(x) = x$ in terms of squared error.

8.5 Conclusion

The common form of the bias-variance decomposition is not valid in a Bayesian model. When parameters are internalized as random variables in the model, squared bias plus variance is larger than mean squared error for all reasonable estimators. The gap between the MSE and $\mathsf{Bias}^2 + \mathsf{VAR}$ is accounted for by a statistic which measures how much information the estimator carries about its target parameter.

This means that there is no bias-variance dilemma in Bayesian statistics, at least not for a fixed amount of data: Any statistical situation with a given prior distribution on the source parameter and a given likelihood model for the data channel automatically prescribes a set of optimal estimators. This means that the bias-variance trade-off takes on a very different interpretation when the parameters are random variables rather than fixed but unknown constants.

This does not necessarily provide an argument in favour of Bayesian over frequentist methods or *vice versa*. It is not clear whether the benefits of Bayesian clarity outweighs the flexibility of the frequentist approach. But if we do represent our uncertainty as a probability distribution, it is important to keep in mind that all questions, including questions about estimation methods, have definite answers that follow mechanically from the features of the models. In a consistently Bayesian philosophy, there is, strictly speaking, no room for leaps of faith.

In the ongoing and vitriolic debates about Bayesian inference, induction, and the possibility of assumption-free knowledge, this fact might be worth keeping in mind. Bayesian models yield clear-cut answers, but to get more out of your calculus, you generally have to put more in there as well.

Chapter 9

Fear and Loathing in Statistics

Almost from its very conception, the field of statistics has been split into warring camps which were hardly willing to recognize even the legitimacy of their competitors, often fiercely attacking their scientific integrity and conception of proper reasoning. How could this debate ever get so out of hand? This chapter attempts at an answer from a historical perspective: I explain the difference between the two conceptions of statistics by attributing them to two different rationalities, each with its own historical origin. I thus trace the pedigree of frequentist statistics to a "bureaucratic" rationality which began to spread in the 17th century, and the success of Bayesian statistics to an "entrepreneurial" rationality which gained a foothold as high technology began to play a more important economic role in the late 19th century. Based on this historical reconstruction, I conclude that the disagreement over the proper foundations for statistics is so intimately connected with the very concept of rationality that they are unlikely to be resolved by rational means.

9.1 Introduction

Throughout the 20th century, statistics has been plagued by a surprisingly vitriolic debate about the proper principles for problems of inference under uncertainty.

One tradition, stemming from Bayes and Laplace, has argued that all matters of uncertainty should be treated with the calculus of probabilities. Another tradition, with roots in the work of Bernoulli but more recently represented by Fisher, Neyman, and Pearson, has argued that the use of probability distributions should be restricted to cases involving actual randomness in nature itself. Mostly though not always, this divide has coincided with two different interpretations of the concept of probability, either as tallying device for comparing relative plausibilities, or as an objective property of the physical world.

This may sound like a relatively fine point of interpretation. It is therefore somewhat surprising to read the heated rhetoric that has surrounded this debate. The frequentists, on their side, have long argued that the use of probability distributions to express all kinds of uncertainty was "extremely arbitrary" (Fisher, 1922, p. 325) and "useless for scientific purposes" (Fisher, 1947, p. 7). The Bayesians, in turn, have answered that the frequentist alternative builds on "an incomplete or fragmentary justification or even no justification at all" (de Finetti, 1972, p. 161), and that it only approximates "arbitrarily and imperfectly" what one can do "uniquely and optimally" with the Bayesian tools (Jaynes, 2003, p. xvii).

The aim of this chapter is to investigate how this debate has managed to get so out of hand. Why has the controversy not been settled, and why does it continue to provoke such heated responses? What could possibly provoke distinguished Fellows of the Royal Society to hurl words like "charlatanry" (Keynes, 1921, pp. 418, 438) at each other over what appears to be little more than a choice of mathematical model? How can professional scientists and statisticians continue, even to this day, to feel "cheated and frustrated" (Wang, 1993, p. vii) by what they see as misapplication of statistical methodology?

My approach to answering this question is essentially historical. After briefly sketching the mathematic outline of what the disagreement is about, I will attempt to document a historical link between the two schools of statistics and two different events in the history of rationality. For the frequentist school, this genealogy focuses on the emergence of the moneyed and educated middle classes in early modern Europe, with their increasingly confident class consciousness. For the Bayesians, I trace the roots of their particular brand of rationality to the second industrial revolution, that is, the increasing importance of the electrical and chemical industries in the late 19th century.

In both cases, my goal will be to show how a certain mathematical calculus reflects a certain philosophical anthropology, or more precisely, a rationality. These two conflicting rationalities differ on what they take as admissible arguments and valid conclusions, and they consequently both see their competitors as hopelessly irrational. Since the roots of the debate thus reach all the way down into the very concept of rationality itself, it is unlikely to be settled by rational means.

9.2 Mathematical Background

The central topic of probability theory is the question of how to compute the probability of various observations once the relevant parameter values are given. For instance, given that a certain coin has a bias θ, we might want to know the probability that it comes up heads five times in a row.

Statistics, on the other hand, is about the reverse problem: We observe a data set x, and we want to use this observation to conclude something about an unknown parameter θ. One way of thinking about this problem is to imagine an indexed family of probability distributions,

$$\{P_\theta \,|\, \theta \in \Theta\}.$$

We imagine that one of these distributions is selected, and an observation x is drawn from it. The statistician then observes x but not θ and wants to say something interesting about the distribution P_θ on the basis of that information.

As an example, imagine that I have a bag of chips enumerated

$$1, 2, 3, \ldots, \theta.$$

I then draw a chip at random from this bag and show it to you. In the terminology introduced above, we thus have a family of uniform distributions, $P_\theta(x) = 1/\theta$, one for each θ (cf. Table 9.1). This family can be indexed by θ, the number of chips in the bag. As a statistician, your task is then to make some kind of decision based on the observation x. You might for instance take a guess at the value of θ, narrow down the range of possibilities, place a bet, or decide whether a specific value of θ should count as plausible or not.

Barring a few minor complications, this is the nature of all statistical inference problems. An observation is made according to some probability distribution, and you are interested in making some kind of decision whose soundness depends on which distribution the data came from. In other words, you get to see x, and you want to find something intelligent to say about P_θ.

As stated, this problem does not have a single, unambiguous answer. Typically, the observation x could logically have come from many different distributions, and there is no single, universally accepted way of summarizing what x tells you about P_θ. For instance, you might draw a chip numbered $x = 17$ in the problem described above; but this observation in itself does not coerce you into any specific course of action, or commit you to any definite conclusion except for the trivial observation that $\theta \geq 17$. In order to reach more substantial conclusions, you need to make more substantial assumptions.

In order to turn a statistical situation into a well-defined math problem, we need to enrich our representation of the situation, and there is no consensus on how to best do this. However, two main schools exist: The classical approach, which recommends using methods that have provably good properties for all values of θ; and the Bayesian approach, which suggests that we should express our uncertainty by imposing a probability distribution on θ. In the following subsections, I will discuss each of these separately, and then consider some of the controversies surrounding them.

9.2.1 Across the Gap

Historically, the first mathematical approach to statistics was proposed by Bernoulli in his influential book *Ars Conjectandi*, which was published posthumously

$P_\theta(x)$	$x = 1$	$x = 2$	$x = 3$	$x = 4$	$x = 5$	$\cdots$
$\theta = 1$	1	0	0	0	0	$\cdots$
$\theta = 2$	$1/2$	$1/2$	0	0	0	$\cdots$
$\theta = 3$	$1/3$	$1/3$	$1/3$	0	0	$\cdots$
$\theta = 4$	$1/4$	$1/4$	$1/4$	$1/4$	0	$\cdots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$

Table 9.1: The distributions associated with the problem of inferring the number θ based on a randomly selected element x from the set $\{1, 2, \ldots, \theta\}$.

in 1713. Bernoulli treated the problem of guessing the bias of a coin based on a large number of tosses; or more generally, the problem of estimating of the probability $\theta = P_\theta(A)$ of a fixed event A on the basis of its empirical frequency (Bernoulli, 2006, Part Four).

Bernoulli's approach was based on an theorem about frequencies: In a very long series of experiments, it is highly unlikely that the relative frequency of a fixed event A deviates much from its probability. For instance, if we buy a large batch of seeds, and if any given seed is fertile with probability θ, then the percentage of fertile seeds in our batch is probably close to θ. In fact, this percentage will fall between $\theta - \varepsilon$ and $\theta + \varepsilon$ with a probability that approaches 1 as the sample size grows.

This is good news for statistics, since it implies that probabilities can be measured empirically. Specifically,

> ... so many experiments can be taken that it becomes any given number of times (say, c times) more likely that the number of fertile observations will fall between these bounds than outside them ... (Bernoulli, 2006, p. 337)

In order to appreciate the content of this claim, we should remember that k, the number of fertile observations, is itself a random variable. Bernoulli's theorem therefore states that this variable follows a probability distribution which clusters closely around θ when the sample is large (cf. Fig. 9.1). Specifically, in a series of n trials,

$$\forall\, \varepsilon > 0,\; c > 0\; \exists\, n\; \forall\, \theta \in [0,1] :\; P_\theta \left\{ \left| \frac{k}{n} - \theta \right| \geq \varepsilon \right\} \;\leq\; \frac{1}{c+1}.$$

The probability of a large deviation between k/n and θ thus dips below any positive constant $1/(c + 1) > 0$ when n is large enough.

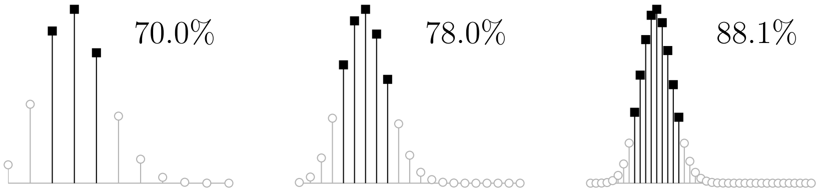


Figure 9.1: According to Bernoulli's theorem, the probability mass of a binomial distribution clusters increasingly around its mean as the sample size goes up.

This holds regardless of the value of θ, and Bernoulli explicitly described how to compute large enough values of $n = n(\varepsilon, c)$ for a given ε and c. For instance, with $\theta = 30/50$ and $\varepsilon = 1/50$, he explained,

> . . . it is inferred that if 25,550 experiments are taken, it will be more than 1000 times more likely that the ratio of the number of fertile observations to the number of all the observations will fall between these bounds, 31/50 and 29/50, than outside them. On the same understanding, if c is set equal to 10,000 or 100,000, it may be seen that it will be more than ten thousand times more probable, if there are 31,258 experiments, and more than a hundred thousand times more probable, if there are 36,966, and so forth to infinity, continually adding to the 25,550 another 5708 experiments. (Bernoulli, 2006, p. 339)

The statement is a theorem of ordinary probability theory, asserting something about a random variable k/n given the value of a parameter θ. However, because the bound holds for all the probability distributions in the family $\{P_\theta \mid \theta \in [0, 1]\}$, it provides us with an estimate that we can rely on without knowing in advance which distribution the data was drawn from. In other words, Bernoulli's theorem allows us to upper-bound the probable error committed by the approximation

$$\theta \;\approx\; \frac{k}{n}$$

without making any reference to the specific value of θ. Bernoulli's theorem thus provides us with a uniform warranty for a certain measuring device, guaranteeing that we can always probe the value of a probability $P_\theta(A)$ by means of the frequency k/n.

The theorem is thus a bridging theorem, relating the unobservable world of probabilities to the observable world of frequencies. This allowed Bernoulli to cut through the vicious circle of defining probabilities in terms of "probable" frequencies. Instead, he could assign the probabilities a role as unobservable quantities that are indirectly revealed through observable data.

9.2.2 Probabilities, Unfixed

Half a century after the publication of Bernoulli's book, a paper by the Scottish reverend Bayes was read before the Royal Society of London (Bayes, 1763). In this paper, which like Bernoulli's book was published after his death, Bayes suggested an alternative method for dealing with statistical problems.

His approach was to treat the unknown bias of a coin as if it "had been at first unfixed" (p. 393) and was then assigned a new value at random. This allowed him to treat the uncertainty about the parameter θ by means of the calculus of probabilities. The parameter θ would be internalized into the model alongside the observable quantity k.

More specifically, Bayes suggested the following analogy: Suppose first we throw a ball onto a square table in such a way that "there shall be the same probability that it rests upon any one equal part of the plane as another." We then record where that ball comes to rest (Postulate 1, p. 385). After that, we throw another n balls across the table, counting them as successes if they land to the right of the first ball, and as failures if they land to the left (Postulate 2, p. 385; see also Fig. 9.2).

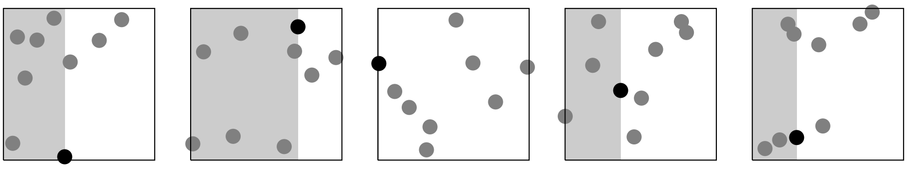


Figure 9.2: Parameter values and data sets drawn according to the Bayesian randomization device. In the examples shown here, the sample size is $n = 8$, and the data sets contain $k = 3, 2, 8, 5, 4$ successes, respectively.

Using this ball-throwing device is equivalent to drawing the parameter θ from a uniform distribution on the unit interval. θ is then a random variable whose distribution is described by the probability density

$$p(\theta) \;=\; \begin{cases} 1 & \theta \in [0,1], \\ 0 & \text{otherwise.} \end{cases}$$

Once we have used this distribution to select a value for θ, we can perform n coin tosses with a coin of bias θ and count the number of successes. Thus, having drawn θ from the uniform distribution $p(\theta)$, we draw k, the number of successes, from the binomial distribution $P_\theta(k) = P(k \mid \theta)$.

This changes the terms of the problem. Now, rather than treating θ as a logical variable bound by a universal quantifier, we have incorporated it into the probabilistic model. This means that it can be handled by the same tools as k. Both k and θ are random variables, qualitatively equal and subject to the same methods of inference.

$p(x,\theta)$	$x=1$	$x=2$	$x=3$	$x=4$	$x=5$	$\cdots$	$p(\theta)$
$\theta=1$	$1/2$	0	0	0	0	$\cdots$	$1/2$
$\theta=2$	$1/8$	$1/8$	0	0	0	$\cdots$	$1/4$
$\theta=3$	$1/24$	$1/24$	$1/24$	0	0	$\cdots$	$1/8$
$\theta=4$	$1/64$	$1/64$	$1/64$	$1/64$	0	$\cdots$	$1/16$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
$p(x)$	.693	.193	.068	.026	.011	$\cdots$	1

Table 9.2: By imposing a prior probability distribution on the parameter, a family of conditional distributions can be transformed into a table of joint probabilities. Given a specific observation, a posterior distribution can then be obtained by normalizing the relevant column.

It also means that we are no longer working with a family $\{P_\theta \mid \theta \in \Theta\}$ of distributions indexed by $\Theta = [0,1]$. Instead, we now have a single probability distribution P that covers the entire two-dimensional space $\Omega \times \Theta$, where $\Omega = \{0,1,2,\ldots,n\}$. Everything in this model, the parameter and the data, is thus imagined to be the outcome of a single, large experiment. Every assertion that we can formulate about k and θ has a specific probability.

Consider for instance the marginal probability $P(k)$. This is the probability of obtaining k successes in n trials, but not relative to any parameter value in particular. In a classical statistical model, such free-floating probabilities do not exist, since the model does not allow us to compute averages of the conditional probabilities $P_\theta(k) = P(k \mid \theta)$. Although the probabilities $P_\theta(k)$ do exist for every $\theta \in \Theta$, they cannot be aggregated into a single number.

In a Bayesian model, on the other hand, the prior probability distribution over θ allows us to compute weighted averages of any function of θ. The marginal probability $P(k)$ is an average of the conditional probabilities $P(k \mid \theta)$, and it therefore exists. In the specific model described by Bayes, it happens to be equal to $P(k) = 1/(n+1)$.

This can be seen by noting that Bayes' sequential randomization scheme is equivalent to a simpler scheme in which we throw all $n+1$ balls onto the table at the same time; we then select the designated "parameter ball" afterwards. Since all one of the $n+1$ balls are equally likely to be selected as the "parameter ball," the number of balls to the right of the parameter ball is equally likely to be any of the numbers $k = 0, 1, 2, \ldots, n$.

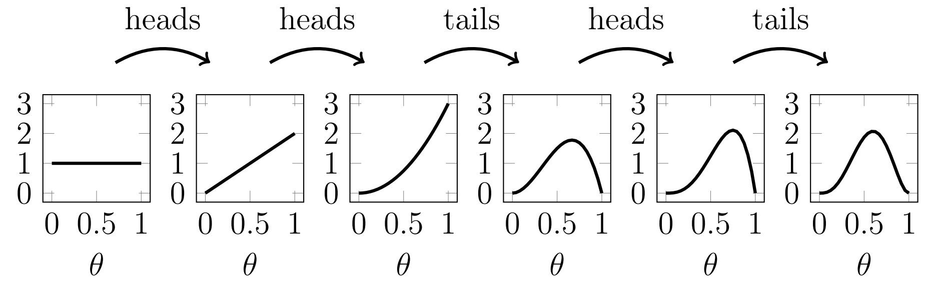


Figure 9.3: Posterior distributions over θ for increasingly large data sets.

Since there are $n + 1$ possible values, this implies that

$$P(k) \;=\; \int_0^1 P(k \,|\, \theta)\, p(\theta)\, d\theta \;=\; \frac{1}{n+1}, \qquad k = 0, 1, 2, \ldots, n.$$

Bayes was aware of this feature of his model and took it as a confirmation of its soundness:

> ... in the case of an event concerning the probability of which we absolutely know nothing antecedently to any trials made concerning it, ... I have no reason to think that, in a certain number of trials, it should rather happen any one possible number of times than another. (Bayes, 1763, p. 393)

Consider now the the posterior probability densities $p(\theta \,|\, k)$. Since the Bayesian assumption essentially reduces the whole model to a single random experiment, it does not draw any theoretical distinction between the conditional distribution of k given θ and the conditional distribution of θ given k. If we think about the joint distribution of these two variables as a big, two-dimensional table (cf. Table 9.2), then these two types of conditional distributions simply correspond to normalized rows and normalized columns, respectively.

Suppose therefore we hold our data point k fixed in order to find the conditional distribution of θ. We can then imagine letting θ slide through the unit interval, tracing out the corresponding values of the joint density

$$p(k, \theta) \;=\; P(k \,|\, \theta)\, p(\theta), \qquad \theta \in [0, 1].$$

If we scale up this function so that it integrates to 1 over the unit interval, it will be equal to the posterior density of θ given k.

Notice that this requires us to think about the binomial probabilities $P_\theta(k) = P(k \,|\, \theta)$ as a function of θ rather than, as we usually do, a function of k. We plug in various possible explanations θ and reweight their prior probability densities by how well they explain the observation k (cf. Fig. 9.3). This provides us with the new distribution over θ.

As this discussion shows, Bayes' approach to the problem of statistical inference was to deliberately shake loose the parameter θ so that it could be treated in terms of probabilities. This intervention reduced a problem of statistical inference to a problem of probability theory. Having thus integrated θ into the calculus of probabilities on equal footing with k, he could legitimately speak of

> ... the chance that the probability of [success] in a single trial lies somewhere between any two degrees of probability that can be named. (Bayes, 1763, p. 376)

Bernoulli would probably not have recognized this as a meaningful statement. For him, the "probability" — i.e., the parameter θ — was a fixed but unknown constant. There could not be any question of assigning a "chance" to a statement like $a \leq \theta \leq b$. Bayes, on the other hand, seems to have thought that such a distinction read too much into the concept of chance, and he apparently saw no serious harm in using a make-believe randomization device to convert one kind of uncertainty into another.

9.3 Aftermath

Almost from its conception, the Bayesian approach to statistics was the topic of fierce criticism. In this section, I briefly sketch the history of this debate.

9.3.1 First Sight of Conflict

Bayes never applied his method to any other problem than the coin flipping scenario discussed above. However, the French mathematician Laplace, who had taken up the use of "inverse probability" independently of Bayes, applied it much more widely in a series of papers read before the French Academy of Sciences (Stigler, 1986, Ch. 3) and in his later book *Essai philosophique sur les Probabilités* (Laplace, 1995). These works dealt both with the problem of inferring the value of an astronomical quantity based on noisy observations, and with discrete problems like the Bayesian coin flipping scenario.

Like Bayes, Laplace had noticed that the assumption of a flat prior over the success rate of an experiment implies that a streak of n successes in a row has probability $1/(n+1)$. Since a streak of $n-1$ successes then has probability $1/n$, the probability of another success at the end of a streak of $n-1$ is therefore

$$\frac{1/(n+1)}{1/n} = \frac{n}{n+1}.$$

In one famous case, he pushed the logic of this argument to its extreme:

> For example, if we place the dawn of history at 5,000 years before the present date, we have 1,826,213 days on which the sun has constantly risen in each 24 hour period. We may therefore lay odds of 1,826,214 to 1 that it will rise again tomorrow. (Laplace, 1995, p. 11)

This example did not fail to provoke a reaction, already in the 19th century. Venn, for instance, cited Laplace's computation and commented that he found it "hard to take a rule such as this seriously" (Venn, 1888, p. 197). Echoing a similar critique by Cournot (Hald, 2007, pp. 76–77), he found that the probabilities suggested by this "Rule of Succession" were at odds with common sense, particularly in extreme cases:

> ... when an event has happened but a few times, we have no certain guide; and when it has happened but once, we have no guide whatever, ... (Venn, 1888, p. 198)

Like Cournot, Venn proposed a rigorous distinction between objective frequencies and subjective degrees of belief. This would allow him to argue that a conclusion based on a single data point had no basis in objective fact, whatever subjective beliefs one might encounter. By showing that Laplace's reasoning led to a counterintuitive result in an extreme case, Cournot and Venn thus felt that they had undermined the general validity of his arguments.

Similar points of critique were raised by other writers, often citing the subjective and arbitrary character of the Bayesian method as a major concern. The Scottish actuary Chrystal, for instance, would also announce that

> ... the so-called laws of Inverse Probability are a useless appendage to the first principles of probability, if indeed they be not a flat contradiction of those very principles. (Chrystal, 1891, p. 423)

In his book on probability, Keynes strengthened this already quite strong language, calling the Bayesian methods "baseless," "ridiculous," and "preposterous" (Keynes, 1921, pp. 425, 436, 443).

However, not everyone accepted these objections. The physicist Jeffreys defended the Bayesian paradigm by asserting that all systems of inductive inference "must involve an *a priori* postulate," thus relying on something that is "believed independently of experience" (Jeffreys, 1933, p. 524). This put everybody in the same boat, since

> ... *if any method appears to avoid such an assumption it must either be erroneous in principle or involve some* a priori *assumption that the author has not stated and possibly has not noticed.* (Jeffreys, 1933, p. 525; emphasis in original)

Since we are all in the business of making arbitrary *a priori* assumptions anyway, Jeffreys suggested, we need not worry too much about the use of prior probabilities. Any solution to a statistical problem, across methodologies, starts by selecting a family $\{P_\theta \,|\, \theta \in \Theta\}$ of distributions without any statistical justification; so why frown at the additional choice of a probability distribution over θ? A coin flipping model, for instance, excludes the possibility of dependencies between the coin flips; why is that better than assigning certain prior probabilities to the range of possibilities?

The answer would often be that the use of prior probability distributions conflated two different representations of uncertainty. Fisher, perhaps the most influential statistician of the 20th century, commented:

> ... advocates of inverse probability seem forced to regard mathematical probability, not as an objective quantity measured by observed frequencies, but as measuring merely psychological tendencies, theorems respecting which are useless for scientific purposes. (Fisher, 1947, pp. 6–7)

By confusing the two meanings of the word "probability," such a methodology would therefore put a calculus of private beliefs in place of truly "rigorous and unequivocal inference" (Fisher, 1947, p. 4). Assigning probabilities to particular facts that are either true or false, for instance, "amounts to assuming that ... our universe had been selected at random" (Fisher, 1922, p. 326). However convenient that assumption may be, it has no basis in evidence of an actual universe-sampling mechanism. For Fisher, this disqualified the approach as an elaborate mathematical make-believe.

9.3.2 First Sight of Compromise

In 1933, Neyman and Pearson published a paper on hypothesis testing which contained a simple but conceptually important lemma. The lemma was formulated in the context of hypothesis testing, that is, the problem of how to best guess which of two probability distributions P_0 and P_1 a data set $x \in \Omega$ was drawn from (Neyman and Pearson, 1933, pp. 298–302).

A solution to such a problem must necessarily come in the form of a decision policy telling the statistician, for each data set, which guess to submit. It can thus be modeled as a function T from the data space Ω to the two possible guesses, $T(x) = 0$ and $T(x) = 1$. Such a testing regime produces two kinds of errors: Guessing $T = 1$ when the data was drawn from P_0, and $T = 0$ when the data was drawn from P_1. The probabilities of these two errors are

$$\alpha(T) \;=\; P_0(T = 1) \qquad \text{and} \qquad \beta(T) \;=\; P_1(T = 0).$$

One family of tests which can always be applied in this context are the likelihood ratio tests. These are the tests that choose a guess R by comparing the ratio P_0/P_1 to a fixed threshold of evidence:

$$R(x) \;=\; 1 \qquad \Longleftrightarrow \qquad \frac{P_1(x)}{P_0(x)} \;\geq\; \tau.$$

The content of Neyman and Pearson's lemma can then be stated as follows: For any arbitrary test T and any likelihood ratio test R,

$$\alpha(T) \;\leq\; \alpha(R) \qquad \Longrightarrow \qquad \beta(T) \;\geq\; \beta(R).$$

In other words, the test T can only perform better than R on one measure by performing worse on the other. Among all the tests with an error rate $P_0(T = 1) \leq \alpha$, the likelihood ratio test thus achieves the lowest error rate $P_1(T = 0)$. The likelihood ratio tests are, in this sense, optimal (Neyman and Pearson, 1933, pp. 300–301; see also Neyman, 1950, Th. 5.1, p. 305–306).

This lemma proved that the frequentist apparatus of hypothesis testing could in fact be rationalized as a Bayesian decision rule: The choice of a specific threshold r could be interpreted as a covert choice of two prior probabilities, $P(\theta = 0)$ and $P(\theta = 1)$, since the threshold

$$\tau \;=\; \frac{P(\theta = 0)}{P(\theta = 1)}$$

then corresponds to a comparison of two joint probabilities:

$$R(x) \;=\; 1 \qquad \Longleftrightarrow \qquad P_1(x)\,P(\theta = 1) \;\geq\; P_0(x)\,P(\theta = 0).$$

Since the posterior probabilities $P(\theta \,|\, x)$ are proportional to the joint probabilities $P(x, \theta) = P_\theta(x)\,P(\theta)$, this decision policy amounts to selecting the hypothesis with the higher posterior probability. The set of reasonable frequentist tests thus corresponds in a one-to-one fashion with the set of *a priori* probabilities for the two hypotheses.

This idea was pursued much more systematically by Wald, who suggested that frequentist methods in statistics could be seen as a kind of worst-case Bayesian analysis: Instead of minimizing the average loss with respect to a fixed *a priori* distribution, the frequentist methodology could be seen as an attempt to identify the prior distribution that would make the Bayesian inference problem as hard as possible, and then solve that problem.

This "minimax" approach, he commented, "seems, in general, to be a reasonable solution of the decision problem when an a priori distribution in Ω does not exist or is unknown to the experimenter" (Wald, 1950, p. 18). It also established a kind of mutual intelligibility between the two approaches:

> There is an intimate connection between minimax solutions and Bayes solutions. ... a minimax solution is, under some weak restrictions, a Bayes solution relative to a least favorable a priori distribution. (Wald, 1950, p. 18)

In a certain sense, these results vindicated both the Bayesian and the frequentist approaches: The Bayesians could say that frequentist statistics had really just been covertly Bayesian all along, and the frequentists could say that they had no objections to a Bayesian solution as long as it would remain robust to any choice of prior. Everyone, it would seem, could be happy.

9.3.3 Back in the Trenches

In fact, the outcome was quite the opposite. The Italian statistician de Finetti was puzzled by Wald's work and felt that it relied on an overly pessimistic "superstition" of a world that wants the worst for us, in the sense of wanting to maximize our losses (de Finetti, 1972, p. 183). According to de Finetti's interpretation, Wald had simply proven that the frequentist analysis rested on a "not yet openly recognized prior opinion" (p. 183), and he suggested that we should be free to choose our prior probabilities as we liked:

> Of course, these marginal elements [i.e., prior probabilities] do not appear in Wald's formulation; their absence there is just what prevents the problem of decision from having the solution that is obvious when the table [of conditional probabilities] is thus completed. (de Finetti, 1972, p. 179)

Jaynes, a fierce polemicist against frequentist methods, made the same case (Jaynes, 2003, pp. 64–66). Why worry about worst-case behavior, he suggested, when all practical experience suggests that things are usually not so bad? The choice ought to be an easy one:

> We are now in possession of proven theorems and masses of worked-out numerical examples. As a result, the superiority of Bayesian methods is now a thoroughly demonstrated fact in a hundred different areas. One can argue with a philosophy; it is not so easy to argue with a computer printout, which says to us: "Independently of all your philosophy, here are the facts of actual performance." (Jaynes, 2003, p. xxii)

Thus, on the Bayesian side, no one seemed interested in a compromise. Wald's suggestions were either rejected as wrong-headed, or taken as proof of the superfluity of frequentist statistics, or both.

On the frequentist side, the picture was largely the same, but with the opposite conclusion. Fisher had always been scornful of the work of Neyman and Pearson, and he stuck to this assessment even after parts of it started to make its way into the standard statistics curriculum. In Fisher's view, a scientific theory P_0 should be evaluated by its own lights, not by how favorably it compared to some explicit alternative P_1. In a late article, he explained that

> ... this difference in point of view originated when Neyman, thinking that he was correcting and improving my own early work on tests of significance, ... in fact reinterpreted them in terms of that technological and commercial apparatus which is known as an acceptance procedure. (Fisher, 1955, p. 69)

Rejecting all such "acceptance procedures," Fisher insisted that a hypothesis test ought to evaluate whether an observation could reasonably have arisen from a given "null hypothesis" P_0, and nothing else. No viable alternative was needed in order to discard P_0.

Under such an interpretation, Neyman and Pearson's rate of "false positives" could therefore be given an exact meaning, since it expressed a probability under P_0. Their "false negatives," on the other hand, were nonsensical concepts according to Fisher, since they required the use of an alternative probability measure P_1. This was close to being a contradiction in terms:

> It was only when the relation between a test of significance and its corresponding null hypothesis was confused with an acceptance procedure that it seemed suitable to distinguish errors in which the hypothesis is rejected wrongly, from errors in which it is "accepted wrongly" as the phrase does. (Fisher, 1955, p. 73)

Rather than seeing Wald's work as a possibility for establishing mutual intelligibility, the key players on both sides of the debate thus rejected it as a misunderstanding that had perverted the philosophical purity of their own methodologies.

9.3.4 The Ideology of Statistics

Why did Wald's hybrid approach to statistics never catch on, and why did it never produce the truce that one might have expected? The answer, we are told by several of the participants in these debates, is that the discussion wasn't really about mathematics. It was about ideology.

In a preface to his criticism of the Neyman-Pearson approach, for instance, Fisher made the following revealing remark:

> I shall hope to bring out some of the logical differences more distinctly, but there is also, I fancy, in the background an ideological difference. Russians are made familiar with the ideal that research in pure science can and should be geared to technological performance, in the comprehensive organized effort of a five-year plan for the nation. How far, within such a system, personal and individual inferences from observed facts are permissible we do not know, but it may be safer, and even, in such a political atmosphere, more agreeable, to regard one's scientific work simply as a contributary element in a great machine,

> and to conceal rather than to advertise the selfish and perhaps heretical aim of understanding for oneself the scientific situation. In the U.S. also the great importance of organized technology has I think made it easy to confuse the process appropriate for drawing correct conclusions, with those aimed rather at, let us say, speeding production, or saving money. There is therefore something to be gained by at least being able to think of our scientific problems in a language distinct from that of technological efficiency. (Fisher, 1955, p. 70)

In other words, by computing probabilities of errors of the second kind (i.e., β), Neyman and Pearson had proven themselves to be not only crypto-Stalinists, but also nihilistic American capitalists. By contrast, Fisher's own British temperament, democratic sensibilities, and significance tests presumably inoculated him against such excesses.

Surprisingly, we also find Jaynes invoking the word "ideology" to describe the hidden motives of his opponents:

> ... the superiority of the orthodox method is asserted, not by presenting evidence of superior performance, but by a kind of ideological invective about 'objectivity' ... (Jaynes, 1983, p. 156)

His own position, of course, was entirely "non-ideological" (Jaynes, 2003, p. 14):

> Thus we continue to argue vigorously for the Bayesian methods; but we ask the reader to note that our arguments now proceed by citing facts rather than proclaiming a philosophical or ideological position. (Jaynes, 2003, p. xii)

De Finetti too seemed to think that the Bayesian school was the victim of an irrational mathematical witch hunt; he compared the resistance to expressing beliefs in terms of *a prior* distributions to a prohibition against expressing temperatures in degrees. What this leaves us with, he wrote, is "not ... less subjective but only less rational" (de Finetti, 1974, p. 127).

This rhetoric is difficult to explain if we try to understand the subject of this debate in purely technical terms, or, for that matter, purely aesthetic. The arguments put forward by both sides suggest that many of them felt that something morally fundamental was under threat, and that common decency required them to defend it. In order to understand what lies behind this intense emotional reaction, we need to take a step back and consider the wider historical and philosophical context of these mathematical theories of rational thought.

9.4 A History of Rationality

As the previous section has shown, the Bayesian approach to statistics was met with a surprisingly aggressive response from adherents of the classical or frequentist school, which saw the Bayesian methods as arbitrary and unscientific. The

Bayesians, in turn, adopted an equally aggressive rhetoric from the 20th century onwards, often depicting the frequentist methods as obscure, unprincipled, and defective.

How were things able to run so much off track? To answer this question, we need to put the specifics of the mathematical question aside; we need instead to think about what the participants in this discussion felt and still feel is at stake, and why this perception can produce such an acute sense of threat. This means that we need to put the recent history of rationality in a historical perspective.

9.4.1 The Formalist Theory of Reasoning

Since the turn of the 20th century, the mathematical study of reasoning has often been depicted as a narrow field concerned with a specific set of technical questions. Particularly since the publication of Kolmogorov's book on measure-theoretic probability (Kolmogoroff, 1933), the field has increasingly been presented as a disembodied mathematical discipline whose possible applications in reasoning and decision-making are entirely accidental. According to this formalist philosophy, a mathematical concept like "probability" is nothing more than an empty vessel that can be filled with any content we like: "love, law, chimney sweep, . . . " (Hilbert, 1980, p. 13).[1]

In his introduction to mathematical statistics, Cramér could thus announce that

> . . . largely owing to the work of French and Russian mathematicians, the classical calculus of probability has developed into a purely mathematical theory satisfying modern standards with respect to rigour. (Cramér, 1946, p. vii; see also pp. 145–146)

Doob made a similar remark in his textbook on stochastic processes:

> Probability is simply a branch of measure theory, with its own special emphasis and field of application, and no attempt has been made to sugar-coat this fact. (Doob, 1953, p. v)

Also in logic, many textbooks emphasize that they are concerned with the study of arbitrarily defined mappings between mathematical structures. Concepts like "meaning," "truth," and "proof" are therefore theory-internal notions with a purely technical meaning, as here explained in a relatively recent textbook:

> The very process of *speaking about*, establishing a connection between language and objects, then simply becomes the study of certain

[1] It is perhaps worth mentioning that Kolmogorov himself might not have shared this formalist view; see his discussion of the "relation to the world of experience" in Chapter. I, §2 (Kolmogoroff, 1933, pp. 3–5).

> functions which interpret — or map — the mathematical structure *language* into the semantical structures chosen to be spoken about. (Felscher, 2000, p. 1)

These conceptions of logic and probability theory have a lot of merit from a mathematical standpoint, but they give a somewhat false impression of where those disciplines come from historically. For most of its history, the principles of logic were not seen as arbitrary mathematical conventions, but were rather designed in a deliberate attempt to model the extra-mathematical concept of proper reasoning.

Only by throwing these rationalistic ambitions overboard, or by exporting them to neighboring disciplines, has a field like logic managed to recast itself as a formal study of mathematical patterns without "the slightest 'practical' importance" (Hardy, 1967, p. 101). The remainder of this section will be dedicated to documenting how historically novel this formalist self-image is, contrasting it with the explicitly extra-mathematical ambitions of the traditional approach.

9.4.2 The Social Function of Reasoning

Compared to the formalism of the 20th century, logic and probability theory played a very different role in the intellectual landscape of early modern Europe, both in terms of its content and its intended use. One early 18th-century theologian explicitly selected logic as the mark of "Rational Man," using the characteristically racist idiom of his day:

> I could venture to say, that the *Improvement of Reason* hath raised the learned and the prudent in the *European* World, almost as much above the *Hottentots*, and other Savages of *Africa*, as those Savages are by Nature superior to the Birds, the Beasts, and the Fishes. (Watts, 1726, p. 1; italics in original)

This point of view was also expressed in a book by the Dutch writer van Schurman (1659). She discussed the benefits of scholastic learning in an attempt to answer the question of "whether *a maid* may be a scholar," that is, "whether it be ... expedient, fit, decent" for her to study the letters (van Schurman, 1659, p. 2; all italics are in the original).

She answered this question in the affirmative; not for proto-feminist reasons, but because the quiet, chaste, domestic virtues associated with scholastic learning would be particularly well-suited for the female character. These qualities, she argued, could be cultivated by devoting oneself to the study of "those Arts which have neerest alliance to *Theology* and the *Moral Virtues*," among which she emphasized "especially *Logick*" (van Schurman, 1659, pp. 4–5).

The study of logic was thus not so much justified in terms of the specific contents of the curriculum as its ability to build one's character. This would be a lot to expect from something that was "simply a branch of measure theory," and is suggestive of the higher cultural significance bestowed on the theories of reasoning in this period.

Both of these examples also suggest a different usage scenario for an early modern textbook in logic. Rather than being an encyclopedic compilation of facts about a certain topic, such a book would expressly aim at the "Improvement of Reason," allegedly for the benefit of "both the greater and the meaner *Actions of Life*" (Watts, 1726, pp. 1–2).

Such a program statement is also made very overtly in the opening sentences of the Port-Royal *Logic*, which assert that "an exact reason is generally useful in all aspects and all walks of life." (Arnauld and Nicole, 1996, p. 5). Studying logic will thus make you a better person, helping you avoid not just scientific errors of judgment, but also

> ... the majority of mistakes committed in civil life: unjust quarrels, ill-founded lawsuits, hasty opinions, and badly organized enterprises. (Arnauld and Nicole, 1996, p. 6)

The utopia of logic, it would seem, was thus not a world of certain knowledge, but rather an orderly and efficient society under good governance. The negative counterpoint to "Rational Man" was not madness and error, but overconfident gambling and frivolous lawsuits (pp. 273–274).

This moral function of logic is also attested by several different 18th-century satirists and commentators. Budgell, an English politician, described in an 1711 issue of *The Spectator* what might have been a common experience of his day:

> Upon my calling in lately at one of the most noted *Temple* Coffee-houses, I found the whole Room, which was full of young Students, divided into several Parties, each of which was deeply engaged in some Controversie. ... In short, I observed that the Desire of Victory, whetted with the little Prejudices of Party and Interest, generally carried the Argument to such an Height, as made the Disputants insensibly conceive an Aversion towards each other, and part with the highest Dissatisfaction on both Sides. (Budgell, 1850, p. 291)

Rather than condemn this as sophomoric bickering, Budgell went on to volunteer some sound advice of his own to the young students:

> Avoid Disputes as much as possible. ... give your Reasons with the utmost Coolness and Modesty, two Things which scarce ever fail of making an Impression on the Hearers. Besides, if you are neither Dogmatical, nor shew either by your Actions or Words, that you are full of your self, all will the more heartily rejoice at your Victory. Nay, should you be pinched in your Argument, you may make your Retreat with a very good Grace: You were never positive, and are now glad to be better informed. (Budgell, 1850, p. 291)

By preserving "Coolness and Modesty" and never "falling into a Passion," the young men will thus exhibit a particular kind of maturity and self-control that would demonstrate just how favorably they compared to the "Savages" mentioned by Watts.

9.4.3 Self-Formation and Rationality

Rationality thus seems to have functioned to some extent as a cultural Shibboleth for the educated middle classes of the early modern period. In view of this function, it is also less surprising just how much the logic books from this period read like self-help books, designed to help the reader acquire a set of socially necessary skills.

The subtitles are a telling detail in this respect:

- *The Artes of Logike and Rethorike plainelie set foorth in the English tounge, easie to be learned and practised ...* (Fenner, 1584)
- *The Arte of Logicke. Plainly taught in the English tongue, according to the best approoued Authours. Very necessary for all Students in any Profession, how to defend any Argument against all subtill Sophisters ...* (Blundeville, 1619)
- *... a short Exposition of the Praecepts, by which any one of indifferent capacitie, may with a little paines, ataine to some competent knowledge and vse of that noble and necessary Science.* (Ramus, 1626)
- *An Introduction to the Art of Logick Composed for the use of English Schools, and all such who having no opportunity of being instructed in the Latine Tongue, do however desire to be instructed in this liberal science.* (Newton, 1671)
- *Logic, or, the Art of Thinking: in which Besides the Common, are contain'd many Excellent New Rules, very profitable for directing of Reason, and acquiring of Judgment, in things as well relating to the Instruction of a Mans Self, as of others.* (Anonymous, 1693)

In today's idiom, these titles could probably be translated as *Teach Yourself Logic in Three Weeks* or *How to Win an Argument.* In terms of content, however, a more pertinent comparison might be contemporary books on mindfulness or dieting. Note also how these early modern authors and publishers advertise their work by emphasizing how useful logic is in your personal and professional life, and how easy it has become to learn, now that you can read about it in English. This latter selling point in particular hints at the relative novelty of seeing books of this sort written for and by laypeople.

The self-help nature of this literature is particularly striking in the case of the *Art of Logick* (Coke, 1654; Measell, 1977). While this book does contain some theoretical parts, it also introduces the reader to a variety of practical exercises one can use in order to acquire the right kind of cognitive habits (Coke, 1654, p. 191–222). These include both "social" and "solitary" exercises (p. 209).

The social exercises include debating games or question-and-answer exercises that can be used to train the student's memory and reaction time. For instance, in one conversation game, two players select an appropriate proposition to debate and pledge to "bind themselves to the Laws and Rules of Logick" as well as "agree between themselves of certain foreknown principles" (Coke, 1654, p. 206). A role of "President" or moderator is also sometimes described (p. 208).

The solitary exercises include various kinds of meditations and reading exercises, consistent with the book's program statement:

> The exercise then of Logick consisteth in this, that we frequently think on, & diligently meditate things conformably to the prescriptions and rules of Logick, that is, orderly and distinctly... (Coke, 1654, p. 12)

In one meditation exercise, a concept is selected (e.g., *animal* or *purgatory*), and the student then has to construct a definition using a certain recipe, or sometimes to provide a superconcept, an antonym, or an etymology (p. 192ff). Other recommended exercises include a certain writing exercise in which the student reads a text, extracts its argument, and presents it in syllogistic form (p. 217ff), as well as the practice of regularly meditating on things learned so far (p. 212).

9.4.4 Meditations

We find the same idea in another much more famous example, the 1641 *Meditations on First Philosophy* by Descartes.

Although we are today inclined to read this book as a theoretical work of epistemology, Descartes actually explains in his preface that the book is intended to be used as a instruction manual: He wants to help the reader recreate his meditative experience, and he would "never advise anyone to read it excepting those who desire to meditate seriously with me" (Descartes, 1993, p. 40).

As was the case for *The Art of Logick*, Descartes also puts quite some emphasis on mnemotechnics. At several points, he demonstrates by example how one can reform one's cognitive habits through repeated exercise:

> ... although I notice a certain weakness in my nature in that I cannot continually concentrate my mind on one single thought, I can yet, by attentive and frequently repeated meditation, impress it so forcibly on my memory that I shall never fail to recollect it whenever I have need of it, and thus acquire a habit of never going astray. (Descartes, 1993, p. 79)

Or again, in the second meditation:

> But because it is difficult to rid oneself so promptly of an opinion which one was accustomed to for so long, it will be well that I should halt a little at this point, so that by the length of my meditation I may more deeply imprint on my memory this new knowledge. (Descartes, 1993, p. 58)

Such exercises in self-discipline are meaningless from a purely epistemological perspective. If the goal really were to construct a chain of valid inferences, the reasoner's personal ability to memorize the individual steps would be irrelevant. If, on the other hand, the goal is to cultivate certain habits of thought, it makes perfect sense to "halt a little" at important points and go over the same ground several times (Descartes, 1993, p. 58).

Also the structure of the book, with its accounts of daily meditations, has the character of a protocol for a ritual and reads much like Saint Anselm's 11th-century *Prayers and Meditations*. They also share the emphasis on concentrated, solitary introspection, as indicated by the incantation with which Anselm opens his first meditation:

> Awake, my soul, awake; bestir thy energies, arouse thy apprehension; banish the sluggishness of thy deadly sloth, and take to thee solicitude for thy salvation. Be the rambling of unprofitable fancies put to flight; let indolence retire, and diligence be retained. Apply thyself to sacred studies, and fix thy thoughts on the blessings that are of God. Leave temporal things behind, and make for the eternal. (Anselm, 1872, p. 18)

This practice of religious meditation had roots in the spiritual exercises of the early Christian and ancient Stoic traditions (Hadot, 1995, Chs. 3 and 4). But although it survived the middle ages almost exclusively within the confines of the monastic tradition (Bestul, 2012), the details surrounding Descartes' mediations clearly situate him in an unambiguously bourgeois setting:

> To-day, then, since very opportunely for the plan I have in view I have delivered my mind from every case and since I have procured for myself an assured leisure in a peacable retirement, I shall at last seriously and freely address myself to the general upheaval of all my former opinions. (Descartes, 1993, pp. 45–46)

Even his props seem chosen as if to signal this social affiliation:

> For example, there is the fact that I am here, seated by the fire, attired in a dressing gown, having this paper in my hands and other similar matters. (Descartes, 1993, p. 46)

So while some things point forwards, other things point backwards. Where does that leave Descartes? The best way to read his *Mediations* is probably as a transitional work: One hand, he continued an established tradition which attempted,

> ... by serious mediation, to draw nearer unto God, so to grow in knowledge, and in grace, and to increase in spirituall strength ... (Hinde, 1641, p. 141)

On the other hand, he broke with that tradition by emphasizing deduction over enlightenment, validity over piety, and knowledge over salvation. In this respect, he was at cross purposes with Christian writers like Pope Gregory, who had praised a fellow monk for using meditation "not so much for cunning and knowledge," but rather so "that he might with the wings of contemplation fly unto the kingdom of heaven" (Gregory I, 1911, p. 242). However similar the means, Descartes clearly had different ends.

In sum, it appears from both contextual and internal evidence that the early modern books on rational thought were meant to be used in a very different way from contemporary textbooks in mathematics. Rather than helping you acquire a specific pool of knowledge, they would teach you how to cultivate a certain set of personal qualities. We therefore misread the history of logic when we consider it as a catalogue of mathematical results, and it will often appear as startlingly naive when we do.

9.5 Two Paradigms of Rationality

As the examples in the previous section indicate, rationality is not merely a common biological trait like upright gait or opposable thumbs. Rational decision-making requires discipline and training, and any given society will have institutions and technologies designed to support its implementation of what it means to behave rationally.

By following the paper trail by this history of training and control, we are able to trace the changes that the norms of rationality have undergone over the course of history, comparing what the concept meant at different times. Such comparative history reveals that the ideals of rationality have not been static over time: The temperate and cautious decision-maker we met in the early modern textbooks differs markedly from the decisive warrior-king of the ancient world, or from the shrewd political strategist of the late middle ages.

Thus, while the use of logic in the 17th century at first seems to be a direct continuation of the medieval tradition, a closer inspection reveals subtle differences: Whereas the monastic tradition liked logic because it promoted the Benedictine virtues of patience, silence, obedience, humility, and modesty (Benedict, 1875), the early modern authors liked it for its tendency to promote restraint, self-discipline, patience, acuity, and good judgment. These value systems are not exactly opposites, but they are also not identical. The purpose of a history of rationality is to notice and map out such differences.

This can be difficult, both because the changes are slight and uneven, and because of the constant temptation to discard important details because they seem, in retrospect, philosophically irrelevant. For instance, when Plato in the *Republic* recommends geometry classes and debating games (526c–527c, 534d–535a), we probably recognize that as a vision of an ideal rationality; but when he recommends gymnastics and military experience (537a–b), we might shrug it off as mere politics. In order to be able to write a history of rationality, we need to read in a way that does not force the source material into such "neat little pots" (Latour, 2005, p. 141).

The explanation I have in mind for the division in statistics is an example of such a history which blurs the line between epistemology and social organization. I will describe two ideal types which have been dominant at different times in the history of rationality, the "bureaucratic" and the "entrepreneurial" rationalities. The central claim of this chapter is that these two rationalities are the governing principles behind the two major traditions in the history of statistics. The division between the frequentist and Bayesian approaches are the mathematical expression of a division between two different rationalities. The conflict between the two schools is a philosophical conflict that never announced itself as such.

This story is inevitably going to gloss over a wealth of details, both about the interaction and dispersion of these rationalities, and about the spectrum of different approaches to statistics. I recognize that there is a danger of oversimplification. It will, however, be helpful to start with the caricature and then return to the complications later. In subsection 9.5.4, I will briefly discuss some problems related to the simplified narrative.

9.5.1 Bureaucratic Rationality

The older of the two rationalities I want to discuss is bureaucratic rationality. This is the ideal which was popularized in the early modern period and created by the somewhat surprising appropriation of monastic practices by the increasingly secular middle classes.

The characteristics of this personal and epistemological ideal is an emphasis on methodic and reproducible protocols of decision-making. The values that are prototypically associated with it are rigor, accuracy, accountability, responsibility, reliability, caution, diligence, thoroughness, meticulousness, and restraint.

If we were to choose a patron saint of this brand of rationality, we might choose Pacioli, the Italian monk who popularized the technique of double-entry bookkeeping. This accounting practice, a self-imposed regime of regularly repeated writing, embodies many of the ideals that come with bureaucratic rationality. As Pacioli explains, the goal is to supply the readers with

> … enough rules to enable them to keep all their accounts and books in an orderly way [*ordinatamente*]. For, as we know, there are things needed by any one who wishes to carry on business carefully [*diligentia*]. (Geijsbeek, 1914, p. 33)

By submitting oneself to this regime of self-examination and bookkeeping, one will thus be able to "arrange all the transactions in such a systematic way [*con bello ordine*] that one may understand each one of them at a glance" (Geijsbeek, 1914, p. 33).

With its curious position between medieval monasticism and early modern commerce, this system of "beautiful order" also illustrates how the new bureaucratic rationality managed to appropriate ancient religious practices. This is reflected in Pacioli's work when, for instance, his very practical discussion of how to draw up an inventory suddenly veers off into moralistic tirade about Christian obligations, admonishing the reader to "never forget to attend to religious meditation every morning" (Geijsbeek, 1914, p. 37).

9.5.2 Entrepreneurial Rationality

Outside the monasteries, bureaucratic rationality seems to date back to the turn of the 17th century. The entrepreneurial rationality, on the other hand, is a more recent invention, dating back only to the late 19th century. I will not attempt to document the emergence of this competing rationality in any detailed way here, but I will try to describe its features, deferring the discussion of its origins to another time.

Entrepreneurial rationality is the ideal native to high-tech capitalism, polytechnic education, and the integration of natural science into commercial enterprise. The stereotypical image of the genius inventor, tirelessly working on his latest piece of techno-magic, is closely related to this ideal of rationality. Rather than the reliable desk clerk, its hero is the bold, creative, original individualist with visionary, novel ideas.

A good representative of this ideal is Edison, who did more than anybody to craft the concept of the genius inventor in his lab. Inventing something, in this sense of the word, means doing something for which there is little precedent, gambling on future profits and technological possibilities:

> Quarter horse, half horse, one horse and five-horse machines, he anticipates, will be in great demand for keeping ventilators in motion, swinging fans in restaurants, running sewing machines and turning lathes, and so on. (*Brookport Republican*, 23 September 1880).

Being such a visionary, "popular prejudices and customs" were of course, in Edison's own words, a major obstacle, and he would therefore take a particular pleasure in demonstrating how he had "already done what the advocates of gas say has not been done, and what they do not believe can be done" (*Watertown Re-Union*, 31 October 1878).

On the negative side, the nightmare that haunts this techno-capitalism is not irresponsibility, but failure. Edison was frequently suspected of selling castles in the air, especially by his competitors in the gas industry (*Watertown Re-Union*, 31 October 1878).

He would respond by exhibiting computations, experiments, and prototypes to prove the feasibility of his scheme, at one point boasting of "over a thousand sheets of drawings alone" (*The Sun*, 25 November 1878). "This is no Keely motor business out here," he told one journalist (*The Sun*, 24 January 1880), with a back-handed reference to John Keely, an infamous inventor who repeatedly failed to deliver on the new, mysterious source of energy that his investors and stockholders had been promised.

Invention is therefore always a gamble, a bet on a certain future society. The inventor has as yet nothing to show for the feasibility of his wonderful machine, and he can only hope that time will prove him right, and prove the conservative naysayers wrong.

9.5.3 Rationality and Bureaucracy

The most important reference point for any history of rationality is Weber's classic study *The Protestant Ethic and the Spirit of Capitalism* (Weber, 1992). Some words will therefore be in order about the relation between the narrative I am pursuing here, and the theory proposed by Weber.

According to Weber's analysis, the ascetic work ethic of Protestantism paved the way for much of what we today recognize as capitalist values. This turns the Marxist order of explanation on its head, claiming that the new economy could never have gotten off the ground without the stable and predictable system of bureaucracy to support it:

> For modern rational capitalism has need, not only of the technical means of production, but of a calculable legal system and of administration in terms of formal rules. (Weber, 1992, p. xxxviii)

According to Weber's analysis, this situation was brought about when the early Protestant sects started to rework their Catholic heritage into a new and original work ethic. Under this radicalized religious system, formerly religious practices would be applied in all spheres of life:

> For when asceticism was carried out of monastic cells into everyday life, and began to dominate worldly morality, it did its part in building the tremendous cosmos of the modern economic order. (Weber, 1992, p. 123)

The argument is here not so much that a certain "idea" or "mentality" reached into the material sphere in order to change the course of history. It is rather that new kinds of discipline and organization may be a precondition for new forms of production: For instance, if you want to run a factory according to a fixed timetable, you need to teach your workers to tell the time. If you want accurate public health statistics, you need a caste of reliable government clerks to collect the numbers. Bringing those conditions about is a matter of social reform, and it is not always clear whether to count these factors as "ideal" or "material."

I agree that a new form of rationality took form about the time Weber focused on, and that this change involved a novel use of old religious traditions. I diverge from Weber on a couple of points, however.

First, I apply the concept of rationality in a more general sense, rejecting the equation between bureaucracy and rationality. When Weber speculates that the bureaucratic work ethic will dictate economic behavior "until the last ton of fossilized coal is burnt" (Weber, 1992, p. 123), I think he is making the mistake of seeing rationality as a raw material rather than a system of behavioral norms. He thus comes to depict the process of "rationalization" as, in a certain ominous sense, the end of history. Whether he actually intended to make this claim is unclear (cf. Weiss, 1987), but his text certainly invites such a reading.

Second, as has already become apparent, I emphasize the link with Protestantism less. The changes in rationality that I describe took place in England, France, Belgium, Germany, Italy, and the Netherlands. These countries and territories experienced the Reformation in very different ways, and they adopted or interacted with radical puritanism to very different degrees. It is not clear that the regional differences in religious practice correspond reliably to differences in the conception of rational thought and behavior. A more detailed study of the roots of these secular practices could perhaps clarify what exactly the connection is, but I have instead chosen to put less weight on the historical connection to Christianity.

9.5.4 Conflict and Co-Existence

If you graduate with a degree in engineering from the University of Toronto, you will be invited to participate in a ceremony called "The Calling of an Engineer" (Wedel, 2012; Roddis, 1993). This ceremony is held once a year, the most recent one being on 7 March 2015. It takes place at the university's Convocation Hall, an imposing building with a domed roof, roman columns, and a huge organ built into the back wall.

During this ceremony, you will read a text in which you pledge to devote yourself to proper craftsmanship and to uphold a high professional standard:

> My time I will not refuse; my thought I will not grudge; my care I will not deny towards the honour, use, stability and perfection of

> any works to which I may be called to set my hand.

You end this pledge by

> ... praying, that in the hour of my temptations, weakness and weariness, the memory of this my Obligation and of the company before whom it was entered into, may return to me to aid, comfort and restrain. (Wedel, 2012, p. 118)

After thus being sworn into the engineering profession, you are given a ring which you are supposed to carry as a symbol of your devotion to the vocation. This iron ring is deliberately austere in design so as to remind you of the potentially catastrophic consequences of sloppy workmanship. In fact, popular mythology has it, the first batch of rings were made from iron salvaged out of the rubble of the Quebec bridge, which had experienced two catastrophic collapses, in 1906 and 1917, both due to engineering errors (Roddis, 1993; Evans, 2011). Since you wear the ring on the little finger of your working hand, its rough edges will scratch against the paper as you work, quietly reminding you of the pledge you once gave after your graduation. Not surprisingly, the engineering societies which administer the ritual are at some pains to explain that its purpose is "integrity and ethics, *not* any specific religious or political agenda" (Evans, 2011, slide 20).

Canadian engineers can thus, in a quite literal sense, marry their profession. This devotional attitude can be contrasted with the philosophy found at the Massachusetts Institute of Technology, not too far from Toronto.

If you took your engineering degree at that university, you might have followed a course called "Street-Fighting Mathematics" with professor Mahajan. According to the title of the textbook he wrote for this course, it taught you "the art of educated guessing and opportunistic problem-solving" (Mahajan, 2010). That amounts to not worrying too much about what might go wrong:

> Too much mathematical rigor teaches *rigor mortis*: the fear of making an unjustified leap even when it lands on a correct result. Instead of paralysis, have courage—shoot first and ask questions later. Although unwise as a public policy, it is a valuable problem-solving philosophy ... (Mahajan, 2010, p. xiii)

This style of reasoning has occasionally been the subject of parody as well as fascination: "Anything I can throw weighs one pound. One pound is one kilogram" (Munroe, 2014, p. 48). This exaggerated example is, however, not that far from the recommendation actually given by Weinstein and Adam (2009), who open a chapter called "How to solve problems" with the instruction "STEP 1: Write down the answer" (p. 2). While we might not want public officials to shoot from the hip, it can thus be a mark of real power in an engineer to know how to get a quick ballpark figure, even at the cost of validity and reliability.

This power derives from “the ability to make rough approximations, inspired guesses, and statistical estimates from very little data” (Morrison, 1963, p. 627). The statistics that goes with this skill is of course Bayesian. If you took professor Gallager’s 2011 course on stochastic processes while you studied electrical engineering at MIT, you were told so explicitly. Prior probabilities, he explained in one lecture, are neither better nor worse than any other modeling assumption:

> You use models to try to understand certain things about reality. And you assume as many things as you want to assume about it. And when you get all done, you either use all the assumptions or you don’t use them. (Gallager, 2011, p. 5)

They might be wrong, of course, but so might everything. The only cure against bad assumptions is better assumptions: “We can change the model. We can do whatever we want with a model” (p. 5).

As these examples show, the two major rationalities are not uniformly dispersed, and neither is completely dominant even in a relatively homogenous field like engineering. Often a single business or a single institution will contain branches and people that represent different epistemic virtues.

However, broader tendencies do exist: Entrepreneurial rationality has been gaining territory since, just to pick a number, the founding of MIT in 1861. No one would have published a book in the year 1700 explaining how to solve math problems “on the back of a cocktail napkin” (Weinstein and Adam, 2009); and most 17th-century natural philosophers would have frowned at the suggestion that “We can do whatever we want with a model.” This is a philosophy that belongs to the era of J. P. Morgan, not the Bank of Amsterdam.

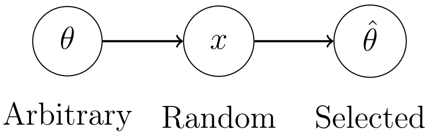
9.6 Estimates and Estimators

How do the two paradigms of rationality discussed above relate to statistics? To get a better sense of this question, it will be useful to focus on a representative case study. In the following, I will therefore primarily talk about the problem of parameter estimation, an inference problem which is particularly well-suited to bring out the differences between the two schools.

9.6.1 What is Estimation?

In statistics, estimation means roughly the same as “guessing.” Estimation is the problem of guessing a number when we don’t actually have enough information to identify it (Hodges and Lehmann, 1964, Ch. 8). This contrasts with related but slightly different problems like hypothesis testing, regression, density estimation, confidence interval construction, and so on.

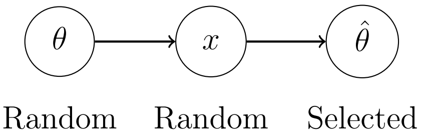
In a frequentist setting, the process of estimation can be modeled by assuming that we are given a sample x drawn from one of the distributions in an indexed family $\{P_\theta \,|\, \theta \in \Theta\}$. We are then interested in computing some quantity $f(\theta)$ which depends on θ. Often, this quantity is a parameter of the distribution P_θ, like the mean of a Gaussian. In that case, $f(\theta) = \theta$, and our goal is to use the data set x to pick a good guess $\hat{\theta}(x)$ estimating the value θ:



In the frequentist setting, the index or parameter θ is imagined to be fixed although unknown, and the data set x is supposed to follow the distribution P_θ. Since the data set x is random, any function of x will be random too. When an estimation problem is repeated twice, holding P_θ fixed but letting x vary randomly, the two data sets may thus produce different estimates $\hat{\theta}(x)$. For instance, if a coin with bias θ is flipped $n = 10$ times, then the frequency $\hat{\theta}(k) = k/n$ will typically vary between data sets even though θ is held fixed.

Evaluating an estimation function $\hat{\theta}$ is thus a matter of tracking properties of the probability distribution of the random variable $\hat{\theta}(x)$. We might, for instance, be interested in measuring how far the mean $E[\hat{\theta}]$ is from θ, or how probable it is to exhibit a deviation of $|\hat{\theta} - \theta| > \varepsilon$ for a given $\varepsilon > 0$. Since $\hat{\theta}(x)$ is a well-defined random variable, these questions can be answered precisely for every particular choice of θ. We might therefore hope to find a function $\hat{\theta}$ which has uniformly good properties for all choices of θ.

From the Bayesian perspective, such a comparison across data sets but within parameter settings is nonsensical. In a Bayesian model, uncertainty can only be modeled as randomness, so the pair (x, θ) is imagined to be sampled from a single joint distribution. For each individual estimation problem, the statistician thus selects a particular prior probability distribution for θ. After observing x, the posterior distribution over θ can then be used to evaluate various proposed guesses $\hat{\theta}$ in terms of average distances, error probabilities, or other measures of average loss:



By imposing a prior probability distribution on θ, this methodology allows the statistician to precisely compare the relative importance of various undesirable outcomes. By design, this approach cannot include any robustness analysis; since θ is assumed to be a random variable with a known distribution, there is no need to consider the consequences of particular, malicious choices of θ. The only way to hold θ fixed is to make it known, an operation that would obviously obviate the need for an estimation procedure.

The mathematical differences between the two methodologies are thus relatively clear in this context. The frequentist approach recommends estimators $\hat{\theta}$ that lead to low average-case loss for all choices of θ (averaging over x); the Bayesian approach recommends estimates $\hat{\theta}$ that have low average-case loss given the x we actually observed (averaging over θ). In the following, I will explain how this mathematical difference relates to the philosophical distinction I drew in the previous section.

9.6.2 The Philosophy of Estimation

A first hint at the philosophy behind the frequentist method of estimation is given by Kendall in his encyclopedic textbook on frequentist statistics:

> Now a single sample, considered by itself, may be rather improbable, and any estimate based on it may therefore differ considerably from the true value of θ. It appears, therefore, that we cannot expect to find any method of estimation which can be guaranteed to give us a close estimate of θ on every occasion and for every sample. We must content ourselves with formulating a rule which will give good results "in the long run" or "on the average," or which has "a high probability of success" — phrases which express the fundamental fact that we have to regard our method of estimation as generating a population of estimates and to assess its merits according to the properties of this population. (Kendall, 1951, p. 1)

Somewhat surprisingly from a naive perspective, what matters in estimation is thus not getting the correct answer; rather, the central concern is with the design of reliable protocols, rules, or methods. Even stronger yet:

> Our problem is not to find estimates, but to find Estimators. We do not reject a method because it gives a bad result in a single case (in the sense that the estimate differs materially from the true value). We should only reject it if it gave bad results in the long run, that is to say, if the population of possible values of the estimator were seriously discrepant with the value of θ. (Kendall, 1951, p. 2)

What philosophy underlies this emphasis on long-term effects? A quite explicit answer is given by Fisher, who opens his book on experimental methods by explaining that the purpose of statistics is to provide protection:

> In the foregoing paragraphs the subject-matter of this book has been regarded from the point of view of an experimenter, who wishes to

> carry out his work competently, and having done so wishes to safeguard his results, so far as they are validly established, from ignorant criticism by different sorts of superior persons. (Fisher, 1947, p. 3)

This suggests that the frequentist methodology should be seen less as a method for learning the truth and more as a kind of protective rite. We follow a statistical procedure because it is the better habit, not necessarily because we believe it gives the right result in a particular case. From a anthropological perspective, we might describe it as a "strategic ritual," an act that allows you to say that you made your choices for the right reasons even when they have the wrong consequences.

9.6.3 The Integrity of an Estimator

The same procedural interpretation is also expressed indirectly by other authors. Rudolf Carnap, who in spite of his Bayesian tendencies was extremely suspicious of unjustifiable subjectivity, would thus explain by a moralistic analogy how one ought to make proper inferences:

> When we wish to judge the morality of a person, we do not simply look at some of his acts; we study rather his character, the system of his moral values ... Similarly if we wish to judge the rationality of a person's beliefs, we should not look simply at his present beliefs. ... We must rather study the way way in which the person forms his beliefs on the basis of evidence. (Carnap, 1971, p. 22)

Again, process is thus emphasized over product; what distinguishes the rational belief is the use of certain rational habits of belief formation. We are thus not far from the early modern admonition to "frequently think on, & diligently meditate things conformably to the prescriptions and rules of Logick, that is, orderly and distinctly" (Coke, 1654, p. 12).

This ritualistic conception of truth is closely related to the concept of bureaucratic rationality, and specifically to its implementation in common-law reasoning. In his *Commentaries on the Laws of England* Blackstone explicitly cited possible future consequence as a criterion for evaluating a decision policy:

> ... however *convenient* these [trials without jury] may appear at first, ... let it again be remembered, that delays, and little inconveniences in the forms of justice, are the price that all free nations must pay for their liberty in more substantial matters; ... the precedent may gradually increase and spread ... (Blackstone, 1769, p. 344)

Rather than having the judge focus primarily on the contingencies of the particular case when making a decision, common-law reasoning thus shifts the focus to an infinite series of future cases. As Blackstone famously remarked,

> . . . all presumptive evidence of felony should be admitted cautiously: for the law holds, that it is better that ten guilty persons escape than that one innocent suffer. (Blackstone, 1769, p. 352)

In this sense, the emphasis of common-law reasoning closely resembles the frequentist concern about the long-term effects of a fixed policy, even though the focus is on implicit precedent rather than explicit rules. Or to paraphrase Mahajan (2010, p. xiii), the best problem-solving approach may not be the best public policy.

Of course, the imperative to control one's impulses and govern one's thoughts and inferences are central to all conceptions of rationality, not just the bureaucratic ideal. However, the types of impulses that are suppressed and cultivated, and the language in which they are described, often betrays a more specific allegiances. Consider for instance Edwards' praise of Fisher's approach to statistics, which according to him

> . . . allows us to do most of the things which we want to do, whilst restraining us from doing some things which, perhaps, we should not do. (Edwards, 1972, p. 212)

If this is not explicitly moralistic enough, consider Fisher's characterization of the ideal researcher:

> . . . the criteria by which [a hypothesis] is approved require a certain honesty, or integrity, in their application. (Fisher, 1955, p. 75)

Integrity, honesty, and restraint are not random selections from the menu of positive qualities one can cite. They reveal a notion of rationality, an ideal embodied by the reliable, predictable, meticulous clerk who makes every decision in an irreproachable manner. This ideal is not universal, but represents a highly particular implementation of rationality, intimately connected with "the heroic age of capitalism" (Weber, 1992, p. 67).

9.6.4 Does Rationality Play Dice?

Against the ideal of rigorous principles of mechanical thought, the Bayesian tradition has proposed a "Rational Man" who acts on what the data so clearly states without worrying about precedent or worst-case scenarios.

In one sense, this understanding of Bayesian statistics goes all the way back to Laplace's remark that probability theory amounts to nothing more than "common sense reduced to a calculus" (Laplace, 1995, p. 124). But the same intuition was expressed by other authors as the debates about the proper foundations of statistics started heating up in the late 19th century. In an 1893 lecture, the Scottish actuary Govan thus responded to Chrystal's fierce attack on Bayesian methods with the following example:

> … let us say we have two bags before us, one containing six white balls, the other five black balls and one white. There is nothing to indicate which is which. We draw from one of the bags chosen at random a ball which proves to be white. It is difficult to believe that any man in the possession of his faculties, say if his life depended on his guessing aright from which bag the ball had come, would hesitate to guess the former. (Govan, 1921, p. 209)

The scenario is characteristic of many of the Bayesian counterattacks on the frequentist critiques: By placing the inference problem in the context of a one-shot gambling situation, the emphasis is shifted from the concern about the infinite repetition of a potentially catastrophic decision to the potentially catastrophic consequences about worrying too much about the future when your problems are in the present.

Jaynes invoked a similar argument, using Laplace's computation of the mass of Saturn as an example:

> Laplace … announced his result as follows: "… it is a bet of 11000 against 1 that the error of this result is not 1/100 of its value." In the light of present knowledge, Laplace would have won his bet; another 150 years' accumulation of data has increased the estimate by 0.63 percent. (Jaynes, 1983, p. 156)

This might be so, but it is typical of the debate, and of Jaynes' writing in particular, that he fails to realize that this is not a relevant criterion of success for the competing school of thought.

9.6.5 The Frequency Controversy

A similar contrast comes up in relation to the question of how to interpret the concept of probability on a philosophical level: What does it really mean that the probability of a coin coming up heads is one half?

Authors like Venn and von Mises were adamant about maintaining a strict frequentistic interpretation of this concept. Under this interpretation, the probability $P(A)$ of an event $A \subseteq \Omega$ must always admit of an interpretation in terms of a limiting frequency of the event $A^{\mathbb{N}}$ in an infinitely repeated experiment, $x \in \Omega^{\mathbb{N}}$. From such a perspective, probabilities are, strictly speaking, inapplicable to individual propositions such as the mass of Saturn being or not being in a certain range. Von Mises explains:

> We can say nothing about the probability of death of an individual even if we know his conditions of life and health in detail. The phrase 'probability of death', when it refers to a single person, has no meaning at all for us. (von Mises, 1957, p. 11)

This philosophical stance would puzzle and annoy many writers in the Bayesian school. Jeffreys, for instance, would complain that such a conception of probability had "no practical application whatever," since

> When an applicant for insurance wants to choose ... a policy ... his probability of living to 65 is an important consideration. The limiting frequency in an infinite series of people is of no direct interest to him. (Jeffreys, 1973, p. 195)

For "practical men," Jeffreys argued, such a speculative notion of probability would thus do none of what they wanted it to do.

But his restriction is not an arbitrary one, von Mises would say; indeed, a probability that did not express a frequency would be an empirically meaningless number, even if it were derived using a formalism that happened to be mathematically coherent:

> The theory of probability cannot be applied to this problem any more than the physical concept of work can be applied to the calculation of the 'work' done by an actor in reciting his part in a play. (von Mises, 1957, p. 15)

This distinction between probabilities as frequencies and probabilities as degrees of belief has received much attention in the philosophical literature because it can be formulated as a question of metaphysics. If I am correct in the hypothesis that I am proposing here, however, this metaphysical dispute is merely the shadow of a more deeply rooted discrepancy in the conception of what would count as a legitimate rational thought.

9.6.6 The Thunderbolt of Truth

In the frequentist theory of estimation, the statistician vaccinates him- or herself against criticism by using an unassailable decision mechanism that does not make any reference to the particulars of the situation or the decision-maker. This approach involves a reliance on worst-case reasoning and hypothetical repetitions of the same problem.

The Bayesian methodology, on the other hand, freely makes reference to subjective hunches, worries little about the worst case, and aims at getting the best out of the situation at hand given all the information available. Not surprisingly, these two paradigms tend to disagree about proper protocol and even the criteria by which results should be judged.

In this section, I have argued that these disagreements are a symptom of a very deeply rooted difference in the underlying ethics. Different conceptions of the ideal rational person tend to suggest different mathematical methods and models, and it is hard to motivate or defend these choices without revealing the philosophical ideas behind.

Whether you find statistics ought to investigate the long-term effects of behavioral policies or squeeze all the information possible out of a singular event will depend on your sympathies. Both possibilities implement, in the form of a mathematical calculus, a relationship with the truth.

This also indicates that rationality did not freeze over when the early modern period matured into the American and French revolutions. It continued to adapt, and in some ways Bayesian statistics seems to have revived parts of the medieval conception of truth as a precious object you have to wrestle from the world by a bold show of willpower. You win the game against truth by exhibiting superior computations rather than by going through protective rituals.

This kind of relationship with the truth is an example of what Foucault suggested might have been the precursor of the early modern scientific truth:

> We could call this discontinuous truth the truth-thunderbolt, as opposed to the truth-sky that is universally present behind the clouds. We have, then, two series in Western history of truth. The series of constant, constituted, demonstrated, discovered truth, and then a different series of the truth ... which is not found but aroused and hunted down: production rather than apophantic. ... This kind of truth does not call for method, but for strategy. (Foucault, 2006, p. 237)

For the desk clerk with clearly defined responsibilities and possibilities of promotion, such an aggressive relationship with the truth is risky, frivolous, and not worth the cost. For the ambitious high-tech entrepreneur, on the other hand, the ritualistic relationship to the truth is merely a fossilized version of an originary moment of inspiration.

It would be a mistake to think that we are in a position to evaluate which of the two sides in this conflict have the stronger hand. What our current situation calls for is neither nostalgia nor glee, but a realistic understanding of the tectonics of rationality as the ground shifts beneath our feet. In a commentary on the transition from disciplinary to post-disciplinary forms of control, the philosopher Deleuze commented:

> There is no need to ask which is the toughest or most tolerable regime, for it's within each of them that liberating and enslaving forces confront one another. ... There is no need for fear or hope, but only to look for new weapons. (Deleuze, 1992, p. 4)

Weapons or not, this holds for the changes in rationality as well. Once we understand that the two paradigms of statistics are the mathematical expression of two different rationalities, we will get beyond the inclination to ask "Which one got it right?"

9.7 Conclusion

According to a persistent myth, the recent increase in the popularity of Bayesian statistics is due to "the development of fast computers and readily available software" (Van Dongen, 2006, p. 91). According to this myth, the Bayesian methodology was held back for many years by the lack of efficient techniques for solving important subproblems such as numerically evaluating high-dimensional integrals (cf. Box and Tiao, 1973, p. 65; Krauss et al., 1999, p. 166; Link and Barker, 2009, p. 8).

There is no doubt that certain complex models, Bayesian or otherwise, have become computationally feasible only after the advent of fast computers. However, this development does not in itself explain the philosophical transition that we are currently seeing, for several reasons:

- Computational problems were never a central objection, even though they occasionally were an obstacle. As we have seen, even Bayes' simple coin flipping model with a flat prior was subject to suspicion and criticism.

- At least as far back as the early work of "Student" (1908), frequentist inference too has involved a quite substantial amount of computational complexity, often including integration problems that do not admit of closed-form solutions.

- Many frequentist methods, including bootstrapping and support vector machines, are "prodigious computational spendthrifts" (Efron, 1982) and would not have been possible without an increase in computing power.

- In terms of the chronology, this explanation also fails to explain what might have motived people like Jeffreys and de Finetti to advocate the use of Bayesian methods at a time when the first electric computers were still only a couple of years old.

To the extent that the Bayesian approach is on the rise at the moment, there is thus something more going on than the mere fact that certain previously unsolvable integrals are now solvable. Hearts and minds have changed too.

Technology has certainly played a role in this process, but not in the direct way that the numerical-integration myth supposes. Rather, the most prominent effect of the last century of information technology has been a redefinition of what it means to make rational decisions:

> Technology expands the potential reach, scale, and monitoring capacity of a decision-maker by magnifying the potential consequences of his or her choices. Direct management via digital technologies makes a good manager more valuable than in earlier times, when executives had to share control with long chains of subordinates and could affect only a smaller range of activities. (Brynjolfsson et al., 2014, p. 51)

In other words, technology is making bureaucratic power largely obsolete. Rather than training a caste of clerks to act like machines, we can now employ actual machines. The bottleneck in production, surveillance, and control is therefore no longer the systematic collection of reliable data, but rather the competitive effort to extract profits from the oceans of data that digital record-keeping leaves behind.

These are changes in our decision-making practices, and it is not surprising that they have been accompanied by changes in our norms of rationality. Throughout the 20th century, the dominance of frequentist statistics has been increasingly challenged by the Bayesian alternative. In fields like biology or psychology, which were once bastions of the frequentist paradigm, there are now vocal minorities that favor Bayesian approaches. This trend will likely continue as long as the social norms of the surrounding society do.

One day we may wake up and have trouble remembering what all the fuss was about. Why did we worry so much about the specter of subjectivity? How did we ever come believe that the alternative to universalism was anarchy? Why did we think it was such a deep and serious problem if people formed different beliefs from identical evidence? And when did we stop caring?

The emergence of such a consensus is likely to be followed by a smug sense of progress: We once were lost, but now are found. As I have tried to argue in this chapter, however, such an understanding of the situation would mistake history for destiny. Rationalities can compete, but they cannot argue. When the very legitimacy of your mode of argumentation is under threat, there is no way to defend yourself by means of argument.

Summary

In this dissertation, *The Kid, the Clerk, and the Gambler*, I have presented a series of case studies in linguistics, psychology, and statistics. These case studies have taken up a variety of theories, concepts, and debates, and have in each case attempted to shed new light on these topics by consistently focusing on foundational issues. What these foundational issues were has depended on the specific topic at hand.

In my discussion of cognitive metaphor theory in Chapters 2 and 3, I thus took issue with the strong emphasis on direct, bodily experience and attempted to show that this physicalistic origin of linguistic behavior should be seen as one among many sources, and probably not the most important one. In my discussion of the syntax-semantics distinction in Chapter 4, on the other hand, I criticized an overly logical "filing cabinet" theory of comprehension and argued that much of the EEG evidence supporting it could be explained more elegantly in terms of probabilistic reasoning and noisy-channel coding.

The next series of case studies combined various conceptual frameworks that are usually seen as separate. In Chapter 5, I thus defined a quantitative concept of relevance based on classical ideas from information theory and decision theory, and I hinted at ways it could be applied to situations in computer science and microeconomics. In Chapters 6 and 7, I selectively transplanted components from epistemic logic into a Bayesian reasoning system in order to construct a flexible modeling language which might have applications in cognitive science and artificial intelligence.

The last two case studies focused on issues regarding the philosophy of statistics and the limits of rationality. In Chapter 8, I discussed the bias-variance trade-off, a conventional way of thinking about the "loading" of statistical estimators, and I explained why this trade-off had no coherent interpretation in Bayesian statistics, even though it seems, from the outside, to support certain tenets of that methodology. In Chapter 9, sketched a history of the protracted war between frequentist and Bayesian statistics and reinterpreted this debate as a conflict between two norms of rationality expressed in mathematical form.

Samenvatting

Dit proefschrift, *Het Kind, de Klerk, en de Gokker*, bevat een reeks aan casestudies uit de taalkunde, psychologie en statistiek. Ik heb geprobeerd om een nieuw licht te werpen op deze onderwerpen door mijn focus consequent op fundamentele methodologische en filosofische aspecten te houden. Deze casestudies zijn op verscheidene theorieën en begrippen gericht, afhankelijk van de specifieke thema's.

In mijn bespreking van cognitieve metafoortheorie in Hoofdstuk 2 en 3 heb ik de sterke nadruk op directe, lichamelijke ervaring ter discussie gesteld, en heb ik geprobeerd aan te tonen dat deze lichamelijke bron van taalgebruik moet worden gezien als één van meerderen, en waarschijnlijk niet de belangrijkste. In mijn bespreking van het onderscheid tussen syntaxis en semantiek, aan de andere kant, heb ik een overdreven logische "archiefkast-theorie" van taalbegrip bekritiseerd. In Hoofdstuk 4 heb ik dus laten zien dat een groot deel van de EEG-data die vaak wordt aangevoerd als bewijs voor de psychologische realiteit van deze concepten beter kunnen worden verklaard in termen van probabilistische redenering en verwante begrippen uit de wiskundige communicatietheorie.

De daarop volgende reeks aan casestudies zijn gebruikt om verschillende conceptuele kaders samen te brengen die normaliter vaak als apart worden gezien. In Hoofdstuk 5 heb ik een kwantitatief begrip van relevantie gedefinieerd, gebaseerd op klassieke ideeën uit de informatietheorie en besliskunde, en heb ik mogelijke toepassingen in de informatica en in de micro-enonomie aangeduid. In Hoofdstuk 6 en 7 heb ik concepten uit de epistemische logica vertaald naar een Bayesiaans redeneringssysteem dat kan worden toegepast om sociale fenomenen te modelleren in de cognitieve wetenschappen of in de kunstmatige intelligentie.

De laatste twee casestudies richten zich op kwesties in de filosofie van de statistiek en de grenzen van de rationaliteit. In Hoofdstuk 8 heb ik een populaire manier om de "geijktheid" van een schattingsmethode te kwantificeren besproken, de splitsing van schattingsfouten in onzuiverheid en variantie. Ik heb uitgelegd waarom deze analyse geen coherente interpretatie heeft in de Bayesiaanse statistiek, ook al lijkt het, van buiten af, zekere aspecten van deze methodologie te ondersteunen. In Hoofdstuk 9 heb ik een overzicht geschetst van de langdurige oorlog tussen de frequentistiche en de Bayesiaanse statistiek, en ik heb dit conflict proberen te verklaren als een geschil tussen twee normen van rationaliteit, uitgedrukt in wiskundige vorm.

Bibliography

Lucía Amoruso, Carlos Gelormini, Francisco Aboitiz, Miguel Alvarez González, Facundo Manes, Juan F. Cardona, and Agustin Ibanez. N400 ERPs for actions: building meaning in context. *Frontiers in human neuroscience*, 7, 2013.

John R. Anderson. The Place of Cognitive Architectures in a Rational Analysis. In Kurt VanLehn, editor, *Architectures for Intelligence*, pages 1–24. Psychology Press, 1991.

J. Yu Angela and Peter Dayan. Uncertainty, Neuromodulation, and Attention. *Neuron*, 46(4):681–692, 2005.

Anonymous. *Logic; or, The Art of Thinking.* T. B., London, UK, 1693. This is an English translation of the Port-Royal *Logic* (Arnauld and Nicole, 1996).

St. Anselm. *Book of Meditations and Prayers.* Burns and Oates, London, 1872. Translated from the Latin by M. R.

Konstantine Arkoudas and Selmer Bringsjord. Propositional attitudes and causation. *International Journal of Software and Informatics*, 3(1):47–65, 2009.

Antoine Arnauld and Pierre Nicole. *Logic or the Art of Thinking.* Cambridge texts in the history of philosophy. Cambridge University Press, Cambridge, UK, 1996. Edited by Jill Vance Buroker.

Solomon Asch. On the use of metaphors in the description of persons. In Heinz Werner, editor, *On Expressive Language*, pages 29–38. Clark University Press, Worchester, MA, 1955.

Robert J. Aumann. Agreeing to Disagree. *The Annals of Statistics*, pages 1236–1239, 1976.

M. Avriel and A. C. Williams. The value of information and stochastic programming. *Operations Research*, 18(5):947–954, 1970.

Giosuè Baggio and Peter Hagoort. The balance between memory and unification in semantics: A dynamic account of the n400. *Language and Cognitive Processes*, 26(9):1338–1367, 2011.

Chris L. Baker, Rebecca R. Saxe, and Joshua B. Tenenbaum. Bayesian Theory of Mind: Modeling Joint Belief-Desire Attribution. In Thomas F. Shipley Laura Carlson, Christoph Hoelscher, editor, *Expanding the Space of Cognitive Science: Proceedings of the 33rd Annual Meeting of the Cognitive Science Society*, pages 2469–2474, Austin, TX, 2011. Cognitive Science Society.

Alexandru Baltag, Lawrence S. Moss, and Sławomir Solecki. The Logic of Public Announcements, Common Knowledge, and Private Suspicions. In *Proceedings of the 7th conference on Theoretical aspects of rationality and knowledge*, pages 43–56, 1998.

Simon Baron-Cohen, Alan M. Leslie, and Uta Frith. Does the autistic child have a "theory of mind"? *Cognition*, 21(1):37–46, 1985.

Thomas Bayes. An Essay Towards Solving a Problem in the Doctrine of Chances. *Philosophical Transactions of the Royal Society of London*, 53:370–418, December 1763. Communicated by Mr. Price, in a Letter to John Canton, A. M. F. R. S.

Richard Bellman. On the theory of dynamic programming. *Proceedings of the National Academy of Sciences of the United States of America*, 38(8):716, 1952.

St. Benedict. *The Rule of Our Most Holy Father St. Benedict, Patriarch of Monks.* R. Washbourne, London, 1875. From the old English edition of 1638.

Shlomo Bentin, Gregory McCarthy, and Charles C. Wood. Event-related potentials, lexical decision and semantic priming. *Electroencephalography and clinical Neurophysiology*, 60(4):343–355, 1985.

Jacob Bernoulli. *The Art of Conjecturing, Together with Letter to a Friend on Sets in Court Tennis.* Johns Hopkins University Press, Baltimore, MA, 2006. Translated with an introduction and notes by Edith Dudley Sylla.

Thomas H. Bestul. *Meditatio*/meditation. In Amy Hollywood and Patricia Z. Beckman, editors, *The Cambridge Companion to Christian Mysticism*, chapter 7, pages 157–166. Cambridge University Press, Cambridge, UK, 2012.

Richard M. Billow. Metaphor: A Review of the Psychological Literature. *Psychological Bulletin*, 84(1):81, 1977.

Max Black. Review of *Metaphors We Live By*. *The Journal of Aesthetics and Art Criticism*, 40(2):208–210, 1981.

William Blackstone. *Commenataries on the Laws of England*, volume IV. Clarendon Press, Oxford, UK, 1769.

David Blackwell. *Basic Statistics.* McGraw-Hill, New York, 1969.

David Blackwell and M. A. Girshick. *Theory of Games and Statistical Decisions.* John Wiley and Sons, Inc., New York, 1954.

Leonard Bloomfield. *Language.* Henry Holt and Company, New York, 1933.

Thomas Blundeville. *The Arte of Logicke. Plainly Taught in the English Tongue, according to the best approoued Authours.* William Stansby, London, UK, 1619.

Frank Boers. When a Bodily Source Domain Becomes Prominent: The Joy of Counting Metaphors in the Socio-Economic Domain. In Raymond W. Gibbs, Jr. and Gerard J. Steen, editors, *Metaphor in Cognitive Linguistics*, Amsterdam Studies on the Theory and History of Linguistic Science, Series 4, pages 47–56. John Benjamins, 1999.

Pia Borlund. The concept of relevance in IR. *Journal of the American Society for information Science and Technology*, 54(10):913–925, 2003.

Ina Bornkessel-Schlesewsky and Matthias Schlesewsky. An alternative perspective on “semantic P600” effects in language comprehension. *Brain Research Reviews*, 59(1):55–73, 2008.

Lera Boroditsky and Michael Ramscar. The Roles of Body and Mind in Abstract Thought. *Psychological Science*, 13(2):185–189, 2002.

George E. P. Box and George C. Tiao. *Bayesian Inference in Statistical Analysis.* Addison-Wesley Publishing Company, Reading, MA, 1973.

Mark J. Brandt, Hans IJzerman, and Irene Blanken. Does Recalling Moral Behavior Change the Perception of Brightness? *Social Psychology*, 45(3):246–252, 2014.

Torben Braüner. Hybrid-Logical Reasoning in the Smarties and Sally-Anne Tasks. *Journal of Logic, Language and Information*, 23(4):415–439, 2014.

Michel Bréal. *Semantics: Studies in the Science of Meaning.* William Heinemann, London, 1900.

Roger Brown. *Words and Things.* Free Press, Glencoe, IL, 1958.

Erik Brynjolfsson, Andrew McAfee, and Michael Spence. New World Order. *Foreign Affairs*, 29, 2014.

Eustace Budgell. Contentious Conversation of Gentlemen of the Long Robe—Advice and Disputes. In *The Works of Joseph Addision. Complete in Three Volumes. Embracing the Whole of the "Spectator," &c.*, volume 1, chapter 197, pages 290–291. Harper and Brothers, New York, NY, 1850.

Wray L. Buntine. Operations for Learning with Graphical Models. In *Journal of Artificial Intelligence Research*, volume 2, pages 159–225, 1994.

Rudolf Carnap. Inductive Logic and Rational Decisions. In Rudolf Carnap and Richard C. Jeffrey, editors, *Studies in Inductive Logic and Probability*, volume 1, chapter 1, pages 5–32. University of California Press, Los Angeles, CA, 1971.

Noel Carroll. Visual metaphor. In Jaako Hintikka, editor, *Aspects of Metaphor*, pages 189–218. Martinus Nijhoff, Dordrecht, 1994.

Daniel Casasanto. Conceptual affiliates of metaphorical gestures. Paper presented at the International Conference on Language, Communication, and Cognition, 2008.

Daniel Casasanto and Tania Henetz. Handedness Shapes Children's Abstract Concepts. *Cognitive Science*, 36(2):359–372, 2012.

Daniel Casasanto and Kyle Jasmin. Good and Bad in the Hands of Politicians: Spontaneous Gestures during Positive and Negative Speech. *PLoS One*, 5(7): e11805, 2010.

Daniel Casasanto and Kyle Jasmin. The hands of time: Temporal gestures in english speakers. *Cognitive Linguistics*, 23(4):643–674, 2012.

Ernst Cassirer. *Language and Myth.* Harper, Oxford, 1946.

Herman Chernoff. A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations. *The Annals of Mathematical Statistics*, pages 493–507, 1952.

Noam Chomsky. *Syntactic Structures.* Walter de Gruyter, The Hague, 1957.

George Chrystal. On Some Fundamental Principles in the Theory of Probability. *Transactions of the Actuarial Society of Edinburgh*, 2:420–439, 1891.

Alan Cienki. Why study metaphor and gesture? In Alan Cienki and Cornelia Müller, editors, *Metaphor and Gesture*, pages 5–26. John Benjamins, Amsterdam, 2008.

Zachary Coke. *The Art of Logick or The entire Body of Logick in English.* Robert White, London, UK, 1654. Contrary to what the title page says, this book was printed in 1653 and written by Henry Ainsworth; see Measell (1977).

William S. Cooper. A definition of relevance for information retrieval. *Information storage and retrieval*, 7(1):19–37, 1971.

Thomas M. Cover. An achievable rate region for the broadcast channel. *IEEE Transactions on Information Theory*, 21(4):399–404, 1975.

Thomas M. Cover and J.A. Thomas. *Elements of information theory.* Wiley-Interscience, 1991.

Richard T. Cox. Probability, Frequency and Reasonable Expectation. *American journal of physics*, 14(1):1–13, 1946.

Harald Cramér. *Mathematical Methods of Statistics.* Princeton University Press, Princeton, NJ, 1946.

Verner Dahlerup, Harald Juul-Jensen, and Lis Jacobsen. *Ordbog over det Danske Sprog.* Gyldendal, Copenhagen, 1918–1956.

Fred J. Damerau. A technique for computer detection and correction of spelling errors. *Communications of the ACM*, 7(3):171–176, 1964.

Bruno de Finetti. La Prévision: Ses lois Logiques, ses Sources Subjectives. *Annales de l'institut Henri Poincaré*, 7(1):1–68, 1937.

Bruno de Finetti. *Probability, Induction and Statistics: The Art of Guessing.* John Wiley & Sons, London, UK, 1972.

Bruno de Finetti. Bayesianism: Its Unifying Role for Both the Foundations and Applications of Statistics. *International Statistical Review*, 42(2):117–130, 1974.

Gilles Deleuze. Postscript on the Societies of Control. *October*, pages 3–7, 1992.

René Descartes. *Meditations on First Philosophy.* Routledge, Abingdon, UK, 1993. Edited and with an introduction by Stanley Tweyman.

Joseph L. Doob. *Stochastic Processes.* John Wiley and Sons, Inc., New York, NY, 1953.

June E. Downey. Literary Synesthesia. *The Journal of Philosophy, Psychology and Scientific Methods*, pages 490–498, 1912.

Anthony William Fairbank Edwards. *Likelihood.* Cambridge University Press, Cambridge, UK, 1972.

Bradley Efron. *The Jackknife, the Bootstrap and other Resampling Plans.* SIAM, 1982.

Greg Evans. The Ritual of The Calling of an Engineer. Informational slideshow by the Corporation of the Seven Wardens, Camp One, January 2011. Available at `http://alumni.utoronto.ca/s/731/images/editor_documents/Iron_Ring_Ceremony_Information_Session_2011%20H%20Edamura.pdf`.

Ronald Fagin and Joseph Y. Halpern. Reasoning about knowledge and probability. *Journal of the ACM*, 41(2):340–367, 1994.

Jerome Feldman. *From Molecule to Metaphor: A Neural Theory of Language.* MIT press, Cambridge, MA, 2006.

Christiane Fellbaum, editor. *WordNet: An Electronic Lexical Database.* MIT press, Cambridge, MA, 1998.

Walter Felscher. *Set Theoretical Logic – The Algebra of Models*, volume 1 of *Lectures on Mathematical Logic.* Gordon and Breach Science Publishers, Amsterdam, the Netherlands, 2000.

Dudley Fenner. *The Artes of Logike and Rethorike, plainelie set foorth in the English tounge, easie to be learned and practised.* R. Schilders, Middelburg, 1584.

Eva M. Fernández and Helen Smith Cairns. *Fundamentals of psycholinguistics.* John Wiley & Sons, 2010.

Ronald A. Fisher. Mathematical probability in the natural sciences. *Technometrics*, 1(1):21–29, 1959.

Ronald Aylmer Fisher. On the Mathematical Foundations of Theoretical Statistics. *Philosophical Transactions of the Royal Society of London, Series A*, pages 309–368, April 1922.

Ronald Aylmer Fisher. Theory of statistical estimation. *Mathematical Proceedings of the Cambridge Philosophical Society*, 22(05):700–725, 1925.

Ronald Aylmer Fisher. *The Design of Experiments.* Oliver and Boyd, 4th edition, 1947.

Ronald Aylmer Fisher. Statistical Methods and Scientific Induction. *Journal of the Royal Statistical Society, Series B*, 17(1):69–78, 1955.

Luciano Floridi. Understanding epistemic relevance. *Erkenntnis*, 69(1):69–92, 2008.

Charles Forceville. Metaphor in Pictures and Multimodal Representations. In Raymond W. Gibbs, Jr., editor, *The Cambridge Handbook of Metaphor and Thought*, pages 462–482. Cambridge University Press, Cambridge, UK, 2008.

Michel Foucault. *Psychiatric Power: Lectures at the Collège de France, 1973–1974*. Palgrave MacMillan, Basingstoke, UK, 2006.

W. Nelson Francis and Henry Kucera. *Computational analysis of present-day American English*. Brown University Press, Providence, RI, 1967.

Michael Franke and Gerhard Jäger. Pragmatic Back-and-Forth Reasoning. In Salvatore Pistoia Reda, editor, *Pragmatics, Semantics and the Case of Scalar Implicatures*, pages 170–200. Palgrave Macmillan, New York, NY, 2014.

Michael D. Franzen. *Reliability and Validity in Neuropsychological Assessment*. Springer, 2000.

Sabine Frenzel, Matthias Schlesewsky, and Ina Bornkessel-Schlesewsky. Conflicts in language processing: A new perspective on the n400–p600 distinction. *Neuropsychologia*, 49(3):574–579, 2011.

Angela D. Friederici, Shirley-Ann Rueschemeyer, Anja Hahne, and Christian J. Fiebach. The Role of Left Inferior Frontal and Superior Temporal Cortex in Sentence Comprehension: Localizing Syntactic and Semantic Processes. *Cerebral cortex*, 13(2):170–177, 2003.

Karl Friston. A theory of cortical responses. *Philosophical transactions of the Royal Society B: Biological sciences*, 360(1456):815–836, 2005.

Robert Gallager. Transcript of Discrete Stochastic Processes, Lecture 21: Hypothesis Testing and Random Walks. MIT Open Courseware, Spring 2011. Available at `http://ocw.mit.edu/courses/electrical-engineering-and-computer-science/6-262-discrete-stochastic-processes-spring-2011/video-lectures/lecture-21-hypothesis-testing-and-random-walks/goT94BheP3E.pdf`.

John B. Geijsbeek, editor. *Ancient Double-Entry Bookkeeping: Lucas Pacioli's Treatise reproduced and translated with reproductions, notes and abstracts from Monzoni, Pietra, Mainardi, Ympyn, Stevin and Dafforne*. John B. Geijsbeek, Denver, CO, 1914.

Raymond W. Gibbs, Jr. Why Many Concepts Are Metaphorical. *Cognition*, 61 (3):309–319, 1996.

Raymond W. Gibbs, Jr. *Embodiment and Cognitive Science*. Cambridge University Press, Cambridge, UK, 2005.

Raymond W. Gibbs, Jr. Evaluating conceptual metaphor theory. *Discourse Processes*, 48(8):529–562, 2011.

Edward Gibson, Leon Bergen, and Steven T. Piantadosi. Rational Integration of Noisy Evidence and Prior Semantic Expectations in Sentence Interpretation. *Proceedings of the National Academy of Sciences*, 110(20):8051–8056, 2013.

Gerd Gigerenzer and Henry Brighton. Homo heuristicus: Why biased minds make better inferences. *Topics in Cognitive Science*, 1(1):107–143, 2009.

Jacob Glazer and Ariel Rubinstein. A study in the pragmatics of persuasion: a game theoretical approach. *Theoretical Economics*, 1(4):395–410, 2006.

Irving John Good. Some history of the hierarchical bayesian methodology. *Trabajos de estadística y de investigación operativa*, 31(1):489–519, 1980.

Noah Goodman, Vikash Mansinghka, Daniel Roy, Keith Bonawitz, and Daniel Tarlow. Church: a language for generative models. In *Proceedings of the 24th Conference in Uncertainty in Artificial Intelligence*, pages 220–229, 2008.

Noah D. Goodman, Chris L. Baker, Elizabeth Baraff Bonawitz, Vikash K. Mansinghka, Alison Gopnik, Henry Wellman, Laura Schulz, and Joshua B. Tenenbaum. Intuitive Theories of Mind: A Rational Approach to False Belief. In *Proceedings of the twenty-eighth annual conference of the cognitive science society*, pages 1382–1387, 2006.

John Govan. The Theory of Inverse Probability, with Special Reference to Professor Chrystal's Paper "On Some Fundamental Principles in the Theory of Probability.". *Transactions of the Faculty of Actuaries*, 8:207–230, 1921. Read before the Actuarial Society of Edinburgh in 1893.

Joseph E. Grady. THEORIES ARE BUILDINGS Revisited. *Cognitive Linguistics*, 8(4):267–290, 1997.

Joseph E. Grady. A Typology of Motivation for Conceptual Metaphor: Correlation vs. Resemblance. In Raymond W. Gibbs, Jr. and Gerard J. Steen, editors, *Metaphor in Cognitive Linguistics*, pages 79–100. John Benjamins, Amsterdam, 1999.

Pope Gregory I. *The Dialogues of Saint Gregory the Great.* Philip Lee Warner, London, 1911.

Kilem Li Gwet. *Handbook of Inter-Rater Reliability: The Definitive Guide to Measuring the Extent of Agreement Among Multiple Raters.* Advanced Analytics, Gaithersburg, 3rd edition, 2012.

Pierre Hadot. *Philosophy as a Way of Life: Spiritual Exercises from Socrates to Foucault.* Blackwell, Oxford, UK, 1995. Edited and with an introduction by Arnold A. Davidson.

Anders Hald. *A History of Parametric Statistical Inference from Bernoulli to Fisher, 1713–1935.* Sources and Studies in the History of Mathematics and Physical Sciences. Springer Science+Business Media LLC, 2007.

Sandra Handl and Hans-Jörg Schmid. *Windows to the mind: Metaphor, Metonymy and Conceptual Blending.* Mouton de Gruyter, Berlin, 2011.

G. H. Hardy. *A Mathematician's Apology.* Cambridge University Press, Cambridge, UK, 1967. With a foreword by C. P. Snow.

John C. Harsanyi. Games with Incomplete Information Played by "Bayesian" Players, I–III: Part I. The Basic Model. *Management Science*, 14(3):159–182, 1967.

Verena Haser. *Metaphor, Metonymy, and Experientialist Philosophy: Challenging Cognitive Semantics.* Mouton de Gruyter, Berlin, 2005.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* Springer, New York, 2nd edition, 2009.

David Hilbert. Hilbert an Frege 29. 12. 1899. In *Gottlob Freges Briefwechsel mit D. Hilbert, E. Husserl, B. Russell sowie ausgewählte Einzelbriefe Freges*, volume 321 of *Philosophische Bibliothek*, chapter XV/4. Felix Meiner Verlag, Hamburg, 1980.

William Hinde. *A Faithfull Remonstrance of The Holy Life and Happy Death, of Iohn Bruen.* R.B., London, 1641.

Ebba Hjorth, Kjeld Kristensen, Henrik Lorentzen, and Lars Trap-Jensen, editors. *Den Danske Ordbog.* Gyldendal, Copenhagen, 2003–2005.

J. L. Hodges and E. L. Lehmann. *Basic Concepts of Probability and Statistics.* Holden-Day, Inc., San Francisco, CA, 1964.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American statistical association*, 58(301):13–30, 1963.

John C. J. Hoeks, Laurie A. Stowe, and Gina Doedens. Seeing words in context: the interaction of lexical and sentence level information during reading. *Cognitive Brain Research*, 19(1):59–73, 2004.

John M. Hull. *Touching the Rock: An Experience of Blindness.* SPCK, 1990.

Edwin T. Jaynes. Confidence Intervals vs Bayesian Intervals (1976). In R. D. Rosenkrantz, editor, *Papers on Probability, Statistics, and Statistical Physics*, volume 158 of *Synthese Library*, chapter 9, pages 149–209. Kluwer Academic Publishers, Dordrecht, the Netherlands, 1983.

Edwin T. Jaynes. *Probability Theory: The Logic of Science.* Cambridge University Press, 2003.

Harold Jeffreys. Probability, Statistics, and the Theory of Errors. *Proceedings of the Royal Society of London, Series A*, 139:523–535, 1933.

Harold Jeffreys. *The Theory of Probability.* Oxford University Press, 1939.

Harold Jeffreys. *Scientific Inference.* Cambridge University Press, Cambridge, UK, 3rd edition, 1973.

Mark Johnson. *The Meaning of the Body: Aesthetics of Human Understanding.* University of Chicago Press, Chicago, 2007.

Mark Johnson and George Lakoff. Why cognitive linguistics requires embodied realism. *Cognitive linguistics*, 13(3):245–264, 2002.

Thomas Kauffman, Hugo Théoret, and Alvaro Pascual-Leone. Braille Character Discrimination in Blindfolded Human Subjects. *Neuroreport*, 13(5):571–574, 2002.

John L. Kelly. A new interpretation of information rate. *IRE Transactions on Information Theory*, 2(3):185–189, 1956.

Maurice G. Kendall. *The Advanced Theory of Statistics*, volume II. Charles Griffin and Co., Ltd., London, UK, 3rd edition, 1951.

John Maynard Keynes. *A Treatise on Probability.* MacMillan and Co., London, UK, 1921.

Boaz Keysar and Bridget Bly. Intuitions of the Transparency of Idioms: Can one Keep a Secret by Spilling the Beans? *Journal of Memory and Language*, 34 (1):89–109, 1995.

Boaz Keysar, Yeshayahu Shen, Sam Glucksberg, and William S. Horton. Conventional Language: How Metaphorical Is It? *Journal of Memory and Language*, 43(4):576–593, 2000.

Albert Kim and Lee Osterhout. The independence of combinatory semantic processing: Evidence from event-related potentials. *Journal of Memory and Language*, 52(2):205–225, 2005.

Paul Kline. *Handbook of Psychological Testing.* Routledge, 2nd edition, 2013.

Corinna Klinge, Brigitte Röder, and Christian Büchel. Increased Amygdala Activation to Emotional Auditory Stimuli in the Blind. *Brain*, page awq102, 2010.

Herman H. J. Kolk, Dorothee J. Chwilla, Marieke van Herten, and Patrick J. W. Oor. Structure and limited capacity in verbal working memory: A study with event-related potentials. *Brain and Language*, 85(1):1–36, 2003.

A. Kolmogoroff. *Grundbegriffe der Wahrscheinlichkeitsrechnung.* Julius Springer, Berlin, Germany, 1933.

Zoltán Kövecses. *Metaphor in Culture: Universality and Variation.* Cambridge University Press, Cambridge, UK, 2005.

Stefan Krauss, Laura Martignon, and Ulrich Hoffrage. Simplifying Bayesian Inference: The General Case. In Paul Thagard Lorenzo Magnani, Nancy J. Nersessian, editor, *Model-Based Reasoning in Scientific Discovery*, pages 165–180. Springer Science+Business Media, 1999.

Saul A. Kripke. Semantical Considerations on Modal Logic. *Acta Philosophica Fennica*, 16:83–94, 1963.

Solomon Kullback and Richard A. Leibler. On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86, 1951.

Gina R. Kuperberg. Neural mechanisms of language comprehension: Challenges to syntax. *Brain Research*, 1146:23–49, 2007.

Gina R. Kuperberg, Tatiana Sitnikova, David Caplan, and Phillip J. Holcomb. Electrophysiological distinctions in processing conceptual relationships within simple sentences. *Cognitive Brain Research*, 17(1):117–129, 2003.

Boris Abramovich Kushner. *Lectures on Constructive Mathematical Analysis*, volume 60 of *Translations of Mathematical Monographs.* American Mathematical Society, 1984.

Marta Kutas and Kara D. Federmeier. Thirty years and counting: finding meaning in the N400 component of the event-related brain potential (ERP). *Annual review of psychology*, 62:621–647, 2011.

Marta Kutas and Steven A. Hillyard. Reading senseless sentences: Brain potentials reflect semantic incongruity. *Science*, 207(4427):203–205, 1980.

George Lakoff. *Women, Fire, and Dangerous Things: What Categories Reveal About the Mind.* University of Chicago Press, Chicago, 1987.

George Lakoff. The Neural Theory of Metaphor. In Raymond W. Gibbs, Jr., editor, *The Cambridge Handbook of Metaphor and Thought*, pages 17–38. Cambridge University Press, Cambridge, UK, 2008.

George Lakoff and Mark Johnson. *Metaphors We Live By.* University of Chicago press, Chicago, 1980a.

George Lakoff and Mark Johnson. The Metaphorical Structure of the Human Conceptual System. *Cognitive Science*, 4(2):195–208, 1980b.

George Lakoff and Mark Johnson. *Philosophy in the Flesh: The Embodied Mind and Its Challenge to Western Thought.* Basic books, New York, 1999.

George Lakoff, Jane Espenson, and Alan Schwartz. Master Metaphor List, Second Draft Copy. Manuscript circulated by the Cognitive Linguistics Group of the University of California at Berkeley, 1991.

Barbara Landau and Lila R. Gleitman. *Language and Experience: Evidence from the Blind Child.* Harvard University Press, Cambridge, MA, 1985.

Susanne K. Langer. *Philosophy in a New Key: A Study in the Symbolism of Reason, Rite, and Art.* The New American Library, 1948.

Pierre-Simon Laplace. *Philosophical Essay on Probabilities*, volume 13 of *Sources in the History of Mathematics and Physical Sciences.* Springer Science+Business Media LLC, New York, NY, 1995. Translated from the fifth French edition of 1825 by Andrew I. Dale.

Bruno Latour. *Reassembling the Social: An Introduction to Actor-Network-Theory.* Clarendon Lectures in Management Studies. Oxford University Press, Oxford, UK, 2005.

Ellen F. Lau, Colin Phillips, and David Poeppel. A cortical network for semantics: (de) constructing the n400. *Nature Reviews Neuroscience*, 9(12):920–933, 2008.

William H. Leatherdale. *The Role of Analogy, Model, and Metaphor in Science.* North-Holland Publishing Co., 1974.

Vladimir I. Levenshtein. Binary Codes Capable of Correcting Deletions, Insertions and Reversals. In *Soviet Physics Doklady*, volume 10, page 707, 1966.

Samuel R. Levin. *The semantics of metaphor.* Johns Hopkins University Press, Baltimore, 1977.

William A. Link and Richard J. Barker. *Bayesian Inference: with ecological applications.* Academic Press, Boston, MA, 2009.

R. Duncan Luce and Howard Raiffa. *Games and Decisions: Introduction and Critical Survey.* John Wiley and Sons, Inc., New York, 1957.

David MacKay. *Information theory, inference and learning algorithms.* Cambridge university press, 2003.

Hannah Macpherson. Articulating Blind Touch: Thinking through the Feet. *The Senses and Society*, 4(2):179–193, 2009.

Mathias Winther Madsen. An Introduction to Multi-Agent Statistics. In Jakub Szymanik and Rineke Verbrugge, editors, *Proceedings of the Second Workshop Reasoning About Other Minds: Logical and Cognitive Perspectives*, volume 1208, pages 16–20. CEUR Workshop Proceedings, `http://ceur-ws.org/Vol-1208/`, 2014a.

Mathias Winther Madsen. A Quantitative Measure of Relevance Based on Kelly Gambling Theory. In *Pristine Perspectives on Logic, Language, and Computation*, pages 124–141. Springer, Dordrecht, the Netherlands, 2014b.

Mathias Winther Madsen. Shannon Versus Chomsky: Brain Potentials and the Syntax-Semantics Distinction. In Henk Zeevat and Hans-Christian Schmitz, editors, *Bayesian Natural Language Semantics and Pragmatics*, number 2 in Language, Cognition, and Mind, pages 117–144. Springer, Dordrecht, the Netherlands, 2015a.

Mathias Winther Madsen. On the Consistency of Approximate Multi-agent Probability Theory. *Künstliche Intelligenz*, pages 1–8, 2015b.

Burkhard Maess, Christoph S Herrmann, Anja Hahne, Akinori Nakamura, and Angela D Friederici. Localizing the distributed language network responsible for the n400 measured by meg during auditory sentence processing. *Brain research*, 1096(1):163–172, 2006.

Sanjoy Mahajan. *Street-Fighting Mathematics.* MIT Press, Cambridge, MA, 2010. Foreword by Carver A. Mead.

Hosam M. Mahmoud. *Pólya Urn Models.* CRC press, 2008.

Lawrence E. Marks. *The Unity of the Senses: Interrelations Among the Modalities.* Academic Press, New York, 1978.

Matthew S. McGlone. Concepts as metaphors. In Sam Glucksberg, editor, *Understanding Figurative Language: From Metaphors to Idioms*, pages 90–107. Oxford University Press, Oxford, 2001.

Matthew S. McGlone. Hyperbole, Homunculi, and Hindsight Bias: An Alternative Evaluation of Conceptual Metaphor Theory. *Discourse Processes*, 48(8): 563–574, 2011.

David McNeill. *Hand and Mind: What Gestures Reveal About Thought.* University of Chicago Press, 1992.

James S. Measell. The Authorship of The Art of Logick (1654). *Journal of the History of Philosophy*, 15(3):321–324, 1977.

Brian P. Meier and Michael D. Robinson. Why the Sunny Side Is Up: Associations Between Affect and Vertical Position. *Psychological science*, 15(4):243–247, 2004.

Brian P. Meier, Michael D. Robinson, and Gerald L. Clore. Why Good Guys Wear White: Automatic Inferences About Stimulus Valence Based on Brightness. *Psychological Science*, 15(2):82–87, 2004.

Lotfi B. Merabet, Roy Hamilton, Gottfried Schlaug, Jascha D. Swisher, Elaine T. Kiriakopoulos, Naomi B. Pitskel, Thomas Kauffman, and Alvaro Pascual-Leone. Rapid and Reversible Recruitment of Early Visual Cortex for Touch. *PLoS One*, 3(8):e3046, 2008.

Pamela S. Morgan. Competition, Cooperation, and Interconnection: 'Metaphor Families' and Social Systems. In Gitte Kristiansen and René Dirven, editors, *Cognitive Sociolinguistics: Language Variation, Cultural Models, Social Systems*, pages 482–516. Mouton de Gruyter, Berlin, 2008.

Philip Morrison. Fermi Questions. *American Journal of Physics*, 31(8):626–627, 1963.

Randall Munroe. *What If? Serious Scientific Answers to Absurd Hypothetical Questions*. Houghton Mifflin Harcourt, Boston, MA, 2014.

Gregory L. Murphy. On Metaphoric Representation. *Cognition*, 60(2):173–204, 1996.

John Newton. *An Introduction to the Art of Logick*. E. T. and R. H., London, UK, 1671.

J. Neyman. *First Course in Probability and Statistics*. Henry Holt and Company, New York, NY, 1950.

J. Neyman and E. S. Pearson. On the Use and Interpretation of Certain Test Criteria for Purposes of Statistical Inference: Part I. *Biometrika*, 20A(1/2): 175–240, 1928.

J. Neyman and E. S. Pearson. On the Problem of the Most Efficient Tests of Statistical Hypotheses. *Philosophical Transactions of the Royal Society of London, Series A*, 231:289–337, 1933.

Sander Nieuwenhuis, Gary Aston-Jones, and Jonathan D. Cohen. Decision Making, the P3, and the Locus Coeruleus–Norepinephrine System. *Psychological bulletin*, 131(4):510, 2005.

Mante S. Nieuwland and Jos J. A. Van Berkum. Testing the limits of the semantic illusion phenomenon: ERPs reveal temporary semantic change deafness in discourse comprehension. *Cognitive Brain Research*, 24(3):691–701, 2005.

Rafael E. Núñez and Eve Sweetser. With the Future Behind Them: Convergent Evidence from Aymara Language and Gesture in the Crosslinguistic Comparison of Spatial Construals of Time. *Cognitive science*, 30(3):401–450, 2006.

Andrew Ortony. Beyond Literal Similarity. *Psychological Review*, 86(3):161, 1979.

Charles Egerton Osgood, George J. Suci, and Percy H. Tannenbaum. *The Measurement of Meaning.* University of Illinois press, Urbana, 1957.

Lee Osterhout and Phillip J. Holcomb. Event-related potentials and syntactic anomaly: Evidence of anomaly detection during the perception of continuous speech. *Language and Cognitive Processes*, 8(4):413–437, 1993.

Alvaro Pascual-Leone, Souzana Obretenova, and Lotfi B. Merabet. Paradoxical Effects of Sensory Loss. In Alvaro Pascual-Leone, Vilayanur Ramachandran, Jonathan Cole, Sergio Della Sala, Tom Manly, Andrew Mayes, Oliver Sacks, and Narinder Kapur, editors, *The Paradoxical Brain*, pages 14–39. Cambridge University Press, Cambridge, UK, 2011.

Judea Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausble Inference.* Morgan Kaufmann, San Francisco, CA, 1988.

Pragglejaz Group. MIP: A Method for Identifying Metaphorically Used Words in Discourse. *Metaphor and symbol*, 22(1):1–39, 2007.

Peter Ramus. *The Art of Logick.* Antony Wotton, London, UK, 1626.

I. A. Richards. *Interpretations in Teaching.* Harcourt Brace, New York, 1938.

W. M. Kim Roddis. Structural Failures and Engineering Ethics. *Journal of Structural Engineering*, 119(5):1539–1555, 1993.

Tim Rohrer. Understanding through the Body: fMRI and ERP Investigations into the Neurophysiology of Cognitive Semantics. Paper presented at the Seventh International Cognitive Linguistics Conference, 2001.

Frank Rösler, Peter Pütz, Angela Friederici, and Anja Hahne. Event-related brain potentials while encountering semantic and syntactic constraint violations. *Journal of Cognitive Neuroscience*, 5(3):345–362, 1993.

Edward Walter Samson. Fundamental natural concepts of information theory. Technical Report E 5079, Air Force Cambridge Research Center, 1951.

Jona Sassenhagen, Matthias Schlesewsky, and Ina Bornkessel-Schlesewsky. The P600-as-P3 hypothesis revisited: Single-trial analyses reveal that the late EEG positivity following linguistically deviant material is reaction time aligned. *Brain and language*, 137:29–39, 2014.

Donald A. Schon. *Displacement of Concepts.* Tavistock Publications, London, 1963.

Claude E. Shannon. A mathematical theory of communication. *Bell system technical journal*, 27:379–423 and 623–656, 1948.

Tatiana Sitnikova, Phillip J Holcomb, Kristi A Kiyonaga, and Gina R Kuperberg. Two neurocognitive mechanisms of semantic integration during the comprehension of visual real-world events. *Journal of Cognitive Neuroscience*, 20 (11):2037–2057, 2008.

Michael K. Smith. Metaphor and mind. *American Speech*, 57(2):128–134, 1982.

Michael Spence. Job market signaling. *The quarterly journal of Economics*, 87 (3):355–374, 1973.

Dan Sperber and Deirdre Wilson. *Relevance: Communication and Cognition.* Blackwell, Oxford, UK, 1986.

Dan Sperber and Deirdre Wilson. Précis of *Relevance: Communication and Cognition. Behavioral and Brain Sciences*, 10(04):697–710, 1987.

Peter Spirtes. Directed cyclic graphical representations of feedback models. In P. Besnard and S. Hanks, editors, *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, pages 491–498, San Mateo, 1995. Morgan Kaufmann Publishers.

Laura Stearns. Watermelon, honeydew, and antelope: An ERP study of semantically anomalous but phonologically rxpected words in sentences. Master's thesis, Wellesley College, May 2012.

Charles Stein. Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 197–206, 1956.

Keith Stenning and Michiel Van Lambalgen. *Human Reasoning and Cognitive Science.* MIT Press, Cambridge, MA, 2008.

Gustaf Stern. *Meaning and Change of Meaning: With Special Reference to the English Language.* Indiana Unversity Press, Bloomington, 1931.

Stephen M. Stigler. *The History of Statistics: The Measurement of Uncertainty Before 1900.* Harvard University Press, Cambridge, MA, 1986.

Barbara M. H. Strang. Review of *Metaphors We Live By*. *The Modern Language Review*, 77(1):134–136, 1982.

Student. The Probable Error of a Mean. *Biometrika*, pages 1–25, 1908.

Andreas Stuhlmüller and Noah D. Goodman. Reasoning about reasoning by nested conditioning: Modeling theory of mind with probabilistic programs. *Cognitive Systems Research*, 28:80–99, 2013.

Samuel Sutton, Margery Braren, Joseph Zubin, and ER John. Evoked-potential correlates of stimulus uncertainty. *Science*, 150(3700):1187–1188, 1965.

Eve Sweetser. *From Etymology to Pragmatics: Metaphorical and Cultural Aspects of Semantic Structure.* Cambridge University Press, Cambridge, 1990.

Paul Thibodeau and Frank H. Durgin. Productive Figurative Communication: Conventional Metaphors Facilitate the Comprehension of Related Novel Metaphors. *Journal of Memory and Language*, 58(2):521–540, 2008.

Owen Paul Thomas. *Metaphor, and Related Subjects.* Random House, 1969.

Stephen Ullmann. *The Principles of Semantics.* Jackson, Son & Company, Glasgow, 1951.

Johan van Benthem, Jelle Gerbrandy, and Barteld Kooi. Dynamic Update with Probabilities. *Studia Logica*, 93(1):67–96, 2009.

Stefan Van Dongen. Prior specification in Bayesian statistics: Three cautionary tales. *Journal of Theoretical Biology*, 242(1):90–100, 2006.

Marieke van Herten, Herman H. J. Kolk, and Dorothee J. Chwilla. An ERP study of P600 effects elicited by semantic anomalies. *Cognitive Brain Research*, 22 (2):241–255, 2005.

Marieke Van Herten, Dorothee J. Chwilla, and Herman H. J. Kolk. When heuristics clash with parsing routines: ERP evidence for conflict monitoring in sentence perception. *Journal of Cognitive Neuroscience*, 18(7):1181–1197, 2006.

Anna Maria van Schurman. *The Learned Maid; Or, Whether a Maid May be a Scholar?* John Redmayne, London, UK, 1659.

John Venn. *The Logic of Chance.* Macmillan, London, UK, 3rd edition, 1888.

Robert R. Verbrugge and Nancy S. McCarrell. Metaphoric Comprehension: Studies in Reminding and Resembling. *Cognitive Psychology*, 9(4):494–533, 1977.

Geerat J. Vermeij. *Privileged Hands: A Scientific Life*. Macmillan, 1997.

Andrew Viterbi. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Transactions on Information Theory*, 13 (2):260–269, 1967.

Richard von Mises. *Probability, Statistics and Truth*. The MacMillan Company, New York, NY, 2nd edition, 1957.

Abraham Wald. *Sequential analysis*. John Wiley and Sons, Inc., New York, 1947.

Abraham Wald. *Statistical Decision Functions*. Wiley, New York, NY, 1950.

Chamont Wang. *Sense and Nonsense of Statistical Inference: Controversy, Misuse, and Subtlety*. Marcel Dekker, Inc., New York, NY, 1993.

Larry Wasserman. *All of nonparametric statistics*. Springer, 2006.

Isaac Watts. *Logick: Or, the Right Use of Reason in the Enquiry after Truth*. John Clark, London, UK, second edition, 1726.

Max Weber. *The Protestant Ethic and the Spirit of Capitalism*. Routledge, London, UK, 1992. Translated by Talcott Parsons. With an introduction by Anthony Giddens.

Jill Weckerly and Marta Kutas. An electrophysiological analysis of animacy effects in the processing of object relative sentences. *Psychophysiology*, 36(5):559–570, 1999.

Kip A. Wedel. *The Obligation: A History of the Order of the Engineer*. AuthorHouse, Bloomington, IN, 2012.

Lawrence Weinstein and John A. Adam. *Guesstimation: Solving the World's Problems on the Back of a Cocktail Napkin*. Princeton University Press, Princeton, NJ, 2009.

Johannes Weiss. On the Irreversibility of Western Rationalization and Max Weber's Alleged Fatalism. In Scott Lash and Sam Whimster, editors, *Max Weber, Rationality and Modernity*, chapter 7, pages 154–163. Allen & Unwin, London, UK, 1987.

Henry M Wellman, David Cross, and Julanne Watson. Meta-Analysis of Theory-of-Mind Development: The Truth about False Belief. *Child development*, 72 (3):655–684, 2001.

Heinz Werner. *Die Ursprünge der Metapher*. W. Engelmann, Leipzig, 1919.

Heinz Werner. Change of meaning: A study of semantic processes through the experimental method. *The Journal of General Psychology*, 50(2):181–208, 1954.

Heinz Werner and Bernard Kaplan. *Symbol Formation*. Wiley, 1963.

Benjamin Lee Whorf. *Language, Thought and Reality: Selected Writings of Benjamin Lee Whorf*. Massachusetts Institute of Technology, 1956. ed. by Carroll, John B.

Joseph M Williams. Synaesthetic Adjectives: A Possible Law of Semantic Change. *Language*, pages 461–478, 1976.

Nicole L. Wilson and Raymond W. Gibbs, Jr. Real and Imagined Body Movement Primes Metaphor Comprehension. *Cognitive Science*, 31(4):721–731, 2007.

Jean-Daniel Zucker and Christophe Meyer. Apprentissage pour l'anticipation de comportements de joueurs humains dans les jeux à information complète et imparfaite: les «mind-reading machines». *Revue d'intelligence artificielle*, 14 (3-4):313–338, 2000.

Titles in the ILLC Dissertation Series:

ILLC DS-2009-01: **Jakub Szymanik**
Quantifiers in TIME and SPACE. Computational Complexity of Generalized Quantifiers in Natural Language

ILLC DS-2009-02: **Hartmut Fitz**
Neural Syntax

ILLC DS-2009-03: **Brian Thomas Semmes**
A Game for the Borel Functions

ILLC DS-2009-04: **Sara L. Uckelman**
Modalities in Medieval Logic

ILLC DS-2009-05: **Andreas Witzel**
Knowledge and Games: Theory and Implementation

ILLC DS-2009-06: **Chantal Bax**
Subjectivity after Wittgenstein. Wittgenstein's embodied and embedded subject and the debate about the death of man.

ILLC DS-2009-07: **Kata Balogh**
Theme with Variations. A Context-based Analysis of Focus

ILLC DS-2009-08: **Tomohiro Hoshi**
Epistemic Dynamics and Protocol Information

ILLC DS-2009-09: **Olivia Ladinig**
Temporal expectations and their violations

ILLC DS-2009-10: **Tikitu de Jager**
"Now that you mention it, I wonder...": Awareness, Attention, Assumption

ILLC DS-2009-11: **Michael Franke**
Signal to Act: Game Theory in Pragmatics

ILLC DS-2009-12: **Joel Uckelman**
More Than the Sum of Its Parts: Compact Preference Representation Over Combinatorial Domains

ILLC DS-2009-13: **Stefan Bold**
Cardinals as Ultrapowers. A Canonical Measure Analysis under the Axiom of Determinacy.

ILLC DS-2010-01: **Reut Tsarfaty**
Relational-Realizational Parsing

ILLC DS-2010-02: **Jonathan Zvesper**
Playing with Information

ILLC DS-2010-03: **Cédric Dégremont**
The Temporal Mind. Observations on the logic of belief change in interactive systems

ILLC DS-2010-04: **Daisuke Ikegami**
Games in Set Theory and Logic

ILLC DS-2010-05: **Jarmo Kontinen**
Coherence and Complexity in Fragments of Dependence Logic

ILLC DS-2010-06: **Yanjing Wang**
Epistemic Modelling and Protocol Dynamics

ILLC DS-2010-07: **Marc Staudacher**
Use theories of meaning between conventions and social norms

ILLC DS-2010-08: **Amélie Gheerbrant**
Fixed-Point Logics on Trees

ILLC DS-2010-09: **Gaëlle Fontaine**
Modal Fixpoint Logic: Some Model Theoretic Questions

ILLC DS-2010-10: **Jacob Vosmaer**
Logic, Algebra and Topology. Investigations into canonical extensions, duality theory and point-free topology.

ILLC DS-2010-11: **Nina Gierasimczuk**
Knowing One's Limits. Logical Analysis of Inductive Inference

ILLC DS-2010-12: **Martin Mose Bentzen**
Stit, Iit, and Deontic Logic for Action Types

ILLC DS-2011-01: **Wouter M. Koolen**
Combining Strategies Efficiently: High-Quality Decisions from Conflicting Advice

ILLC DS-2011-02: **Fernando Raymundo Velazquez-Quesada**
Small steps in dynamics of information

ILLC DS-2011-03: **Marijn Koolen**
The Meaning of Structure: the Value of Link Evidence for Information Retrieval

ILLC DS-2011-04: **Junte Zhang**
System Evaluation of Archival Description and Access

ILLC DS-2011-05: **Lauri Keskinen**
Characterizing All Models in Infinite Cardinalities

ILLC DS-2011-06: **Rianne Kaptein**
Effective Focused Retrieval by Exploiting Query Context and Document Structure

ILLC DS-2011-07: **Jop Briët**
Grothendieck Inequalities, Nonlocal Games and Optimization

ILLC DS-2011-08: **Stefan Minica**
Dynamic Logic of Questions

ILLC DS-2011-09: **Raul Andres Leal**
Modalities Through the Looking Glass: A study on coalgebraic modal logic and their applications

ILLC DS-2011-10: **Lena Kurzen**
Complexity in Interaction

ILLC DS-2011-11: **Gideon Borensztajn**
The neural basis of structure in language

ILLC DS-2012-01: **Federico Sangati**
Decomposing and Regenerating Syntactic Trees

ILLC DS-2012-02: **Markos Mylonakis**
Learning the Latent Structure of Translation

ILLC DS-2012-03: **Edgar José Andrade Lotero**
Models of Language: Towards a practice-based account of information in natural language

ILLC DS-2012-04: **Yurii Khomskii**
Regularity Properties and Definability in the Real Number Continuum: idealized forcing, polarized partitions, Hausdorff gaps and mad families in the projective hierarchy.

ILLC DS-2012-05: **David García Soriano**
Query-Efficient Computation in Property Testing and Learning Theory

ILLC DS-2012-06: **Dimitris Gakis**
Contextual Metaphilosophy - The Case of Wittgenstein

ILLC DS-2012-07: **Pietro Galliani**
The Dynamics of Imperfect Information

ILLC DS-2012-08: **Umberto Grandi**
Binary Aggregation with Integrity Constraints

ILLC DS-2012-09: **Wesley Halcrow Holliday**
Knowing What Follows: Epistemic Closure and Epistemic Logic

ILLC DS-2012-10: **Jeremy Meyers**
Locations, Bodies, and Sets: A model theoretic investigation into nominalistic mereologies

ILLC DS-2012-11: **Floor Sietsma**
Logics of Communication and Knowledge

ILLC DS-2012-12: **Joris Dormans**
Engineering emergence: applied theory for game design

ILLC DS-2013-01: **Simon Pauw**
Size Matters: Grounding Quantifiers in Spatial Perception

ILLC DS-2013-02: **Virginie Fiutek**
Playing with Knowledge and Belief

ILLC DS-2013-03: **Giannicola Scarpa**
Quantum entanglement in non-local games, graph parameters and zero-error information theory

ILLC DS-2014-01: **Machiel Keestra**
Sculpting the Space of Actions. Explaining Human Action by Integrating Intentions and Mechanisms

ILLC DS-2014-02: **Thomas Icard**
The Algorithmic Mind: A Study of Inference in Action

ILLC DS-2014-03: **Harald A. Bastiaanse**
Very, Many, Small, Penguins

ILLC DS-2014-04: **Ben Rodenhäuser**
A Matter of Trust: Dynamic Attitudes in Epistemic Logic

ILLC DS-2015-01: **María Inés Crespo**
Affecting Meaning. Subjectivity and evaluativity in gradable adjectives.