THE UNIVERSITY
of EDINBURGH

Cognition at the Symbolic Threshold: The role of abductive inference in hypothesising the meaning of novel signals

Justin William Bernard Sulik

PhD
Linguistics and English Language
University of Edinburgh
2014

Abstract

Humans readily infer the meanings of novel symbols in communicative contexts of varying complexity, and several researchers in the field of language evolution have explicitly acknowledged that inference plays a key role in accounting for the evolution of symbolic communication. However, in this field at least, there has been very little investigation into the nature of inference in this regard. That is, evolutionary linguists have yet to address the following questions if we are to have a fuller picture of how humans came to communicate symbolically:

1. What kinds of inference are there? Specifically,
    - i Diachronically, what forms of inference are comparatively simpler in evolutionary terms, and thus shared with a wider range of species? What forms of inference are more complex, and limited to humans or to us and our closest relatives?
    - ii Synchronically, if humans are capable of several kinds of complex inference, how do we know which particular kind of inference is being applied in solving a given problem?
2. How do symbol-learning problems vary? Specifically,
    - i What makes a particular symbol-learning problem more or less complex in terms of the kind of inference needed to solve it?
    - ii How would the communicative context of our pre-linguistic ancestors have been different from that of a human child learning words from its linguistic parent?

This dissertation takes a step towards answering these questions by investigating a little-known form of inference called 'abduction' (or insightful hypothesis generation), which has thus far been wholly overshadowed in language evolution by a much better understood form called 'induction' (or probabilistic hypothesis evaluation). I will argue that abduction and induction are both comparatively complex in the diachronic terms expressed above in 1.i, and while induction is useful in accounting for how modern children learn words from linguistic adults, abduction is more important in situations like those that would have faced our pre-lingistic ancestors as they first began to use symbols. That is, I will argue on both theoretical and empirical grounds that abductive inference was an evolutionary milestone as our ancestors crossed what Deacon (1997) calls the symbolic threshold.

Declaration

I hereby declare that this thesis is of my own composition, and that it contains no material previously submitted for the award of any other degree or professional qualification. The work reported in this thesis has been executed by myself, except where due acknowledgement is made in the text.

Justin Sulik

Acknowledgements

First and foremost I have to thank my flatmate and landlady, Lindsay Mucka, for her support and encouragement over the last few years. Without her kindness and incredible generosity, this thesis would never have been written. Apart from the necessities of life, she has also made sure that this poor student has had his proper share of champagne and fancy dinners.

My supervisor, Kenny Smith, joined this project after 1st year. At that point, I had a number of partly formed theoretical claims, but no idea how to test any of them empirically. It is thanks entirely to him that I have become a keen experimentalist, full of the joys of R. His insightful comments and suggestions have helped limit the amount of Peircean waffling or Fodorian gloom that permeated earlier drafts of this work. I'm writing these acknowledgements on the laptop he kindly lent me when my laptop gave up the ghost just four days before my deadline.

I'm very grateful to my talented sister, Lena, for producing so many figures for me (figs. 1.1, 1.6, 2.1, 2.3, 2.4, 2.5, 2.7) and responding patiently when I demanded change after change. She and my parents have always been full of encouraging words, which has meant a lot to me, and I thank them for always being my guinea pigs when I needed to check a new experiment for bugs.

I appreciate all the time put in by the volunteers who helped me pilot experiments or produce stimuli, particularly Dave, Matthew, Isobel, Nic, Claire, Simon, and Jim. I'm thankful to my supervisors in 1st year, Andrew Smith, Jim Hurford and Matthew Chrisman, for their guidance in getting this work off the ground. I value immensely working in a group such as the LEC, particularly during our January retreats.

Finally, I thank the Skye Foundation for funding the first year of this degree.

Contents

List of Figures

List of Tables

Introduction

Brief outline

Symbolic communication is a central feature of human language (Deacon, 1997; Jackendoff, 1999), purportedly distinguishing our words from the communicative signals produced by animals. In the terms of Deacon (1997), people crossed the 'symbolic threshold' when our ancestors became able to communicate symbolically: we infer the meaning of novel symbols with comparative ease, while even our closest relatives must be laboriously trained to understand novel signs. Attempting to explain how we evolved to do this, though, raises a host of logically prior questions. What is a symbol? How should we characterise the cognitive faculties that allow us to learn them? Are all symbol-learning tasks equally complex? To what extent, if any, do we share symbolic signs and the relevant cognitive faculties with other species? What is the process by which symbols are created? Currently, none of these prior questions have uncontroversial answers, so any account of how we evolved to be a symbolic species is thus based on shaky foundations or is likely to be incomplete. The aim of this thesis is to provide theoretical and empirical support for an answer to these questions.

I outline and demonstrate the relevance to these questions of a little-studied and poorly understood form of inference, 'abduction' (Peirce, 1935). Abduction is hypothesis generation and it plays an important role in explaining how humans learn in novel or unconstrained contexts (I unpack these terms in the body of the thesis). I argue that symbols are just signs that require abductive inference, and that communication at the symbolic threshold would have involved novel or unconstrained contexts, in which case a full explanation of how our species evolved to communicate must include abduction.

In this introduction, I'll briefly describe a number of accounts addressing the larger issue of learning novel symbols, and show how each of them runs into difficulty, though they each contribute something useful. I'll then indicate how my answer to the questions above will provide a framework constraining these accounts, filling the gaps, and yoking them together. Thereafter, I'll show how this proposal is empirically testable.

Bayesian induction explains behaviour as the optimal solution to a probabilistic problem, and has had marked success in explaining word learning in laboratory conditions (Tenenbaum et al., 2006; Xu and Tenenbaum, 2007). These laboratory conditions, however, are highly contextually constrained: the set of possible meanings is made small and manifest by experimental design. Induction can explain why we settle on a particular hypothesis about meaning when we are given a set of possible hypotheses, but cannot explain where this set of hypotheses comes from in the first place: that is the role of abduction. While word learning in modern human children is scaffolded by linguistic parents (a child's attention might be directed to a particular object, for instance, constraining the space of possible hypotheses) and is thus amenable to an inductive account, I will argue that the communicative context of our pre-linguistic ancestors was less constrained, and that abduction was thus a necessary complement to induction.

Relevance theory is a pragmatic account of how we infer what speakers intend to communicate to us (Sperber and Wilson, 1995). It proposes a deductive mechanism that retrieves assumptions one-by-one from our world knowledge until a relevant interpretation is found. However, it fails to distinguish the task of interpreting novel gesture from the task of understanding a conventional utterance. I will argue that novelty does in fact make a difference to how we understand speaker intention, that deduction thus cannot be the whole answer, and that abduction therefore plays a role in understanding communicative intentions.

These accounts of word learning and pragmatic inference appeal to induction and deduction, types of inference, and I suggested that a third form of inference, abduction, fills the gap left by context and novelty. Explaining how it does this requires an account of insight and analogy. Insight is a distinct cognitive mechanism that involves the formation of novel connections within our representation of a problem, and analogy involves the formation of novel connections between representation of a novel problem and repre-

sentation of a familiar problem. Both, I will argue, are well suited to novel and unconstrained contexts. However, there seems to be no point of contact between analogy and insight on one hand and deduction or induction on the other. I suggest that abduction is what links these, in that analogy and insight are creative mechanisms underlying abductive hypothesis generation, while abduction provides the input to deductive and inductive mechanisms for hypothesis evaluation, telling us which previously generated hypothesis is the most rational solution to the problem.

The process of symbolisation describes how non-symbolic signs become increasingly abstract with use (Garrod et al., 2007; Fay et al., 2010). These accounts, though, focus on the sign itself and not on how we understand the sign: they make brief reference to induction but leave the matter there. I argue that abduction is needed in the initial stages of any such process, though it becomes less important as signs are grounded in use, since a shared history of use provides a constraining context for communication.

Though I provide detailed arguments for all the above points, based on theoretical considerations and on a review of previous empirical evidence, I then turn to provide my own empirical evidence for my claims. I unpack a number of features of the communicative context at the symbolic threshold, provide evidence for some measurable diagnostics, then use these diagnostics to determine the extent to which abduction is implicated in word-learning problems as those features are manipulated.

Chapter outlines

Part I: Theory

Chapter 1: Symbols In order to know how symbols came to be or what makes human symbolic communication different from animal communication, we need to have a clear idea of what symbols are. This chapter seeks to provide a definition of 'symbol' that is suitable for enquiring into the question of symbol origins. It investigates common definitions (that symbols are arbitrary signs, or conventional signs) and rejects these as being uninformative in an evolutionary context. It proposes, instead, that symbolic communication is special because it requires complex forms of inference, where complexity depends on

the communicative context.

Chapter 2: An Inferential Hierarchy The question of inferential complexity is addressed by positing an evolutionary hierarchy, allowing us to compare our inferential abilities with animals. The hierarchy ranges from the simplest forms of inference, shared with many species, to the more complex, at which even our nearest relatives perform poorly. The stage at which symbolic communication becomes possible is identified, and the symbol-learning abilities of chimpanzees will be discussed.

Chapter 3: Abduction, Induction and Insight Abductive inference is defined as novel hypothesis generation, and several key features of this form of inference are identified, including the fact that abduction is more insightful than induction and deduction. Abduction is shown to be necessary, on theoretical grounds, for crossing the symbolic threshold, in that it copes with novelty in unconstrained contexts. Since induction, rather than abduction, is the focus of many current approaches to the evolution of symbols, this chapter makes explicit comparisons between the two, showing how both are essential for learning symbols, and identifying what kinds of symbol-learning problems make comparatively larger demands on each form of inference.

Part II: Empirical Data

Chapter 4: Diagnostics of Abductive Inference I examine behavioural diagnostics of insight problem solving, and show experimentally that these are more typical of abductive than of inductive or deductive inference, and can thus be used to distinguish abductive from deductive and inductive processes.

Chapter 5: Word Learning and Hypothesis Spaces I argue that one crucial difference between child language learning and the symbolic threshold is the degree to which the hypothesis space is constrained by linguistic adults. I then present an experiment showing that abduction generates hypotheses when none are given, but that when participants are presented with a set of hypotheses, abduction plays less of a role. I also show that as context becomes increasingly uncon-

strained, abduction plays a larger role. Inductive accounts are thus incomplete explanations of symbol origins.

Chapter 6: Word Learning and Predictability from Context I argue that the communicative context of a novel symbol is more informative for modern humans than it would have been at the symbolic threshold. I then show experimentally that less informative contexts require comparatively more abduction.

Chapter 7: Iconicity and Precedence in Word Learning I examine accounts suggesting that symbols evolved from iconic signs through a process of symbolisation. I present an experiment showing that earlier stages in this process require comparatively more abduction than later stages, so abduction would have played an important role at the symbolic threshold. I also present a statistical analysis of the role of precedent and iconicity in this process, and show that it is the lack of precedence, rather than the presence of iconicity, that predicts how abductive a particular problem will be.

Part III: Conclusions

Chapter 8: Conclusions I review the arguments made throughout, showing that there are a number of reasons it is sensible to define 'symbol' relative to abductive inference, distinguishing abduction from other forms of cognition in humans and non-human animals, showing that hypothesis generation is not explicable by induction, and that abduction is implicated in novel word-learning tasks that are contextually unconstrained, as would have been the case at the symbolic threshold.

Aims and intentions

One general requirement in what follows is that definitional or theoretical distinctions should, at least potentially, be supportable by an empirical differences, and that the relationship between the two should be spelled out clearly. Most of the first three chapters will be unpacking just what the definition of 'symbol' consists in and how we might decide objectively whether

something is a symbol, while the following chapters provide empirical evidence for the claims set out in the theoretical chapters. The theoretical chapters are not intended to be an exhaustive account of the relevant theories (either philosophical or semiotic): the discussion may borrow ideas from these areas, but the focus will be on examining what aspects of the theoretical background could possibly be expected to make an empirical difference to an explanation of the symbolic threshold.

In much current work, the focus has been on the nature of the sign itself, or on the nature of social interaction while using signs. The theoretical background here shows what an internal or cognitive counterpart to these external considerations might look like. The experimental side thus represents a new approach to understanding the role of inference in different kinds of communicative problems. While abduction is a cognitive faculty, I will identify behavioural measures that can identify the extent to which abduction or induction are involved in various sign-learning tasks. The conclusion is that abduction is more likely than induction to be involved in contextually unconstrained communicative tasks.

Part I

Theory

Chapter 1

Symbols

1.1 Introduction

Only humans communicate symbolically. At least, that is the claim common in talk of language evolution (Pinker, 1994; Deacon, 1997; Noble and Davidson, 1996; Penn et al., 2008; Ramscar et al., 2010), though dissenting voices include Seyfarth et al. (1980), Armstrong (1993), Hurford (2004), El-Hani et al. (2010) and Ribeiro et al. (2006). Taken at face value, the debate here seems to be about what we do when we talk, about what animals do when they behave in certain ways that affect conspecific behaviour, and about the extent to which there is continuity or similarity between our behaviour and theirs.

However, as things stand, this debate says more about what these various authors mean by 'symbol' than it does about similarities or dissimilarities in human and animal communication, because the word is used in so many different ways. Only if we can neaten up the sense of 'symbol' considerably, or at least be aware of who is using it in what way, will we be better situated to discuss the status of communicative behaviour, whether animal, proto-human or human, in the evolution of symbolic communication.

In this chapter, then, I will take a detailed look at some common definitions of the word 'symbol', tease out their implications, unpack how we might decide whether any given behaviour counts as symbolic, and discuss whether the definitions can thus be applied in a suitably scientific way to any communicative behaviour. This is an essential step if we rely on animal behaviour as evidence in symbol evolution, or if we want to examine just

what it was that evolved as our ancestors crossed the symbolic threshold.

Having given the background in the following section (§1.2)[1], the two most common definitions of communicative symbols (that symbols are arbitrary signs in §1.3 or conventional signs in §1.4) will be examined in turn and rejected. By 'rejected' I do not mean that they are wholly incorrect. They might adequately *describe* superficial properties of symbolic communication, but this doesn't necessarily mean they *define* what symbols really are, at least not in a way that could be applied usefully or objectively in offering some insights into how our ancestors evolved to use symbolic communication, or in deciding whether animal communicative behaviour is interestingly like ours in being symbolic. I conclude in §1.5 that symbols are best treated as requiring certain kinds of inference when they are first learned. For the rest of this dissertation, this will allow me to focus on inference as it relates to symbol learning.

1.2 Background to the problem

1.2.1 Basic issues

Jackendoff (1999) traces the evolution of language through a number of stages from the most primitive logical possibility up to modern language. He identifies stage one with a central claim in Deacon (1997): we are able to combine expressions (sounds, gestures or other forms) with content (or meaning) in a way that may well be unique among animals.

Usually this special combination of expression and content is called symbolic, but 'symbol' is a notoriously complex category (Noth, 1995). A range of definitions have been offered by psychology, informatics, philosophy, anthropology, ethology, linguistics, archaeology and semiotics. The problem is that language evolution is a place where all these fields meet, and this can make it difficult to agree what counts as symbolic communication.

Firstly, authors trained in one field may import a definition of 'symbol' from another field, but they (or their readers) might not be aware of subtle differences, of whether the definition is comparatively unorthodox, or what a definition entails. For instance, Deacon (1997) claims to have built his definition of 'symbol' on Peircean semiotics, and it is common for other authors

[1] Read 'chapter 1, section 2'.

to highlight this Peircean substrata in Deacon's thought (e.g. Garrod et al., 2007; Fitch, 2010). Deacon's usage departs in several ways from Peirce's (Lumsden, 2002; Sonesson, 2006), yet Fitch misattributes one of Deacon's innovations to Peirce. Comparable problems beset the term 'arbitrariness', with Joseph (1987) conflating the definitions in Hockett (1960) and Saussure (1959) and Seyfarth et al. (1980) conflating Saussure and Peirce.

Secondly, this surplus of academic heritage may lead to a scatter-gun approach: tying together a number of independent approaches rather than offering a single concise definition appropriate to the subject matter[2]. For instance, Hurford (2007) says at various stages that symbols are signs that do not involve resemblance, do not involve causation, but do involve learning. The relationships between these criteria need exploration to see if any one is more central than the others.

Thirdly, there are sometimes broad leaps between theory and data: it may be unclear just how a proposed test for symbolic communication relates to the offered definition of 'symbol'. At certain points, Noble and Davidson (1996) discuss symbols being conventional, but at other points they claim animal behaviour is not symbolic in that it is not intentional. However, they do not give an explicit account of how intentionality is related to conventionality. If there is a link, we should spell it out; if not, we should consider refocusing our definition on the empirical test rather than the theoretical background.

I think there is a lot that is right about various authors discussed above, but the useful threads of such claims need to be teased apart from the less helpful elements.

1.2.2 Common definitions

The various senses of 'symbol' can be divided into two broad classes: the internal or mental and the public or external. Internal symbols are mental representations (such as concepts) that are manipulated by cognitive processes (Fodor, 2001), while public symbols are communicative signs of a certain sort (such as spoken or written words), the precise nature of which is explored in this chapter. In what follows, I will use 'symbol' without

[2] All of what I say here is subject-specific: I am not arguing that all the above-mentioned disciplines must agree on a definition of 'symbol', just that language evolution needs a clear-cut definition to work with.

any modifiers to refer to public symbols and will use 'representation' as a shorthand for 'mental representation' to refer to internal symbols[3].

The two most common definitions of public symbols are as follows:

(1) Symbols are arbitrary signs

(2) Symbols are conventional signs

Allowing that the following authors may add other criteria to the above, **(1)** is found in Hurford (2004); Penn et al. (2008); Scott-Phillips et al. (2009); Smith (2004); Call (2006). **(2)** is found in Noble and Davidson (1996); Tomasello (1999); Zlatev (2008). A conjunction of **(1)** and **(2)** appears in Saussure (1959)[4]; Peirce (1955)[5]; Eco (1984); Tomasello (1990).

The following sections examine each of these in turn and evaluate their usefulness in deciding whether a human or animal behaviour is symbolic. The most general point in chapters 3-4 will be that these definitions focus on superficial aspects of symbolic communication, and that, at least when accounting for symbol evolution, we should rather focus on the cognition that underlies such communication.

1.2.3 Some semiotic background

A brief outline of some Peircean semiotic terms will allow for useful points of comparison between what may seem like unrelated points in this chapter, and will allow for more fine-grained descriptions than some of the accounts I critique here. Given the nature of Peirce's writing, it is often difficult to work out just what he meant by certain terms, and it is thus impossible to present any account as being definitively Peircean. Rather, it is better to talk of

[3]To be clear: I am using 'representation' here in the sense meant by cognitive science (for instance, some information-bearing unit that plays a causal role within a mind), not in the sense meant by semiotics, where it might be considered a loose synonym for 'signifier' and which would thus not be suitable here, given that I need something to refer to mental units. The term 'concept' is rather more loaded than 'representation' because some writers argue that concepts require language (e.g. Davidson, 1975; Dummet, 1993), but since I am discussing the evolution of language, and since I agree with Hurford (2007) that we need to allow for the possibility of non-linguistic representations if we want to avoid a saltationist account of language evolution, I intend 'representation' to be a more general category than 'concept' to avoid such issues.

[4]To an extent. Saussure uses 'symbol' to mean icon, and 'sign' to refer to what I am calling a symbol. More accurately, then, his claim is that *signs* are arbitrary and, as discussed below, by 'arbitrary' he mostly means 'conventional'.

[5]Peirce speaks of habits or general laws rather than limiting it to convention.

reconstructions of particular stages in the development of Peirce's thought or, charitably, to call a certain point Peircean if it is Peircean in spirit. The outline of the Peircean sign below focuses on mature developments and is based in part on Short (2004). The discussion of the term 'ground', found in earlier writings, is based on a reconstruction in Sonesson (2006).

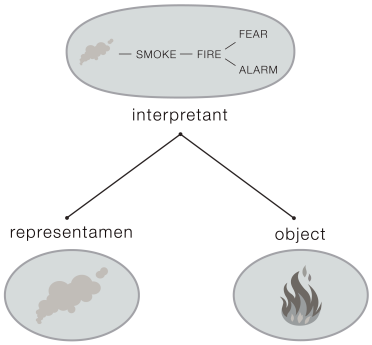
1.2.3.1 Signs

Peirce's sign is irreducibly triadic: it comprises three aspects or parts (interpretant, representamen and object), and the sign is the coming together of all three, rather than any one element (despite occasional terminological infelicities on Peirce's part). The representamen is that thing or event (say, a word, gesture or signal) that stands for something else, its object, and since smoke can stand for fire, the former is a representamen and the latter an object[6]. These two on their own, however, do not make a sign. A sign exists only when these two are connected by a third element, the interpretant.

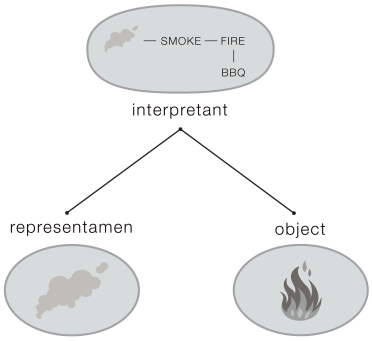
The interpretant is the effect that perceiving the representamen has on whoever is interpreting the sign, typically (but not always) involving the object being brought to mind. If you perceive smoke, that might initiate a series of consequences. First, your representation SMOKE is activated, and this probably in turn activates representation FIRE. Depending on your circumstances (say you've smelled it in the middle of the night in your apartment, rather than while standing around a barbecue) this may in turn have further effects: it may cause an emotion (fear) and may cause you to activate the fire alarm. The various effects that the representamen (perceiving smoke) has on you (activation of SMOKE and FIRE, feeling fear, and giving an alarm) are all interpretants, but two of the interpretants are representations, one an emotion, and one a behaviour. One of the interpretants is a representation of the object, hence bringing it to mind (fig. 1.1 *a*).

It is typical of Peircean semiotics that one interpretant may lead to others, as in this example. Some chains of interpretants may be rather short while others may be quite long. It is also typical that one interpretant

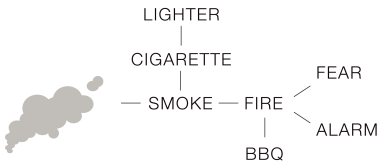
[6]These are not identical to the Saussurean terms 'signifier' and 'signified'. Both are mental entities for Saussure, but Peirce distinguishes various kinds of representamen and object, some of which are mental and some of which are external. These subtypes are not germane to my argument, so it makes no difference whether we consider the representamen to be your internal perception of a word on this page, or the physical printed word itself, or the type of which that word is a token.



(a) The smoke is unexpected indoors at 3am.



(b) The smoke is expected at a barbecue.



(c) A simplified excerpt from the representational structure stored in semantic memory that might underlie the above interpretants.

Figure 1.1: The semiotic triangle: interpretants link smoke to fire.

(sounding the alarm) can simultaneously be a representamen in a further sign: other people (who perhaps didn't smell the smoke) might hear the fire alarm and form similar interpretants. Finally, some interpretants are *potential* rather than necessary: smelling smoke while you're walking up and down the street trying to remember the address of the place you're supposed to be having a barbecue would play out in a different set of interpretants such that smelling smoke wouldn't cause the same fear interpretant (fig. 1.1 *b*). FIRE might be more predictably activated in both cases, presumably because it is closely associated with SMOKE in semantic memory (fig. 1.1 *c*).

1.2.3.2 Grounds

One further step that I think pertinent is the introduction of the term 'ground'. This is not a major feature of mature Peirce, but Sonesson (2006) introduces it into a discussion of the symbolic threshold in Deacon (1997). I define it here, and it will crop up again in my criticisms of arbitrariness or convention, but its relevance to the symbolic threshold is only made explicit in §1.5.2.3.

A weathervane indicates something about the wind to an interpreter, but does not indicate all possible facts about the wind. Rather, the weathervane only makes one aspect of the wind *salient*: its direction. An anemometer, on the other hand, can tell us nothing about the wind's direction, but only has the potential to signify something about the wind's speed. In Sonesson's interpretation, the ground is a relation between relevant aspects of the representamen (the aerodynamic shape of the weathervane, but not its colour or its looking like a rooster) and aspects of the object (the direction of the wind, but not its speed) such that the former can potentially be interpreted as a sign of the latter. The ground 'is really a principle of *relevance*' (Sonesson, 2006, 170, emphasis mine). Alternatively, '[i]n the Ground the object is seen *in a certain respect*, the attention isolates one feature' (Eco, 1997, 61). So despite connotations associated with the word 'ground', what we have here is more of a foregrounding than backgrounding. The key terms introduced here (salience and relevance) will recur frequently throughout what follows.

The ground is described as iconic, indexical or symbolic depending on whether representamen and object are connected by similarity of properties; a physical relationship such as contiguity or causation; or convention

or general law, respectively. A fire causes smoke under certain conditions, regardless whether anyone is around to observe it. Their relationship (an indexical *ground*) is thus prior to signification. It only enters into an indexical *sign* once someone actually interprets it as such: once the representamen has an effect (the interpretant) on an observer, potentially bringing the object to mind. I quite like the phrasing in Sonesson (2006): it is the causal ground that makes smoke an 'apt' representamen for fire.

Figure 1.2: Indexical grounds: tracks (source: http://commons.wikimedia.org/wiki/File:Elephant-tracks.jpg)

We must distinguish iconic/indexical/symbolic grounds from iconic/indexical/symbolic signs. It is wrong to treat everything causal as an indexical sign. Rather, as an indexical ground, it merely has the *potential* to be interpreted as an indexical sign. Consider the difference between a hunter and an urban youth seeing fig. 1.2. The hunter would interpret the circular depressions to be signs of an elephant. The urbanite might spot the tyre tracks, yet not see any signs of an elephant here. The ground is indexical in both cases, yet the elephant spoors are not an indexical sign of an elephant — are not a sign at all, in fact — to the youth. Similarly, vervets do not understand leopard spoor as signs of a leopard (Cheney and Seyfarth, 1990). Presumably the youth can learn to recognise spoor, though, while the vervets are somewhat more constrained by their biology.

1.2.4 Caveats and warning flags

Before looking at these definitions in more detail, I would like to mention a couple of distinctions that I will *not* be referring to much in what follows. Firstly, I will not have much time for exploring terms that various authors claim are somehow more basic than 'symbol'. Secondly, while the terms 'icon' and 'index' are frequently contrasted with 'symbol' in the literature, I do not consider it profitable (for my purposes) to rely overmuch on that distinction,

Sebeok (2001) lists various types of sign[7]: symptom, signal, icon, index, symbol and name. Symptoms are physical markings such as bird plumage, while signals are a matter of most animal communicative behaviour. However, I am more interested in seeing whether a given definition of 'symbol' is good enough to tell us something about language evolution than in trying to decide whether something non-symbolic is a symptom or a signal.

Secondly, the terms 'icon', 'index' and 'symbol' are not mutually exclusive, though Lyons (1977) misattributes this notion to Peirce. That is, a sign can be iconic, indexical and symbolic at the same time (Noble and Davidson, 1996; Keller, 1998). The typical sign on the door of a male bathroom may be considered reasonably iconic, but that does not prevent it from being indexical (in that its physical location is crucial to the meaning conveyed: it means male bathroom *here*, not just male bathroom in general) and symbolic as well (in that the picture resembles a woman in trousers as much as it resembles a man, and it is only convention that distinguishes this sign from the skirted female bathroom marker).

Further, Burling (1993) claims that iconicity is peculiar to humans and Sonesson (2006) argues that symbolicity may actually be phylogenetically prior to iconic and indexical signs. The upshot of all this is that we should be suspicious of negative definitions of 'symbol', claiming that symbols are signs that lack certain properties of icons (such as perceptual similarity) or indexes (such as physical contiguity), though I address this worry in more detail in what follows.

What I am aiming for in this chapter, then, is a way of characterising

[7]Sebeok uses 'sign' in a *very* broad sense here. An alternative would be to use 'sign' just to refer to the terms from 'icon' rightwards on Sebeok's list, where a interpreter distinguishes representamen from object. This terminological difference is unimportant for my purposes here.

what is special about human signs, taking 'symbol' to be typical of human signs. With this background in place, we can now turn to look at common definitions of 'symbol'.

1.3 Are symbols arbitrary?

Semiotics usually takes 'arbitrary' to mean 'unmotivated' (Saussure, 1959; Eco, 1984), so this section explores two senses of 'motivated'. The first sense (§1.3.1) involves perceptual similarity, a feature of icons. A map is motivated: it can be used to work out how to get somewhere because properties of the map correspond to properties of the environment. A second sense (§1.3.2) involves causation, a typical feature of indexes. A weathervane is motivated: it can be used to work out the direction of the wind since there is a physical or causal connection between the two.

But there is no similar motivation behind the symbols 'map' and 'weathervane'[8]. We will thus be dealing, throughout this section, with a negative definition of 'symbol': whatever 'motivated' turns out to mean, symbols are not that.

1.3.1 Motivatedness as perceptual similarity

Hockett (1960) posited a number of design features of language. One of them, arbitrariness, requires that linguistic symbols are non-iconic[9]: that is, there is usually little perceptual similarity between words and their referents, or between representamen and object[10].

First, I show that iconicity is subjective (§1.3.1.1). While this is not a problem for many researchers in iconicity, it is a problematic base for our definition of 'symbol' since we need an objective definition to account for symbol origins. Thereafter, I compare icons based on straightforward

[8] 'Weathervane' can be further broken down into 'weather' and 'vane', and since a weathervane does have something to do with the weather, it might be claimed that 'weathervane' is motivated in a way that 'map' is not. The question of this higher level arbitrariness ('relative arbitrariness,' Saussure 1959) takes us into questions of systematicity in linguistic form, which is quite distinct from the question of arbitrariness in the semiotic ground and which will not concern me further.

[9] There are divergent interpretations, but this quite a common one.

[10] This may sound like Saussure's point about the arbitrariness of the linguistic sign, but is the sort of misconception I identified in §1.2.1. Saussure eventually stated that by 'arbitrary' he meant 'conventional', and not everything non-iconic is conventional.

perceptual similarity with those requiring more sophisticated cognition and argue that a reliance on perception obscures an important cognitive difference, and that this cognitive difference is more important for symbol evolution (§1.3.1.2). Finally, I look at what role iconicity usually has in language evolution, and show this role is of secondary importance, compared to cognition (§1.3.1.3).

1.3.1.1 Iconicity is subjective

The Old English (OE) word for cuckoo was '*geak*' /jæ:ak/, which appears to be less motivated by phonological similarity than the modern form 'cuckoo', a later borrowing into English from Old French *cocu*, derived from Latin *cuculus*. However, the OE form derives from proto-Germanic **gauk-a-z* (Roger Lass, personal communication), the root of which is transparently onomatopoeic.

The transition from **gauk-* to /jæ:ak/ was gradual, so there must have been periods when the onomatopoeic motivation of intermediate forms was debatable: for some speakers it may have been more iconic than for others. If symbols are a matter of arbitrariness and if arbitrariness is a matter of lacking iconicity, we would be forced to claim that the symbolic status of an intermediate form between **gauk-* and /jæ:ak/ depended on how much it reminded any given person of the noise of a cuckoo.

While it may be unproblematic to *describe* some of these forms as being more or less iconic than others, this subjectivity causes two problems for *defining* 'symbol' as unmotivated in this sense, both of which may be illustrated by an analogy.

Focusing on perceptual similarity in such cases is like trying to decide whether people are related purely based on how similar they look. Siblings may well look similar and this may well suffice for every-day purposes, but their parentage or DNA, not their appearance, is what makes them siblings. One implication is that the former is what someone interested in objectivity, such as a judge or scientist, should investigate. A second is that it is the former that explains the latter, not the other way round.

If this is our definition of a symbol, then the symbolic threshold occurred when our ancestors evolved from understanding iconic to understanding non-iconic signs. But since iconicity is subjective, the status of a given sign at an intermediate pre-symbolic stage would have depended on how much it

resembled its referent in the subjective opinion of each individual hominid. So two hominids may have been able to communicate using the same sign while it was symbolic for one and non-symbolic for the other. Unless we were to investigate individual differences in this regard, we would reach a stalemate in accounting for the evolution of symbolic communication. But the relevant individual differences must have been cognitive. If these admit objective measurement, then it is these cognitive differences, not perceptual similarity, that should be central to our definition of 'symbol'. It is also these that would explain how an individual or a species understands either symbolic or non-symbolic signs.

I don't deny that icons may evolve into less motivated symbols over time, and I examine this process in ch. 7, but I demonstrate experimentally that iconicity does not play an explanatory role in the dynamics of the process, though this doesn't prevent it having an important role in how the process begins.

1.3.1.2 Iconicity privileges perception over cognition; linguistic meaning demands a focus on cognition

Developing this idea further, I turn to look at a range of icons which lie along a continuum between comparatively straight-forward perception and cases where perception is merely the input to some rather complex cognition. If our definition of symbols is that they are arbitrary in the sense of non-iconic, then we have no way to distinguish the following examples *vis-à-vis* the symbolic threshold. I argue that this would be problematic given that meaning is a cognitive phenomenon, and that our definition must therefore reflect these differences.

To illustrate these concerns, consider the following. When looking at a clear photo, it is usually easy to recognise the face of someone we know in it. Perceptual similarity allows us to make a connection in our mind between the photo and the person: we can *see* they are the same, though we do not mistake the photo for its referent (unlike pigeons, Sonesson, 2006), so even at this stage there may be something more sophisticated going on in the human brain. Regardless of such a possibility, this is a comparatively simple task and it is more a matter of perception than cognition, in so far as there is a difference.

When looking at a well drawn caricature of that person, there are still

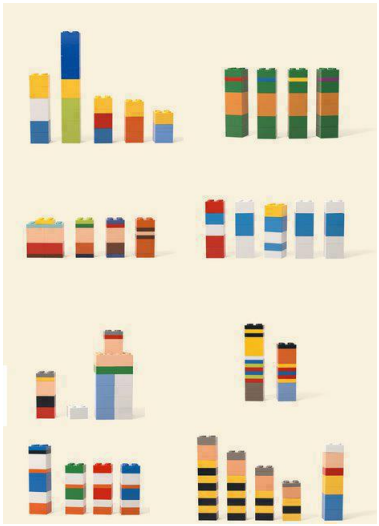


Figure 1.3: A problem of perceptual similarity[11] (Jung Von Matt, 2012).

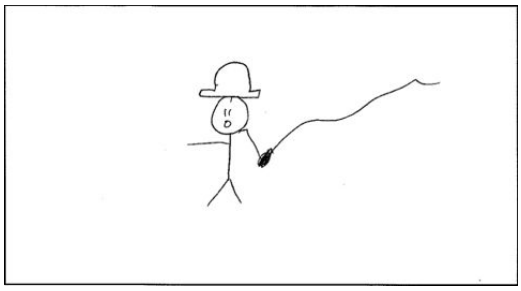


Figure 1.4: A Pictionary clue from an experiment in ch. 7[12]

properties shared by the caricature and the person: we can see they are (somewhat) the same. But the process is not merely a matter of seeing: we make a connection between certain (probably exaggerated, salient) features of the caricature and the referent, and so some cognitive process draws our perceptual attention to those features. This is not as straight-forwardly perceptual as a clear photo.

A step further in the same direction, it can take some people quite a while to recognise what some of the stacks of lego blocks in fig. 1.3 depict, and out of context, one might have even less success with fig. 1.4, drawn

[11] Answers: The Simpsons; Teenage Mutant Ninja Turtles; South Park; The Smurfs; Asterix and Obelix; Bert and Ernie from Sesame Street; Donald Duck with Huey, Dewey and Louie; Lucky Luke and the Dalton Gang.

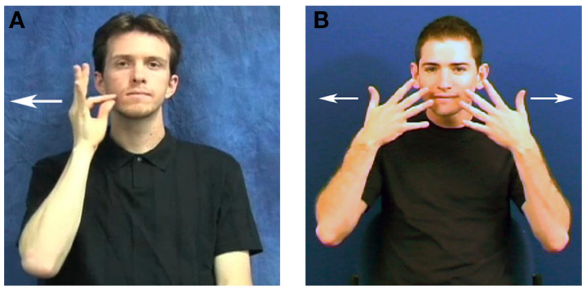


Figure 1.5: (A) ASL and (B) BSL signs for CAT (Perniss et al., 2010).

in a Pictionary-like experiment from ch. 7. Some people see the answer almost immediately, others may feel, after staring blankly at these pictures for a while, a sudden flash of surprise as they hit the answer. This is typically called an 'Aha!' moment, or a flash of insight (the presence of which in such situations I demonstrate experimentally in ch. 5). This may also be conceived of as a non-trivial form of analogy since the mind has to do some work to make a connection between properties of these lego bricks and properties of the things they represent[13].

Two individuals may have similar perceptions of these figures, but differ in how their brains retrieve the relevant memories or make connections between salient elements. Even though (it turned out) I had a visual memory of all these referents prior to seeing this figure, I was unable to retrieve those memories for at least two of the sets, and had to look at the answers.

Fig. 1.5 shows American Sign Language (ASL) and British Sign Language (BSL) signs for CAT. The gesture in each case shares perceptual features with the referent, though these are stylised or abstracted somewhat: despite differing hand shapes, they resemble a cat's whiskers and are thus iconic (Perniss et al., 2010). Nonetheless, Perniss et al. claim that these iconic features are often not accessible to people unfamiliar with those particular languages and anyone who has played charades will know that iconicity is not always cognitively accessible.

The photo, caricature and other figures are all iconic, but the presence of

[12] Answer: Harrison Ford.

[13] Gentner (2010) claims (briefly) that analogy and insight are related, and I will discuss their relationship and their role in symbol learning in ch. 3.

perceptual similarity in these cases is not a sufficient reason to suppose that the same cognitive processes are applied in interpreting them all: perceptual features are inputs to increasingly complex process of analogy, insight or inference in fig. 1.3 and fig. 1.5, but not so in the photo. If symbols are nothing more than arbitrary (i.e. non-iconic) signs, this causes problems for language evolution: such a definition focuses on the perceptual dimension at the expense of the cognitive one.

We have learnt through a number of ingenious experiments (originating with von Frisch, 1994) that one form of bee dance involves moving in a figure of eight, indicating the direction and distance to flowers as a source of food. The angle between the vertical on the hive wall and the waggle run (the crossing of paths in the middle of the figure of eight where the bee waggles its body) corresponds to the angle between the sun and the source of nectar. The duration of the waggle run corresponds to the distance of the source from the hive. This similarity of properties means the ground is iconic.

Vervet monkeys emit three kinds of alarms (bark, grunt or chutter) in response to three kinds of predators (leopards, martial eagles and pythons, respectively, Seyfarth et al., 1980). The monkeys respond to these alarm calls with the same avoidance behaviour they would display to signs of an actual predator (such as a sighting, a leopard growl or an eagle shriek), and the avoidance behaviour to both monkey alarms and predator signals is environmentally appropriate for escaping those predators: in response to an eagle sighting, shriek or alarm, they look upwards and run for cover in dense bush; in response to a leopard sighting, growl or alarm, they run up trees. 'The monkeys behaved as though each alarm call designated specific objects or events in the external world' (Seyfarth et al., 1980, 1090). Seyfarth et al. note that these calls are non-iconic[14].

If symbols are defined as being non-iconic, then we must class the sign language gestures above with bee dances as being non-symbolic and vervet alarms with words as being symbolic. But the way this divides up these communicative behaviours is entirely counter-intuitive if we want to account for how a non-symbolic species evolved into a symbolic one, and I take it that this includes accounting for how humans but not other species (with few exceptions, after laborious training by linguistic humans) can understand

[14]Rendall et al. (2009) point out that an aspect of the alarms is nonetheless non-arbitrary: short bursts capture attention effectively.

sign-language gestures and words. Sign-language gestures may be iconic, like bee dances, but the former is linguistic and the latter not; the same point applies, *mutatis mutandis*, to non-iconic alarm calls and words.

A sensible way to distinguish iconic sign-language gestures from iconic bee dances, or arbitrary words from arbitrary alarm calls in the way required by this evolutionary question, is by appealing to cognitive differences. Our definition of 'symbol' must reflect this fact: we should define the term relative to whatever cognitive processes this perceptual information is an input to.

On a related note, it has been established that perceptual similarity exists in bee waggle dances, but the existence of an iconic ground does not mean that bees always interpret that particular ground in order to find the flowers. Grüter and Farina (2009) summarise a range of experiments showing that different individual bees perceiving the same waggle dance find the relevant flowers for a range of reasons. Some of these bees react to olfactory cues given off by the dancer and then follow an odour trail, rather than the angle of the waggle run. Other bees had visited the indicated source the previous day and simply had their memories of the place reactivated by olfactory cues. A focus on perception, prompted by the definition currently on offer, risks seeing iconic signs where there may be only iconic grounds (cf. §1.2.3.2).

1.3.1.3 Iconicity plays, at best, a supporting role in symbol evolution

The previous subsections have criticised the notion that iconicity could be the basis for a negative definition of symbols, but didn't examine why language evolution researchers might be interested in iconicity at all. Here I outline one role iconicity may play. I then argue that this role is nonetheless subordinate to cognitive considerations.

Hurford speculates about two possibilities for the first proto-linguistic words. One possibility is that

> some hominin ancestor made a random noise while attempting to convey some idea; the hearers were able to make a good guess from the context what idea was meant, and the random noise became arbitrarily associated with that idea. (2010, 121)

I examine in the next two chapters just what this 'good guess' might have been. The other possibility is that

> Some slight element of naturalness [read: motivatedness] in connections between meanings and sounds could have been the bootstrap needed to get such a system up and running. (2010, 127)

That is, a potential role for iconicity in language evolution is that it could have facilitated the learning of connections between representamen and object because this is easier to process (learn, interpret or recognise) than is the case with arbitrary connections, for various reasons. Accounts tend to focus either on the fact that iconicity might be more informative or that there may be cross-modal associations supporting such learning.

As an example of the first of these, consider the Japanese mimetic verbs in table 1.1. Kantartzis et al. (2011) show (using other examples) that iconicity aided acquisition for English speaking children by comparing learning rates for such Japanese verbs when matched with their meanings and learning rates for mismatched verbs.

Japanese mimetic word	English translation
koro	light object rolling
korokoro	light object rolling repeatedly
goro	heavy object rolling
gorogoro	heavy object rolling repeatedly

Table 1.1: Iconicity in Japanese words (Perniss et al., 2010).

Similarly, Thompson et al. (2009) presented ASL users with ASL signs accompanied by one of two pictures, one with the iconic ground made salient and the other lacking such salience. For example, the ASL sign for banana involves miming peeling a banana, so the salient picture showed a half-peeled banana while the non-salient picture showed an intact banana. Subjects pressed keys to indicate whether a simultaneously presented gesture and picture represented the same object, and the results show that subjects reacted more quickly in the salient condition. So salient iconicity facilitated processing. Thompson et al. (2012) show that iconicity aids child BSL signers learn signs.

Moving to the second point, there exist several demonstrations of associations across sensory modalities. Ramachandran and Spence (2001) inves-

tigate the *bouba-kiki* effect. Subjects were presented with a rounder and a spikier shape, and asked which they thought was called *Bouba* and which *Kiki*. They were significantly more likely to associate *Bouba* with the round shape and *Kiki* with the spiky one. Simner et al. (2010) had subjects move sliders which altered phonetic properties of a vowel sound they heard. For instance, lower F1 frequencies correspond to higher vowels. One of four basic tastes (sweet, salty, bitter, sour) were dropped on their tongues and they were asked to manipulate the sliders to produce the sound they thought best matched the taste. Although this seems a highly abstract and subjective task, there were systematic associations. For example, sweet tastes were associated with lower F1 frequencies than sour ones.

When the above two points (ease of processing and cross-modal associations) are applied to language evolution, researchers are often explicit about the fact that iconicity plays a role in so far as it provides 'scaffolding for the cognitive system to connect linguistic form and embodied experience' (Perniss et al., 2010, 12). Similarly, Hurford claims 'the very first learned meaningful expressions would have been sound-symbolically, or synaesthetically, connected to their meanings, facilitating their learning and diffusion through the community' (2010, 127), in line with his second quotation above. In either case, our focus should be on learning, or the cognitive system generally.

The above studies do not focus explicitly on symbols, but if symbols are defined as arbitrary in this sense, then the same conclusion applies: we should define symbols relative to features of cognition, not a lack of perceptual similarity. In other words, iconicity and arbitrariness focus on the ground, while accounting for how humans use linguistic meaning needs an account of how we interpret these grounds, of how our minds handle interpretants, whose role it is to connect representamen and object to produce signification. Just to be clear: the claims here do not argue against the possibility of iconic stage prior to the symbolic threshold. Nor am I denying that this would have involved a change from something iconic to something non-iconic. What I am arguing is that an explanation of this change must make reference to cognitive differences.

1.3.2 Motivatedness as causation

It might be argued that perceptual similarity is too literal an interpretation of Hockett's design feature, and that there are more abstract forms of motivatedness that are relevant, either instead of or in addition to perceptual similarity. I examine one such option, indirect iconicity (§1.3.2.1), before showing that we are really talking about causation here, which I relate to mental processes in communication (§1.3.2.2). I then frame the discussion in terms of chains of interpretants (§1.3.2.3), and conclude that this form of motivation, like the previous, refocuses our attention on cognition. In particular, I discuss the open-endedness of human learning when it comes to symbols.

1.3.2.1 Direct and indirect iconicity

I shall call the literal iconicity of the previous subsection 'direct' and the more abstract form of this subsection 'indirect'. Direct iconicity involves a representamen and its object sharing properties. The angle between the vertical on the hive wall and the waggle rush in a bee waggle dance is the same as the angle between the sun and the flowers; the colours in fig. 1.3 are the same as the colours of the characters they depict.

Indirect iconicity, on the other hand, involves properties of a representamen correlating with *different* properties of the object, in what are called 'graded signals' (Wilson, 1975) or 'analogue communication' (Burling, 1993). These are typical of animal communication, as well as non-linguistic aspects of human communication. For example, Davies and Halliday (1978) show that larger toads of the species *Bufo bufo* have deeper pitched croaks than smaller con-specifics due to differences in their physiology, and that such calls influence female mate choice by signalling the size of the male. This is not direct iconicity, since there is no shared property; rather it is indirect because properties of the signal (frequency) correlate with, and in this instance are caused by, properties of the content (size of the male). Similarly, the length of the second unit of a domestic chicken's alarm call is predictable from the angular size of an aerial object that might turn out to be a predator (Gyger et al., 1987)[15].

[15]Since this section intends to move the discussion away from iconicity and onto psychological considerations, these terms are not intended to be permanent additions to an already tangled lexicon. Rather, they are introduced here to highlight some vagueness

Laughter may convey something about a person's emotional state, and this signal can fall on a continuum from a giggle to a guffaw (Burling, 1999): typically, the funnier the situation, the louder or longer the laugh. These aspects of human communication are non-linguistic in the sense that drawing out the adjective in 'how boooooooring!' may emphasise the degree of emotion without making a semantic distinction between boredom and any other feeling.

Indirect iconicity is relevant to language evolution because it motivates the distinction in Burling (1993) between two systems of human communication: digital language and an analogue gesture-call system such that we share only the latter with other animals. Burling notes that these gesture-call signals are more difficult to control than linguistic signals, and I will return to this question of volitional control by the end of this section.

Despite appearances, though, indirect iconicity actually takes us away from perceptual features and bring us to another form of motivatedness: causation. I will have to unpack this somewhat before returning to the question of arbitrariness and symbols.

1.3.2.2 Indirect iconicity and causation

The toad calls are straightforwardly a matter of physical causation: the laws of physics mean that larger vocal tracts produce deeper sounds, and larger toads have larger vocal tracts. The chicken alarm calls may seem less like physical causation, but the difference between these and the frog calls is one

or overextension in the term 'iconicity' in the language evolution literature. For example, both Burling (1993) and Arbib (2012) acknowledge that there are subtle differences among icons, but both use 'iconic' in a way that fails to distinguish what I call direct and indirect icons, which draws focus away from important cognitive differences that will be explored throughout this dissertation. It may also be worth briefly pointing out why these terms do not reduce to established usages such as imagistic vs. diagrammatic iconicity (Peirce, Collected Papers 2.277, 1903) or primary vs. secondary iconicity (Sonesson, 2006). Peirce's images involve shared perceptual properties and thus correspond to 'direct icons' here, but his diagrams involve internal structure and correspondences between such structures. The examples of indirect iconicity above do not have any such structures and are thus not diagrammatic icons. Sonesson's primary icons can straightforwardly resemble their objects, while the resemblance in secondary icons usually needs to be pointed out before it can be appreciated. For instance, person A might not notice that a certain cloud looks like a whale until person B labels it as such, at which point A might be able to spot the resemblance. This is nonetheless a matter of perceptual properties and thus unlike indirect iconicity, which does not involve shared perceptual features as discussed above. Further, the chicken case is not indexical, though the toad case is, so 'indirect iconicity' is a category independent of 'index'. See §1.3.2.3 for more details.

of degree, not kind.

Causation is a purported relationship between sets of events, such that some (causes) are followed by others (effects). This relationship is often conceived of as a universal law, such that an effect follows with 100% certainty from a cause (Thagard, 2007). However, Thagard points out that we are comfortable applying the term 'causation' even when causes only lead probabilistically to their effects: 'infection by a mycobacterium causes tuberculosis, but many people infected by it never develop the disease' (Thagard, 2007, 15). So rather than stipulate that effects follow causes with 100% certainty, some philosophical positions allow degrees of probability lower than 100%: universal laws are only extreme versions of probabilistic laws, rather than a different kind of law. In what follows, I use 'causation' in this broader sense.

Not all probabilistic relationships are causal, however: Thagard notes that the probability a patient has tuberculosis given that they take the drug Isoniazid is higher than the probability they have tuberculosis given they do not take this drug. But that is because Isoniazid is used in the treatment of tuberculosis and is not a cause of that disease. As the well known chestnut has it, correlation is not causation. In the case of tuberculosis, it is possible to research the biochemical processes by which a mycobacterial infection leads to the disease, while the role of Isoniazid is clear from the relative timings of events: a patient gets sick, goes to the doctor, and *then* gets treated. Though there are other gaps that may need to be filled, such explanatory considerations are helpful in deciding that the mycobacterium plays a causal role while Isoniazid does not.

But what are the relevant explanatory considerations in the chicken case? That depends mostly on the level of analysis relevant to your scientific field[16]. A neurologist might be interested in how some chicken neurons related to perception of a predator cause activation in neurons related to alarm-call behaviour, or how various brain networks interact. A biologist might be more interested in the function of alarm calls with respect to predator avoidance in the evolutionary history of the species. Though I will occasionally refer to neurons, brain networks and biological function throughout this dissertation, I am more interested, on the whole, in the psychological mechanisms linking perception to behaviour.

[16] I examine levels of analysis more explicitly in ch. 2.

Simplifying for the sake of illustration, the chicken perceives a predator, which causes activation of their representation of the predator, and in turn this representation causes avoidance behaviour or alarm calling. That is, there is a probabilistically causal story linking perceptual input to representations and a similarly probabilistically causal story linking representations to behavioural output.

Millikan describes these as pushmi-pullyu representations: 'they at the same time tell what the case is with some part of the world and direct what to do about it' (2005, 20), for instance representing both the predator and the relevant avoidance behaviour. Millikan also gives the bee dance as an example of this, and Hurford (2007) expands this point to vervet alarm calls. All these behaviours are thus fairly predictable given the relevant inputs. Indirect iconicity thus cuts across animal communicative behaviours differently from how direct iconicity did. In the case of direct iconicity, bee dances were not symbolic and vervet alarms were, which I argued was problematic. But in the case of indirect iconicity, both the dance and the alarm are motivated and thus non-symbolic.

Since these communicative behaviours are comparatively predictable, we should decide whether symbols are less predictable, or in what sense they are less predictable. This take on arbitrariness needs an account of causation in cognition underlying communicative behaviour, some principled way of distinguishing comparatively causal from comparatively open-ended interpretation[17]. I argue in the next section that interpretants play an important role here.

1.3.2.3 Causation and interpretants

Interpretants are useful in avoiding a potential confusion. Grounds involving causation are indexical, but not everything causal presumes an indexical ground; nor, indeed, an indexical sign. A leopard growl and the leopard that made it are related in an indexical ground. But the fact that a growl causes activation of a vervet monkey's LEOPARD representation is a matter of the relationship between representamen and interpretant, not representamen and object (fig. 1.6 *a*). On the other hand, a chicken that reacts to an overhead silhouette of an eagle is reacting to an iconic ground (the silhouette

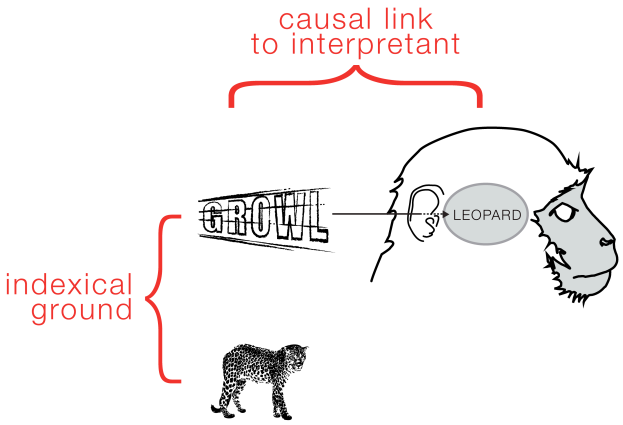
[17] We cannot, however, distinguish clear-cut cases of causation or of arbitrariness, given that I started with a graded notion of causation.

shares perceptual features with the eagle predator). If this representamen causes activation of the chicken's PREDATOR representation, then this causal relation between representamen and interpretant is nonetheless part of an iconic, not indexical, sign (fig. 1.6 *b*). The same may be said for iconic bee dances. Talk of resemblance or causation without this framework risks confusing different aspects of the sign, and Sonesson (2006) identifies such a conflation in Deacon (1997).

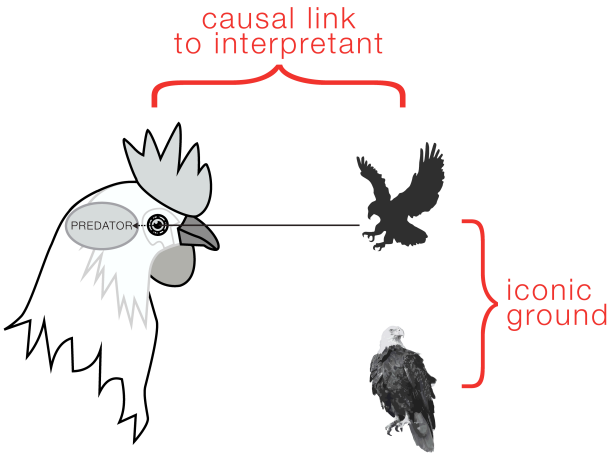
Discussing the synaesthetic and sound-symbolic accounts above, Hurford (2010) highlighted the 'naturalness' of the connection across sensory modalities. Applying this to the example from Simner et al. (2010), there may be a predictable relationship between certain vowel features and tastes. This is not a directly iconic ground (it involves different properties), nor an indexical ground, nor is it even a sign (sweetness does not signify a higher vowel). But it is about probabilistically causal links in cognition, which is the sort of motivatedness I'm examining here. So motivatedness can be abstracted away from signification, as might be expected if it plays a role in getting a signification system started, as Hurford claims.

Compare a child learning English words with an English-speaking adult. If both hear 'fire!', it is fairly predictable that this will cause activation of the adult's interpretant representation FIRE (and this interpretant representation may in turn weakly or strongly cause an interpretant emotion, fear, or interpretant response, escape), though the ground here is conventional. If the fire is visible at the time of utterance, the child's concept FIRE may be activated at that time, but this concept is not an interpretant of the word 'fire' until that word is learnt.

The process of learning is the process of connecting that particular concept to that particular representamen, at which point the concept becomes an interpretant, which means that the conventional ground is now the base for a conventional sign for that child. So while a causal link between representamen and interpretant is provided by evolutionary history in the case of bees, chickens, toads and vervets, it is the result of learning in human communication. The conclusion is that human symbols, in being unmotivated, are more open-ended than animal communication. If symbols are unmotivated signs, then it is the learning process we need to focus on in deciding what counts as a symbol. In particular, we eventually need to discover what was special about learning at the symbolic threshold. I examine this in §1.5



(a) Two causal relationships: one between object and representamen, the other between representamen and interpretant.



(b) An iconic relationship between object and representamen, and a causal one between representamen and interpretant.

Figure 1.6: Causation in grounds is distinct from causation in interpretants.

and in the coming chapters.

For now, though, two sets of terms can help anchor the above discussion. Firstly, Gärdenfors (1995) distinguishes cued and detached representations. The former are triggered by something currently (or recently) present in the environment; the latter are not cued in this way. That is, detached representations allow us to think about things not currently present. Since representations are one kind of interpretant, we can thus distinguish cued and detached interpretants. Under the current sense of 'motivated', perhaps symbols are unmotivated in that they have a detached interpretant: a word allows us to communicate about things not currently present. The question then becomes how interpretants ever get attached, either to representamens or to objects, in learning.

Secondly, Sperber and Wilson (1995) distinguish central and peripheral processes. The latter are involved in processing perceptual input or producing behavioural output; the former are those relating representations to other representations. The cued representations in fig. 1.6 involve only peripheral processes, for instance. The connection between smoke and SMOKE in fig. 1.1 is peripheral, while the connection between SMOKE and FIRE is central. Ignoring the fact that this was an example of an index[18], a central processes is needed to connect SMOKE with FIRE. Perhaps, then, symbols are unmotivated in that they, too, need central, rather than peripheral, processes, to link interpretants with objects. Since causation is a matter of inference, this is an initial suggestion that symbolic meaning might also be a matter of inference.

1.3.3 Conclusions about motivatedness

I have unpacked two accounts of motivatedness upon which we might base a negative definition of symbols as being unmotivated or arbitrary. I argued that the first, iconicity, is unhelpful concerning the symbolic threshold in that it relies on subjective and superficial criteria and yields counter-intuitive or misleading conclusions about evolution. The second, causation, offered more useful points about the threshold. Causation here is not about indexicality in grounds, but rather a matter of relationships between representamen and interpretant, or between interpretants. While it is descrip-

[18] After all, in fig. 1.6, the ground's iconicity or indexicality was orthogonal to the fact that both interpretants were cued by the representamens.

tively accurate to claim that symbols are unmotivated in this sense, the discussion drew our attention to cognition, particularly the open-ended nature of learning, raising the question of what was special about learning at the symbolic threshold.

I don't think much would be gained by trying to rescue motivatedness by claiming that symbols are not iconic, not indexical and learned. This is the scatter-gun approach I mentioned in §1.2.1. A series of negative definitions would fail to narrow in on positive features of symbols, leaving us unsure just what evolved at the symbolic threshold. If just the last of this raft of ideas is the most important, then the addition of iconicity and indexicality muddies the water, especially given the potential confusion between indexes and other causal but non-indexical processes in signification, or about differences between direct and indirect iconicity.

There still remains the possibility, though, that symbols are just conventional signs. This might obviate the need for this cognitive slant, and the possibility must be discussed before I continue with my cognitive approach.

1.4 Are symbols conventional?

Apart from claims about cognition, the previous section drew the focus to the open-endedness of human behaviour as opposed to comparatively more causal animal behaviour. This section examines whether 'conventional' is a good way of describing that open-endedness. That is, we shift away from examining whether symbols are *unmovitated* to look at whether they are signs *motivated* principally by convention.

But what precisely makes something a convention? And what objective criteria can we apply to decide whether something is conventional? 'Conventional practices are a widely accepted feature of the social environment, but what they are seems unclear' (Latsis, 2005, 11). The following subsections take a closer look two influential accounts of convention.

Lewis's approach (§1.4.1) is based on the question of how conventions arise and become stable by rational means, but I will give various reasons for thinking that Lewis conventions alone cannot account for the evolution of a symbolic species without an account of the evolution of particular kinds of inference. Some of these concern *salience*, and others are quite nebulous, prompting a need for a clearer inferential framework. Millikan's take on

convention (§1.4.2) is motivated more by biological than by logical concerns and is thus potentially suitable for the evolution of language, but it centres on imitation and raises the problem of inferences about *relevance*.

Incidentally, it also turns out that conventions are arbitrary in a particular sense (different to the unmotivatedness of the previous section), which I will briefly examine before moving on to look at Lewis and Millikan.

Keller (1998) illustrates this claim, made by Lewis (1969), by imagining a village that has two wells. An outside observer notes that all the villagers only ever use one of the two wells and then asks himself whether the behaviour is conventional. If the observer discovers there is some property of one well that provides a reason not to use it (it is too far away, or the water does not taste nice) then he would have no reason to call the villagers' behaviour conventional. But if both wells are equally suitable though only one is in fact used, then the observer would be justified in hypothesising some convention at work. Perhaps the villagers know they will have the opportunity for gossip if they all congregate at the same well every day, but this is preference does not relate to a property of the well itself: they could gossip at either location, as long as they meet at the same one.

So conventional things are arbitrary in that some other action could equally have been chosen relative to a particular purpose, such as collecting water (Keller, 1998), and if anything motivates one choice over another in such cases, it is something like precedence, or the solution to a coordination problem, or an explicit agreement. That is, arbitrariness in a positive sense (unlike the negative senses of the previous section) focuses on the availability of choices. Convention (whatever it turns out to be) is what settles on and perpetuates a particular choice.

While a vervet cannot choose to give anything other than a bark in response to a leopard, de Saussure's (1959) discussion of arbitrariness dwells on the fact that a dog could equally well have been called '*chien*' or '*Hund*', or that some people choose to call a tree '*arbre*' or '*baum*'. Whether 'dog' sounds like a dog or not is less central to Saussurean semiotics, though I think this has been distorted somewhat by Hockett's influence, as I discussed in §1.2.1.

The village-well example prompts one constraint on this positive sense of 'arbitariness': arbitrariness is relative to a given purpose or function[19]. A

[19]This is not a feature of Lewis's original discussion, but rather an addition by Keller.

coffee cup has to be capable of holding hot liquid and be useful for drinking. Relative to this function, it is arbitrary whether the sides are straight or curved; the colour and pattern are also arbitrary. Relative to a different goal (matching your kitchen colour scheme), the colour would not be arbitrary.

If we are to take Saussure's claim about arbitrariness at face value, and if arbitrariness involves the availability of choices relative to some goal, then his example of different words for 'dog' or 'tree' suggests that a good way to conceive of the primary function of symbolic communication is the bringing to mind of an interpretant representation in one's interlocutor. Relative to this goal, it does not matter if we say 'tree', '*arbre*' or '*Baum*' as long as we have some basis for thinking that our interlocutor knows the convention motivating our usage such that their representation TREE is then activated.

1.4.1 Lewis conventions

.

According to a well known but early formulation, Lewis states:

> A regularity R in the behavior of members of a population P when they are agents in a recurrent situation S is a *convention* if and only if it is true that, and it is common knowledge in P that, in any instance of S among members of P,
>
> 1. everyone conforms to R;
> 2. everyone expects everyone else to conform to R;
> 3. everyone has approximately the same preferences regarding all possible combinations of actions
> 4. everyone prefers that everyone conform to R, on condition that at least all but one conform to R;
> 5. everyone would prefer that everyone conform to R', on condition that at least all but one conform to R';
>
> Where R' is some possible regularity in the behaviour of members of P in S such that no one in any instance of S among members of P could conform to both R' and R. (Lewis, 1969, 78)

According to Keller, Lewis's definition means that 'similar behavior among the members of a group is called "conventional" if, for every individual, the only reason for choosing exactly this kind of behavior is that

each person thinks that the others will do the same' (1998, 137). According to Vanderschraaf, 'a convention is a state in which agents coordinate their activity, not as the result of an explicit agreement, but because their expectations are aligned so that each individual believes that all will act so as to achieve coordination for mutual benefit' (1995, 65).

As an example, consider driving on the left or right side of the road. It does not matter which we do, as long as we do the same as others and expect others to do the same as us. Lewis's criteria 4 and 5 express this requirement of arbitrariness as illustrated above in Keller's example of the village well. That is, Lewis treats conventions as game-theoretic coordination problems[20].

Millikan notes that 'almost no clause of Lewis's analysis has withstood the barrage of counterexamples over the years' (2005, 1). I will show that even if we accept Lewis's definition of convention, it would not be useful to *define* symbols as being conventional in this way at the symbolic threshold, though they may be *described* as being conventional among modern humans.

I begin by outlining two sub-types of convention, one involving co-ordination of action and the other co-ordination of action and belief; I call the latter symbolic convention (§1.4.1.1). I then examine the role of salience in Lewis conventions and show that, while his account may be able to explain how conventions are stable, it needs to complemented by an account of inferential salience to be able to explain how symbolic conventions arise (§1.4.1.2). Thirdly I look at some of Lewis's alternatives to salience, and show that these rely on forms of inference left rather vague in his discussion (§1.4.1.3). The previous two points are concerned with how a given species reaches a convention, but not with how that species evolved to handle convention in the first place; so finally, I show that Lewis conventions are unsuitable for discussing the symbolic threshold (§1.4.1.4).

1.4.1.1 Two types of Lewis convention

Though the above formulation is the most commonly discussed, Favereau (2008) points out that Lewis later revised his definition and that this update is usually ignored. In particular, Lewis later realised that convention is not always just a co-ordination of action, since some types can also require co-

[20]In other words, the availability of choices can be rephrased as requiring that conventions are arbitrary in that they require the existence of more than one strict Nash equilibrium (Sillari, 2008; Cubitt and Sugden, 2003).

ordination of belief (1983) and Favereau argues that these (action alone vs action and belief) are different types of co-ordination problem and thus different types of convention. I call the latter 'symbolic convention', and will refer to 'interpretants' rather than 'beliefs'.

If you drive on the left, I should drive on the left, too: we coordinate action and I can straightforwardly see which side you are driving on. But when it comes to symbolic conventions, I cannot see what you are doing as straight-forwardly. If I say 'dog' and you say 'dog' then we performing the same action, but symbols need more than this: a mynah bird could achieve as much. As discussed in the introduction to §1.4, to achieve our communicative goals, we must also have some expectation of co-ordinated interpretant representations, and I cannot perceive your representations.

It wouldn't matter whether I say 'dog', '*chien*' or '*Hund*' except that I have some expectations about which is most likely to activate your interpretant representation DOG, given that you've been reading this English text. Conversely, if I've previously observed you speak English, then this provides evidence allowing me to guess what interpretants lie behind particular communicative behaviours of yours. But in the absence of a shared language, as at the symbolic threshold, such information is not available.

One of us could provide alternative non-linguistic evidence in the form of gesture, such as pointing or miming. For instance, if you were unsure just what I meant by 'hot', I might wave my hand in front of my mouth, tongue out, to indicate SPICY as opposed to BOILING. I would be picking out a *salient* feature of my representation HOT to bring to your attention.

But this, while non-linguistic, is a comparatively sophisticated form of communication. The next subsection looks at how Lewis argues that salience originates non-symbolic conventions in the *absence* of any such communication, linguistic or otherwise, before returning to such invented signals in the following subsection.

1.4.1.2 Salience and inference

Lewis's project was to show how conventions might arise without explicit agreement, and his answers focus on notions of salience derived from Schelling (1960), as well as on precedent, a particular kind of salience. Schelling (1960) asked a number of respondents where and when they would meet a friend in New York if they had not previously agreed a time and place: this is a

co-ordination problem with a vast multitude of possible strategies.

The majority of his respondents independently responded that they would meet at Grand Central Station and most of these settled on noon. Of all the possible times and places, Schelling claims that Grand Central and noon are somehow salient, standing out from the rest, and Lewis claims that people would expect them to be salient to others, meaning that they can correlate their expectations about mutual behaviour without explicit prior communication. One way in which an action might be salient is precedent: it was previously the solution to a similar problem.

Sillari (2008) and Cubitt and Sugden (2003) interpret this to mean that salience is what originates conventions, while precedent perpetuates them. Lewis argues that, for salience to play this role, the agents in question must be rational agents, by which he means that they must adhere to standards of *inductive* inference and must assume others adhere to similar standards. However, Skyrms (1996) argues that this merely pushes the problem back a step for Lewis, in that he would have to account for how common knowledge of the salient choice or common knowledge of standards of inductive inference arise without a great deal of prior communication. That is, Skyrms objects that Lewis cannot claim conventions arise without language, or at least something quite like it, since salience or inferential standards must be common and known to be common, and Lewis has not shown how this might occur without extensive communication.

Skyrm's alternative is to provide a non-cognitive account, avoiding induction entirely. In a model of a signalling game involving random strings, he shows that, beginning with a state of random variation in the frequencies with which signals are chosen, if more frequent actions are more likely to be replicated by imitation and learning, then a conventional signalling system is certain to emerge, though the agents in the model do not need salience, common knowledge or common standards of inductive inference[21].

Cubitt and Sugden (2003) acknowledge that Skyrms's model produces conventions without cognitive mechanisms for salience, but argue that this does not mean that real-world problems can be solved without salience. In particular, they note that Skyrms doesn't say much about how the replication processes in his model are to be understood, and point out that

[21] A host of models of the emergence of signalling systems exists (for a review, see Kirby, 2002). I focus on Skyrms because he explicitly addresses Lewis conventions.

imitation of a real behaviour requires recognition of just what pattern of behaviour is to be imitated. They argue that mechanisms for recognising patterns of regularity across behaviours require salience[22]. Anyway, Skyrm's suggestion relates to dynamic processes, but doesn't explain away the role of salience in static or once-off decisions such as Schelling's example above.

The question of human symbol origins thus needs a framework capable of evaluating the relationship between cognitive mechanisms for processing salience and other processes insensitive to salience. I propose such a framework in chs. 2 – 3 to show how we might tell whether salience is involved in a particular task, and test empirically in ch. 7 whether, or to what extent, salience-processing inference is required during the process of conventionalisation.

Two views on salience will add key terms to this framework. Aumann (1987) explicitly interprets the above phrase, 'standards of inductive inference,' in Bayesian terms: probabilistic *hypothesis* evaluation. Postema (2008) characterises such accounts as taking some salient feature (or set of features) as given, directly perceivable, or otherwise cognitively obvious (he calls this 'natural salience'), and working out how likely it is that another agent has settled (or will settle) on a particular naturally salient option, given their behaviour.

On the other hand, Postema argues that in many cases salience is not naturally manifest, and just which features in a given problem are salient is a matter to be inferred. 'Salience reasoning is not only reasoning from salience detected, but also reasoning to that which is salient' (Postema, 2008, 45). For example, an inductive account would evaluate how likely it is that someone might meet at Grand Central Station, given that it is the salient option. Postema's salience reasoning, on the other hand, is interested in working out why or how anyone would come up with the idea that the station is a salient choice in the first place.

He claims that *creativity* and *imagination* are features of salience reasoning, and that it is related to *analogy*. So the framework just mentioned must also involve investigating whether creative inference and analogy are different from induction, and if they are, whether symbol origins require them. Since the inductive approach involves reasoning *from* salience and is called 'hypothesis evaluation', it seems that the set of salient options are the

[22] I explore this aspect of imitation in more detail in §1.4.2.

hypotheses in this framework. This suggests the possibility that the creative alternative, since it involves reasoning *to* salience, is a matter of hypothesis *generation* (though this is not how Postema phrases it). I explore this possibility in ch. 3, though for now I just distinguish inductive or natural salience from salience-deciding inference.

1.4.1.3 Alternatives to salience

But salient co-ordination points are not the only possible source of coordinated equilibrium, though they predominate in Lewis's discussion and are a focus of much of the literature on the subject. Almost as an aside, Lewis also suggests that an agent could just creatively invent a signal which, once understood, can serve as precedent for convention. It is important to examine this possibility in detail, since a number of approaches to symbol origins (such as Garrod et al., 2007) begin with something similar; if my above SPICY example were novel, it would be an example of the sort of thing Lewis means here.

In Lewis's example, someone wants to warn others about a patch of quicksand. They decide to partially submerge a scarecrow in the quicksand in the hope that others seeing it will 'catch on' (1969, 158). 'The idea, presumably, is that a stylised representation of a human figure submerged up to its chest will, by a natural association of ideas, prompt thoughts about real human beings being similarly submerged' (Cubitt and Sugden, 2003, 201). But what is the empirical upshot of Lewis's phrase 'catch on' or Cubitt and Sugden's 'natural association of ideas' which 'prompt thoughts' about real humans? How would we know whether any communicative event counts as an example of this? Is it entirely unrelated to salience, or are there some connections? The quoted phrases are suggestive, but need some unpacking.

A couple of semiotic terms will clarify just what aspects of the communicative process we're talking about. The 'ideas' and 'thoughts' referred to above are interpretant representations. Upon seeing the scarecrow in quicksand, one's interpretant SCARECROW is activated. This eventually leads, via interpretant HUMAN, to the ultimate behavioural interpretant: a disposition to avoid the patch of quicksand. So Cubitt and Sugden's suggestion amounts to claiming there is a 'natural association' between interpretants, not between representamen and object, so it's not iconicity that explains

this example.

There are a number of possible ways of understanding these 'natural associations'. The first is predictability (see §1.3.2.3): it could be that activation of SCARECROW predictably leads to activation of HUMAN. A second is semanticity: part of one's world knowledge about scarecrows includes the fact that they resemble typically humans. The last is structural similarity (Gentner, 1983): the thought that a scarecrow is half submerged may share structural properties with the thought that a human is half submerged. I don't think these are equivalent, but neither are they incompatible. They may even be related[23].

Regardless, these 'natural associations' cannot shoulder the explanatory burden alone. Relationships between potential interpretants vary according to context. If you see a scarecrow in a field, I think it less probable that your representation HUMAN would be activated. Even if it were, it would probably not play an explanatory role in your subsequent behaviour, unlike the disposition to avoid the quicksand. Further, there may be similarly natural associations between SCARECROW and STRAW or OZ or FIELD, but these potential interpretants play no role in avoiding the quicksand, though they may play a role in other contexts.

Even if there are 'natural associations' between all of the above interpretants, naturalness alone cannot explain why the link between SCARECROW and HUMAN plays an explanatory role in this particular case, while other 'natural associations' do not. Just which interpretants are involved in interpretation thus also depends on the particular communicative context. Lewis's phrase 'catch on' focuses attention on the interpreter, particularly on their understanding of the communicator's intent. So I must eventually investigate how 'natural associations' interact with inferences about contextually-situated communicative intention. Just as Postema (2008) distinguished natural salience from salience reasoning, it may turn out that we need to distinguish 'natural associations' from associations that are the product of the interpretive process.

Grounds are another semiotic feature of the quicksand example. The submerged-scarecrow representamen is not intended to signify a huge range

[23] When I get around to talking about accessibility and analogy in ch. 3, their interrelationship will become clearer. For now, I'm just underlining why symbol origins require a detailed cognitive background to convention.

of facts about its object, a submerged human, just as a weather vane cannot signify anything other than the direction of the wind. Grounds pick out or make salient particular features of a representamen and particular features of an object, just as my SPICY gesture made salient certain features of 'hot'. So this purported alternative to salience hasn't really stepped very far away from salience.

If symbols are conventional, then symbol origins needs to investigate possible sources of convention. The approach examined here leads to the conclusion that inferences about communicative intent are needed to interpret novel signs, and that these inferences are salience-deciding and sensitive to semantic information or representational structure. *Relevance Theory* (Sperber and Wilson, 1995) is an area deals with precisely such concerns, for which see §1.5.2.

1.4.1.4 Salience and phylogeny

In the preceding few subsections, then, I have examined two sources of convention, namely salient coordination points and a 'natural association of ideas'. Cubitt and Sugden (2003) conclude that both these ways help Lewis explain how conventions might arise without anyone using language to agree explicitly on a salient solution (i.e. how a linguistic species can arrive at convention without explicit agreement), but they admit this does not provide an answer to the question of how conventions arise without language at all (i.e. how a non-linguistic species develops into one capable of symbolic convention).

Regarding this particular problem, Cubitt and Sugden end with the speculation that 'human beings are born with innate tendencies to privilege certain patterns when making inductive inferences' (2003, 203), rather than having to acquire such tendencies, which seems to tie together salience and induction.

Peirce makes a similar claim:

> Nature is a far vaster and less clearly arranged repertory of facts than a census report; and if men had not come to it with special aptitudes for guessing right, it may well be doubted whether in the ten or twenty thousand years that they may have existed their greatest mind would have attained the amount of knowl-

> edge which is actually possessed by the lowest idiot. (CP 2.753, 1883)[24]

Animals, he suggests, also have an innate tendency to privilege certain patterns (such as con-specific signals, or signals of food or predation) over other background information, whereas humans seem capable of learning to do this in a wider range of situations. These claims cohere with the first of Hurford's two options in §1.3.1.3 above (that our ancestors were able to make a 'good guess' in response to a novel signal).

While the introduction to this section glossed 'arbitrariness' as open-endedness in behaviour, a detailed look at symbolic convention has raised the question of how, given such open-endedness, we manage to co-ordinate on anything at all. I am going to end up leading the discussion away from induction, but for now the central point is that how we evolved across the symbolic threshold is inextricably intertwined with how we evolved to make certain kinds of inference. In particular, two background questions will be, firstly, whether salience-deciding inferences differ in degree or in kind from other types; and secondly, whether human rationality is different in degree or in kind from animal rationality.

1.4.1.5 Conclusions about Lewis conventions

While the previous section on arbitrariness (§1.3.2) identified the open-endedness or non-causal nature of our communicative behaviour as being of interest, this section raised the point that settling on co-ordinated signals in the midst of such open-endedness is itself problematic. Lewis's main attempt to solve this problem assumed natural salience and common standards of inference, whereby the vast uncertainty of a communicative problem is reduced to a few options that are obvious to us and that we expect others to grasp. A subsidiary attempt involved creating a novel signal while 'natural associations' among ideas interpret it. Cubitt and Sugden (2003) and Favereau (2008) argue that these are not enough to account for the origins of language. Rather they can just account for how a species capable of language can arrive at a convention without using that language.

[24] Citations to Peirce are traditionally of the form CP X.Y where CP refers to *Collected Papers of Charles Sanders Peirce* (Peirce, 1935), X to *volume* and Y to *paragraph*.

I argued that salience-deciding inferences about communicative intention must be added to the mix, and that how we evolved to make such inferences is central to the story. I distinguished salience reasoning from natural salience and argued that symbolic convention needs the former; I also showed that 'natural associations' need to be situated in a larger inferential communicative context in order to have any explanatory weight. A number of key points for further exploration were identified, including induction (particularly the Bayesian kind), creativity, imagination, analogy, hypothesis generation and evaluation, connections between interpretant representations, and salience- and context-sensitive processes.

However, it might be that Lewis's account is entirely incorrect, rather than incomplete. We should thus turn to alternative characterisations of convention.

1.4.2 Millikan conventions

Compared to Lewis, Millikan (2005) offers a greatly simplified version of convention. Lewis's co-ordinating conventions are merely a subtype of convention as far as Millikan is concerned. For something to be conventional in Millikan's more general sense, it must firstly consist of patterns of behaviour that are reproduced; and secondly, these patterns must proliferate mainly due to 'weight of precedent, rather than due, for example, to their intrinsically superior capacity to perform certain functions' (2005, 2).

By 'reproduced', Millikan does not mean 'always and in every respect'. She points out that the background colour and texture of a photocopy depends on what paper is put into the tray, rather than the background colour and texture of the original. It is still a reproduction of that original, though not in every respect. Further, a cat might only catch a mouse one in ten pounces, but the cat still reproduces this pouncing behaviour even if it is not always successful. For these and other reasons, Millikan's characterisation is somewhat more biological than Lewis's, which relies on reasoned and consistent regularity.

Millikan's second criterion, precedent, means that something must be reproduced, not due to any functional advantage, but simply because that's how it was done before. For instance, she notes that although she learned from her mother how to open jars by dipping the lids in hot water, this behaviour was reproduced because it is better at opening jars, not merely

because of precedent. It is thus not conventional. So to Lewis's requirement concerning the availability of alternatives, Millikan adds the notion of functionality, something I touched on in the coffee-mug example above. Those innate behaviours due to natural selection are not conventional, since they would have been selected for the biological advantages they provide.

To see how we tell whether something is copied due to 'weight of precedent' or functional advantage, I begin with a discussion of cultural transmission in chimpanzee tool use, raising questions about just what counts as imitation and why imitation is different from emulation (§1.4.2.1). This in turn prompts a discussion of two kinds of imitation (§1.4.2.2). I identify which kind of imitation is necessary for symbols, and argue that this requires complex mechanisms for inference about *relevance* (§1.4.2.3).

1.4.2.1 Cultural transmission: imitation and emulation

There are many examples of animal behaviour that are culturally transmitted (for a review of chimpanzee behaviour, see Whiten et al., 1999) and thus seem to fit Millikan's criteria, but cultural transmission is a broad category and in the relevant literature there is an important distinction between imitation and emulation. It will turn out that Millikan conventions require imitation. In this subsection I sketch out the difference between the two and highlight a complication with the latter.

Typically, emulation involves the copying of a result; imitation involves the copying of behaviour (Tomasello, 1990; Horner and Whiten, 2005). If one animal observes a conspecific opening a box to retrieve food, and then uses its own method to open the box, it is emulating. By observing a conspecific, it learns something about the world: that the box contains food and that it is possible to open the box to retrieve the food. It then discovers for itself how to achieve the same result[25], so its box-opening behaviour might not replicate that of the model.

Imitation, on the other hand, requires the replication of conspecific behaviour: in the above case, the actual motions made (whether causally useful or irrelevant) by the model conspecific in opening the box. Millikan conventions thus require imitation, not just emulation, since convention requires that patterns of behaviour be reproduced for non-functional reasons, but

[25]'Results' thus decomposing into 'affordances' and 'causal relationships' (Horner and Whiten, 2005).

emulation doesn't involve replication of behaviour and does include replication of function. If symbols are conventional in this sense, they need a species capable of imitation.

Though apes are capable of cultural transmission (Whiten et al., 1999; Horner et al., 2006), there is a general tendency to paint humans as imitators and apes as emulators. But Horner and Whiten (2005) and Whiten et al. (2009) argue that this is overly simplistic. They show experimentally that chimpanzees are more likely to emulate when they can perceive the interior (and thus the relevant causal relationships) of a box that can be manipulated in relevant and irrelevant ways to retrieve a reward; they are more likely to imitate when they cannot perceive the interior. Human children, however, imitated relevant and irrelevant actions in both conditions, despite the fact that human understanding of causal relationships is superior to that of apes.

However, Boesch and Tomasello (1998) complicate matters by introducing the idea that imitation also involves replicating the *intention* of the model, in addition to the behaviour. If a human perceives another cleaning a window, they represent the behaviour to themselves as 'cleaning the window' rather than 'moving her hand in a circular motion on the surface of the window while holding a cloth' (1998, 598-599). In that case, the observer still imitates the behaviour if they clean the window, even if they do not replicate every motion of every body part. Boesch and Tomasello thus claim that intention plays a role in deciding which aspects of the model behaviour are *irrelevant* or *arbitrary*.

Though this is an appealing addition, it introduces a complication, which can be seen by decomposing 'copying' into the three features discussed so far: result, intent and action. Copying each of these can be independent of the others. If, while exploring an ancient temple, I see a colleague trip such that they knock over one of a series of vases, smashing it to reveal gold coins inside, I might copy this result by purposefully smashing another vase in the series, though I do not copy their intent (it was an accident after all) or their action (they bump into it, I pick it up and drop it). A second colleague might copy me, swinging their machete at yet another vase. Here they copy the result and intention of my behaviour, but not the action. A third colleague retraces our steps to smash all vases we had previously seen. It turns out that all other vases in the temple are not hollow, so he has copied our intentions (with or without copying actions) but not the results

because we have not yet discovered just what counts as a member of the category 'gold-containing vase'. A child accompanying the party interprets all this as a grand game of smashing stuff and copies the pick-up-and-drop action to make his contribution. Here, he copies action and result, but not intention[26].

If imitation were to involve the yoking together of action and intention, this would muddy the waters somewhat because intentions are related to goals and thus, potentially, to results and thus to emulation. On the other hand, if we keep action, intent and result distinct as suggested by the above examples, this raises the possibility that there is more than one type of imitation (i.e. copying of behaviour alone, or copying of behaviour and intention) and that animals, though capable of simpler forms of imitation and thus of conventional behaviour in Millikan's sense, might thus not be sufficiently like us to grant them symbols.

The window-cleaning example introduced a link between intention and *relevance*. But if there are different kinds of imitation, this means there might be different kinds of relevance. So the next subsection looks at kinds of imitation and the role of inference, while the following one distinguishes different kinds of inference about relevance.

1.4.2.2 Types of imitation

The claims here run parallel to Favereau's criticisms of Lewis (§1.4.1.1): he argued that the addition of mental considerations to our conception of convention means that there must be two kinds of Lewis convention: co-ordination of action alone or co-ordination of action and belief; I argued that symbol origins need the latter kind. Replacing 'co-ordination' with 'copying' or 'replication' and replacing 'belief' with 'intention' yields the same claim, *mutatis mutandis*, for Millikan conventions: there is replication of action alone, or of action and intention. Humans may not need precisely the same intentions when using a symbol, but our intentions must overlap at least partially if we are to communicate successfully.

Zlatev (2008, 2009) distinguishes between five levels in a mimesis hierarchy, five increasingly complex behaviours involving copying of some sort,

[26]Imitation of action, but not results or intent, is called overimitation in the literature (e.g. Lyons et al., 2007; Whiten et al., 2009), though this will not be a central concern in what follows.

where each new layer develops upon lower levels but adds new capacities. The third-highest level is triadic mimesis, which involves imitation of communicative intention. The level above that, protolanguage, is where convention comes in. So Zlatev's mimesis hierarchy implies that understanding communicative intentions is a pre-requisite for proto-linguistic convention.

If we define 'symbol' as requiring conventions in the sense of requiring imitation of action and intention, then we must consider whether animals can read intentions, and if so, what kind. Moll and Tomasello (2007) discuss an interesting example drawn from Hare and Tomasello (2004) contrasting chimpanzee understanding of pointing and reaching. What Hare and Tomasello found was that chimpanzees did not successfully interpret a human experimenter pointing to a container with food in (i.e. they did not search there for food), but chimpanzees did search for food in a container when the human experimenter reached unsuccessfully for it[27]. The interpretation of this by Moll and Tomasello is worth quoting in full:

> [I]n the case of reaching, the chimpanzees just need to perceive the goal-directedness of the human's reaching action and 'infer' that there must be something desirable in the container. This task can thus be solved with some understanding of the individual intentionality of the reaching motion. In contrast, to understand pointing, the subject needs to understand more than the individual goal-directed behaviour. She needs to understand that by pointing towards a location, the other attempts to communicate to her where a desired object is located; that the other tries to inform her about something that is relevant for her. (2007, 644)

The authors conclude that chimpanzees are able to read intention in competitive but not in co-operative situations. This means that chimpanzees are not typically capable of symbolic communication since this requires co-operative intention reading. But notice that Moll and Tomasello only partly frame the issue in terms of co-operation. They also partly frame it in terms of inference: chimpanzees are able to infer some things (the desirability of

[27]The visual difference in finger arrangement between reaching and pointing seems not to be salient to chimpanzees (Leavens and Hopkins, 2005).

whatever was in the container) but not able to infer others (the communicative intention behind pointing)[28].

Presumably the latter is, in some sense, more difficult, so we should unpack what makes some inferences more difficult. The following subsection thus examines a range of inferences about goals and intentions, and shows that symbols involve complex inferences about relevance.

1.4.2.3 Imitation and relevance

Recall that Millikan does not require a perfect copy for convention. Rather,

> [a] pattern has been reproduced if its form is derived from a previous item or items having, *in certain respects* the same form, such that had the model(s) been different in these respects the copy would have differed accordingly. A reproduction is never determined by its model in all respects. (Millikan, 2005, 3, emphasis mine)

In Millikan's photocopy example, the background colour of the copy need not be similar to the original, but rather depends on whatever blank paper was put in the drawer of the machine. Different patterns of ink on the original, however, would have led to different patterns on the copy. Relevance, here, is determined mechanically.

But how is relevance determined cognitively? I will argue that this depends on interactions between types of inference and types of imitation, though explaining this will first require a look at claims in Gergely and Csibra (2003). They describe and offer evidence for a distinction between a teleological stance and a mentalistic stance (fig. 1.7). These are interpretational systems, but the former involves representations of observable things (actions, goals states resulting from those actions, and situational constraints); the latter involves representation of non-observables (intentions, beliefs and desires). Goals states are something like the causal relationships and affordances involved in emulation, so recognising a physically perceivable goal state does not require recognising an intention.

[28] Sometimes the authors use scare quotes around 'infer'; other times not, so it is not clear whether they would be committed to this view. These approaches (competitive/co-operative vs inferential) focus on different dimensions and are thus not incompatible, and the discussion of Lewis conventions provided support for the idea that inference is worth pursuing further in this regard, in addition to the competitive/co-operative distinction.

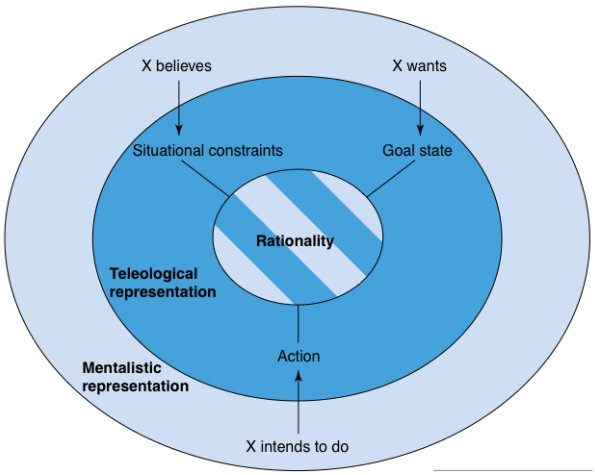


Figure 1.7: Teleological and mentalistic stances (Gergely and Csibra, 2003, 289).

Both stances support inferences: taking a teleological stance, from observing an action one might infer its goal; taking a mentalistic stance, one might infer someone's intention from their desires. The chimpanzees' understanding of reaching in Moll and Tomasello (2007) is an example of a teleological stance: from the reaching behaviour, they can make inferences about the affordances (desirability) of the goal (something in the container). Gergely and Csibra (2003, 2006) review other cases to show that some animals are indeed capable of such inference. They also provide experimental evidence that even one-year-old children can make teleological inferences, though apes' teleological understanding is very limited compared to human infants'.

As humans develop, we eventually become capable of taking a mentalistic stance, which Gergely and Csibra (2003) claim is built on or overlays the earlier teleological stance (hence the concentric circles in fig. 1.7). Humans can make inferences from a teleological stance to a mentalistic stance, while the chimpanzees' failure to understand pointing in Moll and Tomasello (2007) is due to their inability to reach infer the communicative intent motivating the pointing action.

A 'principle of rationality' warrants inferences within each stance (the innermost circle in fig. 1.7), and this underpins how relevance is decided

cognitively. In the teleological case, rationality boils down to efficiency: beings operating under the teleological stance presuppose, when making inferences, that others' actions lead efficiently to goal states, given situational constraints (Gergely and Csibra, 2003).

As an example, Gergely et al. (2002) had preverbal (14 month old) children observe a demonstrator turn on a light-box by bringing her forehead into contact with the switch. In one condition (hands occupied), the demonstrator pretended to be cold and used her hands to keep a blanket wrapped around herself. In the other condition (hands free), the demonstrator's hands remained on the table.

When asked to copy the behaviour, infants could either imitate by using their heads or emulate by using their hands. They were significantly more likely to imitate in the hands-free condition and emulate in the hands-occupied condition. Since the reason the demonstrator used her head was perceivable in the hands-occupied condition, children could *see* that the use of her head was irrelevant in turning on the light, and most were able to rationally emulate the behaviour with their hands. In the hands-free condition, however, 'if infants noticed that the demonstrator declined to use her hands despite the fact that they were free, they may have inferred that the head action must offer some advantage in turning on the light. They therefore used the same action themselves in the same situation' (Gergely et al., 2002, 755).

Gergely and Csibra (2006) claim that relevance is 'cognitively transparent' in the teleological stance, since decisions about efficiency are derivable from observation of physical means-ends relationships in a given situational context. Assuming the demonstrator to be rational is a matter of perceived efficiency, not mental attribution in this case. On the other hand, relevance in the mentalistic stance is 'cognitively opaque', not derivable from these observables without further inferences about intention, desire or belief. In that case, imitation of action plus intention (and thus symbolic convention) requires something more, cognitively, than imitation of action alone.

The final step in this subsection is to show that some inferences about intention are more difficult than others in terms of relevance. Returning to the mimesis hierarchy (Zlatev, 2008, 2009, mentioned in §1.4.2.2 above), the level below triadic mimesis is dyadic mimesis. Triadic mimesis involves understanding communicative intention, 'where the subject *intends* the act

to stand for some action, object or event for an addressee (and for the addressee to recognize this intention)' (Zlatev, 2008, 138); dyadic mimesis is imitation without communicative intention.

Starting with dyadic mimesis, recall the window-cleaning example above (§1.4.2.1). If you observe someone deliberately performing an action you recognise, going from a teleological stance (the goal state of their action is a clean window) to a mentalistic stance (they intend to clean the window) requires a very simple premise schema: 'if someone is performing action x, then they intend to perform action x'. But what counts as relevant imitation is independent of this inference: it won't matter if you move your cloth clockwise while the model wipes anticlockwise as long as you both get the window clean, and you will be able to *see* whether it is getting clean. Relevance (and thus imitation) here is determined by the principle of efficiency, without moving beyond the teleological stance: it is not cognitively opaque here, since it does not depend on a mentalistic inference.

Moving onto triadic mimesis, if I already understand the word 'tree', I know why it can refer to a syntax diagram but not a flower because I know that the relevant feature is the prototypical tree's branching, rather than its being a plant. Knowing a symbolic convention means knowing (to some extent) what counts as relevant in replicating that convention. So learning the convention requires figuring out what is relevant.

But this differs in two ways from the dyadic example. First, while relevance was independent of the mentalistic inference there, it cannot be so here. Secondly, important parts of the mentalistic inference cannot be captured by anything as simplistic as the above premise schema. If you don't know the word, you could still infer that by saying [tri:] they intended to say [tri:], but you will not have discovered anything about the meaning of the word, or about their communicative intention in using it. Working out what is relevant depends on a complex inference relating one non-observable (the communicative intent) with another non-observable (the semantic content) in a particular context.

Csibra (2003) and Gergely and Csibra (2006) state explicitly that this is the sort of pragmatic inference described by Sperber and Wilson (1995) which I have already mentioned in §1.4.1.3, and which I will look at in the next section.

1.4.2.4 Conclusions about Millikan Conventions

I argued that, if symbols are conventions in Millikan's sense, they require imitation, not emulation. I distinguished imitation of action alone from imitation of action and intention, and argued that the latter, in cases of triadic mimesis, requires inferences about relevance as part-and-parcel of inferring speaker intention and word meaning, both of which require a mentalistic, rather than teleological stance. Since relevance is cognitively opaque in the mentalistic stance, we need to shift our attention from properties of the sign (such as whether it is replicated by precedence) to properties of inference. While the distinction between co-operative and competitive behaviour has been discussed in language evolution, I have argued that an as-yet-unexplored dimension worth considering is what kinds of inference are involved in various kinds of interpretation (teleological vs mentalistic; dyadic vs triadic; known vs novel; cognitively transparent relevance vs cognitively opaque relevance).

1.4.3 Summary on symbols as conventional signs

Since symbols are comparatively unmotivated, open-ended behaviour, the previous two sections have examined the question of whether 'conventional' is a good way to describe how that open-endedness is constrained. Lewis and Millikan offer two approaches to this based on rational co-ordination and imitation, respectively. In both cases, I showed that while these may be accurate descriptions, neither can account for symbol origins without sophisticated forms of inference about salience or relevance, two concepts I will eventually show to be related. The two consequences of this are that we must look at just how symbols should be defined inferentially (the next section) and that we need to distinguish complex, highly evolved forms of inference from more basic forms (the next chapter).

1.5 Symbols are inferential

Compared to the more common definitions of 'symbol' above, what follows is rather a minority approach, but some sources do focus on inference as opposed to arbitrariness or convention. These sources do not always refer explicitly to inference, but may phrase their claims in terms of 'judgement' or

'interpretation', which I take to be terms inextricably bound up in questions of inference. Firth, for example, claims that

> in the interpretation of a symbol the conditions of its presentation are such that the interpreter ordinarily has much scope for exercising his own judgement ... hence one way of distinguishing broadly between signal and symbol may be to class as symbols those presentations where there is much greater lack of fit — even perhaps intentionally — in the attribution of the fabricator and interpreter. (1975, 66)

Similary, Deacon (1997) suggests that what distinguishes our understanding of reference from, say, a dog's, is something additional produced in our heads when we interpret an utterance. Deacon explicitly identifies this 'something additional' as a Peircean interpretant, which he glosses as 'whatever enables one to infer the reference from some sign or signs and their context' (1997, 63). Finally, Eco (1978, 1986) stresses that any semiotic phenomenon (not only symbols, but signs more generally) is characterised by the necessity for inference in its interpretation and Ramscar et al. (2010) argue, based on computational models and experimental data that symbol learning is inherently inferential.

In the following subsections, I will first outline some cognitive/theoretical background that discusses the role of inferential interpretation in symbolhood (§1.5.1). Then, a look at an account rooted in pragmatics will support this shift in focus from the nature of the sign to the nature of interpretation (§1.5.2). Finally, I will tie together a number of threads discussed throughout this chapter to provide a definition of 'symbol' (§1.5.3).

1.5.1 Cognitive/theoretical background

Throughout the following, I will keep exposition of technical terms quite brief so as not to hold up the discussion, but key ideas introduced here will be discussed in more detail in coming chapters.

1.5.1.1 Fitch on *The Evolution of Language*

Fitch (2010) contrasts two broad approaches to meaning: the realist and the cognitive. These, he says, focus on different aspects of the semiotic triangle

(fig. 1.1 above). Fitch claims that it is not at all unusual in philosophy and formal semantics to treat word and sentence reference as a real, direct mapping between linguistic units (semiotic representamens) and the world (semiotic objects). Counter to this, the cognitive model allows no direct relationship between such signals and the world. Rather, this relationship is indirect in that it is mediated by the mind (semiotic interpretants). Fitch argues that data from cognitive science (human and animal) leave only the latter model a feasible option and my outline of the Peircean sign (§1.2.3.1) make the same point.

The previous sections about arbitrariness and convention can thus be summarised as arguing that these, as descriptions of an *indirect* relationship, are rather superficial and fail to capture with much clarity what a symbol is or how we might investigate symbol origins. And since these indirect relationships are mediated by the mind, I have been aiming at a refocus on cognition.

Fitch goes on to say it is pre-linguistic concepts that allow for meaning. Hurford (2007) argues for the existence of pre-linguistic concepts in a wide range of animals, and states explicitly that these are evolutionary preconditions for language. For both Fitch and Hurford, the problem of the earliest origins of meaning is one that has been partially solved by evolution: our ancestors and various relations evolved to have pre-linguistic concepts of certain kinds, and evolved cognitive mechanisms that allow such concepts to connect signals on one hand with reality on the other. Two pertinent questions, then, are what kinds of cognitive mechanisms these were, and what kinds of cognitive mechanisms are unique to human language.

Fitch lists biases or constraints on inference; biases or constraints on concept formation; ToM; higher-order intentionality[29]; and cooperation in information sharing. Of these, he says the last three are uniquely human, while many animals are capable of concept formation and inference, to a certain extent. I will shortly make a few comments on his view of inference, but first some illustration might help.

He contrasts what we know of child language learning with Quine's *gavagai* problem (Quine, 1960). An anthropologist observes a tribesman speaking an unknown language say '*gavagai*' as a rabbit runs past. An indeterminate multitude of hypotheses are possible concerning the meaning of

[29] Here in the sense of 'aboutness', rather than 'volition'.

'*gavagai*': it is likely to mean 'rabbit', but could also mean 'food', 'Let's go hunting', or a host of odder things: 'There will be a storm tonight', 'A momentary rabbit stage' or 'An undetached rabbit part'. But despite these *theoretical* possibilities, children in fact tend to generate a much more tightly constrained set of hypotheses:

> The child obviously does not unconsciously process all of Quine's various logical possibilities. Rather, the hypothesis space appears *constrained* in certain ways, and the child simply fails to consider many of these possible meanings. (Fitch, 2010, 126)

These constraints are often, in cognitive or linguistic sciences, framed as a problem of inductive inference (Xu and Tenenbaum, 2007), where talk is common of inductive biases: (typically innate) cognitive mechanisms that constrain hypothesis generation. For instance, studies show that children have a whole-object bias (Markman, 1991): they tend to assume novel labels refer to whole objects like rabbits rather than parts or properties like fluffiness or rabbit limbs. Another example is the basic-level bias (Rosch et al., 1976; Markman, 1989): people are more likely to assume a word refers to an intermediate level in a taxonomic system (like rabbit) rather than superordinate (mammal) or subordinate (Angora) categories.

Fitch draws a connection between such biases and evolutionary salience:

> The child doesn't induce such [Quinean] wacky concepts, for the same reason that a dog does not conceptualize "rabbit" in these ways, but rather as a medium-sizeed, fleet-footed potential prey item. This constraint thus may reflect what the child finds salient in the environment (namely whole objects) and conceives of, prelinguistically as a THING. (2010, 127)

So Fitch claims that inductive biases are what constrain the hypothesis space of possible meanings of a new symbol, or make certain meanings salient, but one problem is that these inductive biases are rather context-dependant. In a novel word-learning task, Christie and Gentner (2010) presented two groups of children with two sets of visual stimuli (e.g. standard 1 and standard 2 in fig. 1.8). In one condition, the children were presented with standard 1 and standard 2 one after the other (consecutive condition); in the other condition, both standards were presented simultaneously (concurrent

condition). Both standards were labelled with a novel word embedded in a syntactic context suggesting it was a noun. The children were then asked to choose between two possible matches for the same label: the relational match showed previously unseen animals in the same spatial arrangement as the standards; the object match showed both previously seen animals without the spatial relationship seen in the standards. In the consecutive condition, children were more likely to chose the object match; in the concurrent condition, the relational match.

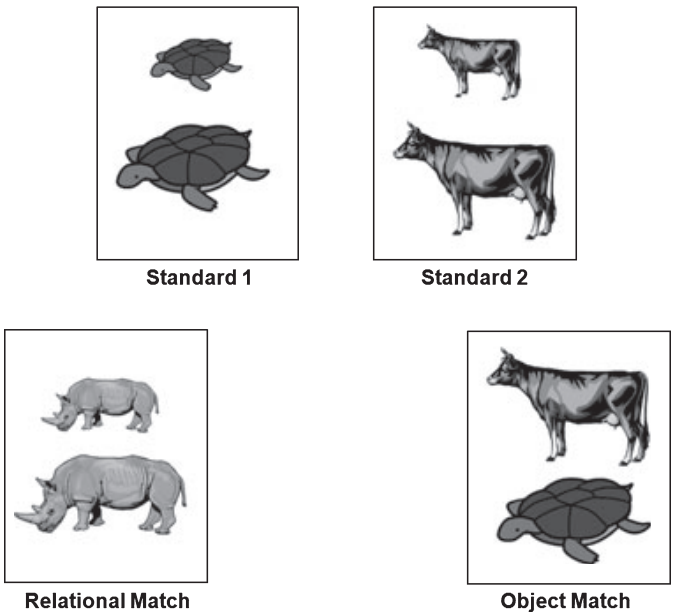


Figure 1.8: Stimuli from Christie and Gentner (2010).

Although this particular experimental design precludes a non-ambiguous object match[30], presumably those who chose the object match thought the label had something to do with either of the objects, while presumably those who chose the relational match thought it has something to do with the relationship between objects. That is, the object bias dominated in the

[30]That is, in the case of the object match, the label's meaning is presumably disjunctive: 'cow or turtle' rather than just 'cow' or just 'turtle'. Or possibly those who chose the object match despite seeing both standards with the same label just ignored one of them. Repeating the experiment with participants old enough to provide such feedback is the only way to know for sure. The participants in Christie and Gentner (2010) ranged from 3;6 to 5;1 in age, and Fitch (2010) warns against treating children's guesses about meaning like adults' guesses.

consecutive condition, but not the concurrent condition. The object bias may thus operate *within* a particular communicative context, but there exist cognitive mechanisms (described by Christie and Gentner as analogy, or relational insight) that evaluate the relationships *between* contexts or decide whether a bias applies to a particular case. In other words, there exist not only context-based inductive inferences of the type described by Fitch, but also context-deciding inferences that may be of a higher order (inferences about inferential context), or more complex, or open-ended, or even of a different nature (requiring analogy or insight). Christie and Gentner explicitly link these terms to hypothesis generation.

It might be possible to avoid this problem by suggesting an increase in the number of biases, such that one bias operates in Christie and Genter's consecutive condition and another in their concurrent condition. But if we have to suppose either an indefinite number of biases or indeterminacy in the application of biases, then we are scarcely in a better position than we would be with Quine's indefinite number of hypotheses. Rather, it makes more sense to allow a limited set of biases, but to suppose their application may depend in turn on more open-ended, context-sensitive inference, or even different kinds of inference, for whose existence we must then search for independent, empirical proof. I provide evidence for the existence of such inference in this dissertation.

Thus far, then, I have been agreeing with Fitch by extending his criticism of the realist model to apply it to a definition of the word 'symbol'. But he simply says that animals, like us, can perform inference and that the distinguishing features of human cognition must thus lie elsewhere (ToM, cooperative communication, higher-order intentionality). I am diverging from his account by suggesting that we need to look at different kinds of, or levels of complexity in, inference as well.

Fitch frames his discussion of two of the above mechanisms (ToM, cooperative communication) in terms of the third (higher-order intentionality) by positing an intentional hierarchy. Objects, plants and simple animals lack intentionality altogether: they do not represent the external world in any way. Some animals, especially vertebrates, exhibit first-order intentionality: they have representations of things (such as prey or predator), and may have beliefs, desires and goals[31]. This, Fitch says, allows us to think that dogs

[31]There are philosophers who disagree with such a claim (e.g. Davidson, 1982), but

have minds, but not a ToM. For that, he says, one requires second-order intentionality: representations of representations[32]. Second-order intentionality is also a requirement for cooperative communication. Finally, pragmatic interpretation requires (at least) third-order intentionality (Sperber and Wilson, 1995).

Fitch's take-home message is that 'it is a deep mistake to treat "*theory of mind*" as a monolithic whole that you either have or don't... A "divide-and-conquer" strategy must be adopted towards the intentional stance and ToM' (2010, 136). By 'divide-and-conquer strategy', he means the above intentional hierarchy: a series of increasingly complex cognitive mechanisms, each building on a lower one, such that the lowest level is evolutionarily the oldest and thus shared with the widest range of species.

I have argued that co-operative information sharing and a ToM are insufficient for dealing with the full range of relevance-deciding inferences in imitation (§1.4.2.2–1.4.2.3). So while Fitch is correct in ascribing inference to animals and in highlighting human uniqueness in terms of higher-order intentionality, I am merely extending his warning against 'monolithic wholes' to inference: there are different levels of inferential complexity that separate different animals, or that separate animals from us, and I will explore this hierarchy in the next chapter. I have also extended his argument against a realist model of the semiotic triangle to argue against 'arbitrary signs' and 'conventional signs' as definitions of symbols, since they focus on indirect signal-world relationships.

1.5.1.2 Tomasello on *The Cultural Origins of Human Cognition*

The discussion of word learning in Tomasello (1999) shows some similarities to Fitch, but focuses more particularly on social aspects of cognition such as ToM and joint attentionality, and is unimpressed by inductive biases. However, a number of terms he uses when talking about word learning are informative.

Our primate relatives are capable of sophisticated cognition, as Fitch admits. However, Tomasello considers humans to be capable of certain kinds of inferences that animals are not. While animals may be capable of

defending Fitch's position would be too tangential just now.

[32]Crockford et al. (2012) argue that chimpanzees have some degree of representation of conspecific knowledge. Once more, it would be too tangential to get into this now.

associating a physical consequent with a physical antecedent, humans can go a step further by making inferences about 'intermediate and often hidden "forces," the underlying [physical] causes and intentional/mental states that are so important to human thinking,' (1999, 19) where these states include things like the intentions of a conspecific.

Tomasello's point, I think, is that animals can infer *that* something might happen, while humans can infer *why* it happens, and the explanatory 'why' is usually cognitively opaque. Though not explicit about whether this is a qualitatively different kind of inference or merely quantitatively different, Tomasello describes it as going beyond purely physical imitation in 'some *creative leap*' (1999, 52, emphasis mine).

Fitch had described the difference between animals and humans in terms of higher-order intentionality; Tomasello agrees that humans are capable of higher-order intentionality and that this plays a role here, but also hints at this difference in inferential ability, which is what I was arguing for above. At any rate, from claims in both Tomasello and Fitch, we can derive the conclusion that symbols operate inferentially. My above call for an inferential hierarchy must eventually also include a look at just what Tomasello might mean by this 'creative leap'.

Tomasello is also sceptical about the role of inductive biases. Like Fitch, he believes that we have ways of avoiding Quine's problem by narrowing down the hypothesis space, but rather than innate biases, he looks instead to replicable joint attentional scenes. A scene is a routine social interaction that might be recognisable to a participant, and that foregrounds certain aspects of the environment for participants in the scene.

He gives the example of an American in an Hungarian train station buying a ticket. The American understands about buying train tickets but doesn't know Hungarian, so upon hearing a novel word while the ticket seller reaches for his cash or offers change, the American makes an inference of the following type: 'if that unknown expression meant X, then it would be *relevant* to the ticketseller's goal in this joint attentional scene' (Tomasello, 1999, 99).

So this involves the generation of an explanatory hypothesis about speaker intention. While limitless Quinean hypotheses are possible, the traveller is likely to generate only that subset relevant to the activity of buying train tickets that he and the Hungarian are jointly attending to. This is more

flexible than an approach based on inductive biases because, once again, 'relevance' is understood in terms of Sperber and Wilson (1995).

This section has concluded, yet again, that we must shift our focus to inference. In particular, it offered initial reasons for thinking that induction is an incomplete account of what goes on in learning new symbols and for considering the relationship between induction, explanatory hypothesis generation, insight, analogy and Tomasello's 'creative leap'.

1.5.2 Pragmatic background

The following outlines an account of pragmatics in Sperber and Wilson (1995), highlighting, among other things, their distinction between context-deciding and context-bound inference (§1.5.2.1). They focus almost exclusively on deductive inference, but admit the existence of more nebulous forms of non-demonstrative (i.e. non-deductive or ampliative) inference. I consider a couple of criticisms suggesting that pragmatic inference is not as deductive as they suppose (§1.5.2.2). I then argue that the question of symbol origins is inherently pragmatic, but that the non-deductive aspects of interpretation come to the fore at the symbolic threshold (§1.5.2.3).

1.5.2.1 Sperber and Wilson on *Relevance*

A core text in modern pragmatics is Sperber and Wilson (1995) on Relevance Theory. The authors contrast an inferential theory of pragmatics with an older code theory[33]. On the code-based account, communicative units are coded and decoded according to pre-existing correspondences: if you hear signal `a`, then interpret it to mean A. The code theory, then, rests rather heavily on conventionalised (or otherwise law-like) correspondences and expects close matches between speaker meaning and hearer understanding. Much animal communication is coded (Sperber and Wilson, 2002), though this is more in the sense of 'law-like' (cf. §1.3.2) than 'conventional'.

In contrast, the inferential theory claims that communication is 'achieved by producing and interpreting evidence' (Sperber and Wilson, 1995, 2): a speaker produces evidence for his intended meaning in a particular con-

[33]The authors claim that a code theory is typical of a semiotic account, which unfortunately shows very narrow knowledge of the semiotic literature. I think Peirce himself would have agreed with many of their basic assumptions, though he would have been more interested than they are in exploring categories of non-demonstrative inference.

text and an interpreter formulates and evaluates hypotheses about what the speaker's intended meaning was, based both on conventional codes and on more open-ended inferential principles that process world knowledge, lexical information and environmental information in the interlocutors' surroundings, finding interpretations relevant to their representation of the world and the discourse. Both inference and codes are required for language, though relevance-based theories such as this construe pragmatics as predominantly inferential. Unlike the code model, there is lots of room for error: many hypotheses are possible, and the hearer may, despite their best efforts, generate one quite different from the speaker's intention.

Sperber and Wilson argue that an utterance or any ostensive (broadly, explicitly communicative) act carries a presumption of its own relevance: hearers assume that things addressed to them are of benefit to them. The benefit of relevance is the positive cognitive effect that accompanies hearers' process of interpretation: they may be able to infer new conclusions that were inaccessible to them before, or they may become more certain or less certain about beliefs they have. They phrase this change in certainty as a change in the *strength* of the beliefs resulting from the process of interpretation. Counteracting this payoff is a processing cost: longer sentences; utterances which require a great deal more inference to understand; or less accessible assumptions require more cognitive effort.

A central claim is that humans are geared towards maximising relevance, achieving the highest cognitive benefit for the least effort. This means that we will not generate all possible hypotheses about an utterance's meaning. Rather, we will unconsciously generate the relevant hypothesis that is the most accessible. Accessibility (or salience, as they sometimes call it) is essentially ease of retrieval from memory.

A set of examples (Sperber and Wilson, 1995, 133-135) will help illustrate these points. In each case, take it as assumed that Peter is implying that he's too tired to cook, and that he thus wishes Mary to make the meal, and that Mary is aware of this.

(1.1) Peter: I'm tired.

Mary: I'll make the meal.

(1.2) Peter: I'm tired.

Mary: The dessert is ready. I'll make the main.

(1.3) Peter: I'm tired.

Mary: The dessert is ready. I'll make an osso-bucco [sic.]

(1.4) Peter: I'm tired.

Mary: The dessert is ready. I'll make the speciality of the Capri restaurant.

In (1), the relevance of Mary's utterance to Peter's implication that he'd like her to cook the meal is obvious. In the other cases, though, he would have to go further, retrieving information related to her utterances from his store of encyclopaedic memory. These are 'assumptions'. But which assumptions must he retrieve? How broad is the context? Sperber and Wilson (1995) consider (but eventually reject) the possibility that there is some way of specifying the context prior to interpretation. The most basic way of doing this would be to specify the context as including implications of previous utterances, and this is what Mary does when she understands that he'd like her to make the meal.

In (2), however, Peter would have to go beyond such implications for Mary's response to be relevant. In order to achieve the same cognitive benefit as in (1), Peter would have to access encyclopaedic information associated with representation MEAL, such as the fact that it may consist of a dessert and a main course. Once he has retrieved this information, it is clear that she is offering to make the meal, so he understands the relevance of her utterance. This has involved slightly higher cognitive cost than (1), though. So far, then, we might attempt to specify the relevant context as consisting of the set of implications of utterances and the assumptions associated with concepts those utterances and implications.

In (3) and (4), however, we have to go still further. If the speciality of the Capri restaurant is osso buco and if osso buco is a main course, then for Mary's responses to be relevant, Peter must access assumptions related to assumptions related to assumptions related to concepts in Mary's utterance. This could extend indefinitely, so if the context is pre-defined, it would have to be one's entire encyclopaedic memory. This, they say, is absurdly unrealistic: we cannot possibly search our entire memory in the short time it usually takes us to understand each other. Rather, they argue, the context is decided as part of the process of interpretation, and

accessibility (or salience, or ease of retrieval from memory) is a key feature of this process, which they say is deductive, on the whole.

Let's look in more detail at how the context is limited in (4). Peter's unconscious interpretive process retrieves assumptions about the Capri restaurant from encyclopaedic memory, but rather than retrieving all the facts he knows about it, the interpretive device retrieves only the most accessible or salient. Just which is most accessible depends on the structure of Peter's memory, so let's assume for now that it is the assumption that their speciality is osso buco. The information that their friend John lives next to the restaurant, being less accessible, will not be retrieved and it would thus not be available to the deductive process. At each such step, previous assumptions can be strengthened or new assumptions can be derived. Each such strengthened or derived assumption is maintained in the memory of the interpretive module, while others are deleted from that particular memory store.

Failing at this stage to find the relevance of Mary's utterance to the concept MEAL in his implication, the process continues, retrieving information associated with OSSO BUCO. Let's say that the most accessible piece of information is that osso buco is made from veal, the second that it comes from Milan and the third most accessible piece of information is that it is a main course. The interpretive process does not access all of these simultaneously. Rather, it retrieves them one-by-one in order of accessibility. The information retrieved first (that it is made from veal) fails to achieve relevance (and presumably information associated with the concept VEAL fails similarly), so this assumption is abandoned and the next examined. Information about MILAN similarly fails, so the process returns to OSSO BUCO and retrieves the next accessible assumption, that it is a main course, which allows further retrieval of the fact that a main course is part of a meal, and thus relevance is achieved: Peter understands that Mary will make a meal, which is what he implied he wanted (fig. 1.9).

Since this depends on the structure of an individual's memories, if Peter's encyclopaedic information about veal included the accessible fact that it is often used to make a main course, then the interpretive process would have played out differently. Alternatively, the information that osso buco is a main course could have been the most accessible piece of information rather than the third, in which case he would have reached the same conclu-

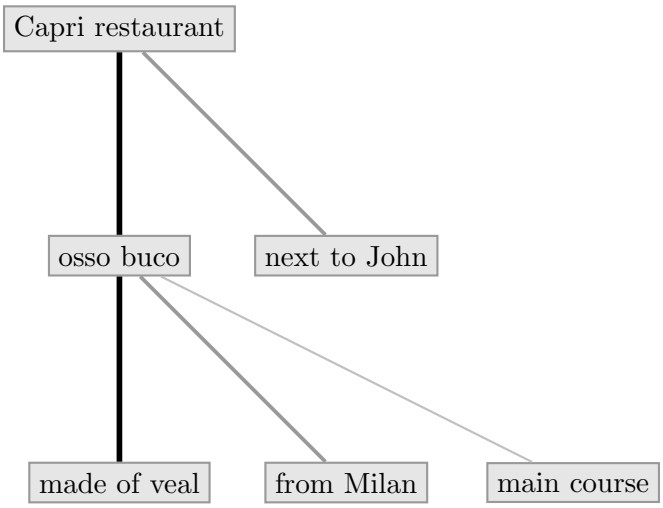


Figure 1.9: The structure of Peter's representations. Darker, thicker lines indicate higher levels of accessibility.

sion with less effort. At any rate, because people have different representational structures, such interpretive processes are not guaranteed to succeed in producing an interpretation that matches what the speaker intended. The authors thus speak of the process being successful or efficient rather than deductively valid. Because of its reliance on accessibility which varies from individual to individual, the process is non-demonstrative on the whole, even though it operates on each step deductively.

So the three main features of all this, then, are that it involves mentalistic inferences about speaker communicative intention, that it is context-deciding rather than context-dependent, and that much depends on accessibility (the ease of retrieval or strength of associations between representations).

I aim to show that the cognitive mechanisms underlying pragmatic inference cannot be as deductive as Sperber and Wilson claim, especially when it comes to the symbolic threshold. This leaves us with only their mention of some poorly understood non-deductive inference to do most of the work. Over the next two chapters I'll unpack what I think this would be like, but for now, a couple of criticisms will give shape to that discussion.

1.5.2.2 Criticisms

The issue here concerns the cognitive mechanisms that are supposed to implement this interpretive process: I argue that the proposed deductive mechanism fails to shoulder the explanatory burden they require from it and conclude that some other type of inference must be responsible. In the next subsection, I look at why the symbolic threshold is especially unlikely to be deductive.

While Sperber and Wilson allow the process to be open-ended, it is open-ended in a stronger or more problematic way than they might be willing to accept. Assume that for Peter, a piece of information that would go on to achieve relevance (that osso buco is a main course) just happened to be less accessible than another, irrelevant piece of information (that osso buco comes from Milan). Since pragmatic interpretation is supposed to proceed stepwise in order of accessibility, MILAN will be retrieved first. Since it fails to achieve relevance, assumptions associated with MILAN will then be retrieved, such as its being in Northern Italy, its being a fashion capital, or its being home to La Scala. But how does the process decide to abandon further pursuit of any such thread of encyclopaedic information? How does it decide when to abandon any subroutine or subsubroutine or subsubsubroutine originating with MILAN and return to OSSO BUCO to retrieve the assumption (its being a main course) that turned out to be pragmatically relevant?

Sperber and Wilson's proposed control mechanism is this: there is lots of stuff in the environment competing for one's attention (and thus for one's cognitive resources) and because we, as a pragmatic species, try get the most cognitive benefit for the least cognitive effort, we would eventually cease to attend to something that costs too much effort. So an interpretive process that drags out in the above way would simply loose out in this competition. Note, however, that this control mechanism determines the fate of the interpretive process *as a whole*: one would simply stop trying to interpret Mary's utterance if led down the garden path in this manner and if something else in the environment seemed to warrant our attention more. Possibly one would just ask her what she meant. This cannot, however, control any subroutines *within* the interpretive process: it cannot account for why an accessible but ultimately irrelevant assumption is eventually abandoned to return to an earlier step with a less accessible but ultimately more relevant

assumption. Fodor (1987) calls this Hamlet's problem: the problem of when to stop thinking. In particular, it is the deductive part of interpretation that runs into Hamlet's problem.

If there are two parallel processes, one exploring MILAN while the other explores OSSO BUCCO, then the latter will achieve relevance sooner or later and the former can terminate whenever that happens. But if we allow two parallel searches, there's no reason there shouldn't be multiple such processes. These could begin with concepts contained in the utterance being interpreted, and earlier activations could influence the course of later activations via some priming effect. That is, rather than proceeding in a deductive linear fashion, it is at least plausible (and, more importantly, testable) that such pragmatic problems are explained by activation spreading from multiple points, depending on the strength of association between representations.

I will show empirically in the next chapter that this is in fact what happens. For now, though, it seems that their blind deductive process founders somewhat against the backdrop of open-ended cognition and if it is this open-ended, non-linear process that explains how a relevant interpretation is eventually found, then deduction cannot always play a dominant role. I am willing to accept that the balance between the two may shift, and that some cases are comparatively deductive, but it will turn out that the above criticisms are particularly problematic at the symbolic threshold. It will also turn out that key terms we've already encountered, such as insight and creative leaps, are often used to describe these cognitive processes.

1.5.2.3 Pragmatics and the symbolic threshold

Sperber and Wilson claim that this pragmatic process interprets not only modern human language, but also non-linguistic signs like novel gesture. Their example of a novel gesture is my holding up a full glass of wine to indicate to you that you needn't open another bottle on my account. They don't actually unpack how deductive interpretation of this would proceed, but I will argue that this wine-glass example is significantly different from symbolic language, though similar to pre-symbolic communication. This still requires pragmatic inference, but I will argue that such cases cannot be deductive.

Consider the *gavagai* example again, and compare it both to a situation where the tribesman in fact says 'rabbit', and also to one where he holds

up a rabbit in the anthropologist's line of sight without saying anything. The latter is like the wine-glass example, while I take the *gavagai* example to be a proxy for the symbolic threshold (their being modern humans notwithstanding) since the participants do not share a code.

In the 'rabbit' case, a rough distinction can be drawn between semantic and pragmatic aspects of interpretation. On the semantic side, hearing 'rabbit' predictably activates the anthropologist's general RABBIT concept. But thinking about rabbits in general cannot be the end of the process. On the pragmatic side, the anthopologist must still infer the tribesman's communicative intent in drawing his attention to this rabbit on this occasion. Pragmatic inference will (probably) settle on a relevant interpretation. It is relevant if it provides some cognitive benefit for the anthropologist, and it is pragmatic in that it produces relevance through a context-deciding inference about speaker intention based in part on the accessibility of assumptions that are associated, however distantly, with the concept RABBIT. This is the part that Sperber and Wilson claim is principally deductive in modern language.

While in the 'rabbit' case the anthropologist's RABBIT representation was predictably activated by a conventional sign, in the silent case it *might* be activated by joint attentional mechanisms, but these cannot possibly be as effective as a conventional sign when it comes to reliably bringing *specific*, *intended* interpretants to mind in one's interlocutors. One way of unpacking this is to exlore how the two cases differ greatly in terms of semiotic ground (§1.2.3.2): the ground is given in the symbolic case, but must be inferred in the silent case.

That is, 'rabbit' will predictably activate RABBIT. This may in turn lead to activation of related representations, but it is rabbitness that is foregrounded compared to these other representations, and knowing the convention means knowing that this is the case. On the other hand, if the tribesman had said 'fluffy', then the fluffiness of the animal would have been made salient for the anthropologist since 'fluffy' would predictably activate FLUFFY.

In the silent case, attentional focus on the animal may well lead to activation of RABBIT. But it might (in addition or instead) lead to activation of FLOPSY, ANGORA, LAGOMORPH, PET or FLUFFY. RABBIT is not foregrounded to the same extent as in the symbolic case since one or more of these other representations could have been activated prior to RABBIT, or

might have been as strongly activated as RABBIT.

Activation of RABBIT might be a statistically more likely outcome than the others and it might thus be argued that the two cases are not very different. But that argument cannot work in general because the intended meaning is not always the whole-object basic-level label (see §1.5.1.1 for similar comments on inductive biases). If RABBIT is more likely to be activated, this can't explain explain cases where the intended meaning is a perceptually less salient feature such as fluffiness, or a subordinate or superordinate category such as ANGORA or MAMMAL: the ostensive gesture doesn't allow as fine-grained direction of attention or manipulation of salience as a symbolic ground does, so this is an example of salience reasoning, rather than natural salience (§1.4.1.2). The ground must be inferred from the context, but the context on this occasion includes the communicative intent of the tribesman, which itself is something requiring pragmatic inference.

So while activation of the appropriate interpretant was the semantic *starting point* for pragmatic inference in the 'rabbit' case, it is *part of or the result of* pragmatic inference in the silent case. In the former case, a coded signal provides access to a particular address in the anthropologist's representational network, from which point pragmatic inference can proceed; in the latter case, working out the salient aspect of the animal cannot be disentangled from working out the tribesman's communicative intention in holding it up. There are thus some respects in which interpreting novel gesture is unlike interpreting language.

I now turn to argue that one such respect is the extent to which the relevance cognitive mechanisms can be deductive: Sperber and Wilson focus on deductive inference in modern language, but this cannot work for cases completely lacking linguistic scaffolding such as novel gesture or the '*gavagai*' case. Like the former, working out a novel ground in the latter requires an inference about the speaker's communicative intention. Consider the following framing inference, though I am not claiming that such premises are explicitly represented in such propositional form, or that the interpreter is aware of any of this.

(1.5) The tribesman has said '*gavagai*'.
I assume he has a rational/relevant[34] reason for doing so.
If '*gavagai*' means RABBIT, then his behaviour would be rational/relevant.
Possibly, then, '*gavagai*' means RABBIT.

I will focus on the first clause of the third premise, which is an hypothesis about the meaning of the word. It is the generation of this hypothesis that, though part of a pragmatic inference about speaker meaning, is non-deductive.

I talked through an example of how Peter reasoned from Mary's offer to make the speciality of the Capri restaurant to the conclusion that she was offering to make a meal. Each step was based on a proposition he believed to be true: the speciality of the Capri restaurant is osso buco; osso buco is a main course; a main course is part of a meal. So each step allowed for deductive inference: if Mary offers to make the speciality of the Capri restaurant, then she necessarily is offering to make part of a meal.

However, a belief of his might turn out to be untrue, or an unintended assumption could nonetheless have been more accessible and relevant, so he could still be wrong on the whole. This, I think, is what Sperber and Wilson mean when they say that interpretation is non-deductive overall, but the interpretive module operates deductively. But even if Sperber and Wilson are correct about each step of Peter's inference being deductive (though I questioned the cognitive mechanisms underlying this in §1.5.2.2), the same cannot be true for the *gavagai* case since one clause contains an hypothesis, not a known fact.

To explore why this step is non-deductive, recall that the hypothesis above is similar to Tomasello's schema in the Hungarian train station (§1.5.1.2). Tomasello claimed humans are more able than any other species to take a creative or insightful leap to infer unobservable or hidden forces (whether physical causes or mentalistic intentions). In (1.5), the hypothesis about the meaning of the word is just the sort of creative leap Tomasello describes because it involves hypothesising an unobservable explanation for observable behaviour and he calls such things creative because the information contained in that hypothesis is not deductively derivable from any

[34] I've included both 'rational' and 'relevant' since this harks back to the discussion about rational imitation (§1.4.2.3).

observables. However, *once* someone comes to believe it, then assumptions can be deductively derived from it as per the osso buco example.

Sperber and Wilson consider creative inference but ultimately reject the notion that this plays an important role in pragmatic inference. They say that such inference is more typical of scientific discovery, and that this is too sporadic and slow a process to explain our quick, common-place pragmatic inferences. But if one posits a continuum between quick inferences that pretty much everyone is capable of as in (1.5) and large-scale scientific insights that are available only to a few geniuses and that involve a total reworking of our understanding of the universe, then Sperber and Wilson's claim could be rephrased as involving a difference of degree, not of kind. That is, hypothesis generation in (1.5) could still involve a (small, easy) creative leap and thus be non-deductive. I take this to be an empirical matter, and will provide experimental support for this claim in part II.

Sperber and Wilson try to avoid the need for these inferential creative leaps by claiming that, while the universe is full of complex facts and sifting through them to form an hypothesis is a complex task, in communication our interlocutors simplify the problem for us by trying to be relevant, scaffolding the task for us. I accept that this could achieve quite a lot in the case of *symbolic communication*, but it cannot do so for crossing the symbolic threshold. In a pre-symbolic species, conventional forms of scaffolding are unavailable because they are conventional. Indeed, Tomasello (1999) defines 'symbol' to be a conventional way of directing attention. The alternative is using novel gesture to direct attention but, as I argued above, novel gesture is different because it is less able to make the relevant aspect of the semiotic object salient to the observer: the ground must be inferred pragmatically. Again, I will show empirically in ch. 6 that manipulating the amount of scaffolding (or how informative the context is) affects the extent to which a pragmatic inference involves a creative or insightful leap.

1.5.2.4 Conclusions about pragmatic inference

I outlined Sperber and Wilson's theory of pragmatic inference because symbol origins are really a pragmatic (as opposed to purely semantic) problem. I highlighted how pragmatic inference is context-deciding and relevance-deciding unlike situations where context is constrained prior to inference or where relevance is determined teleologically rather than mentalistically. The

need for this was prefigured in sections on convention (§1.4.1.3, §1.4.2.3).

I argued above that the symbolic threshold also requires salience-deciding inference given that it involves pragmatic inferences about grounds. This is supported by the discussion about natural salience in Lewis conventions (§1.4.1.2). In what follows, 'pragmatic inference' will mean context-, salience- and relevance-deciding inference about speaker communicative intention.

However, a number of sections above also suggested a role for insightful inference or creative leaps, especially in hypothesis generation. Sperber and Wilson downplay this in favour of deductive inference, but I argued above that novel symbols and gestures are problematic in this regard, and that we thus need to look more closely at these non-deductive kinds of inference.

1.5.3 A definition of 'symbol' suitable for discussing the symbolic threshold

I'll set the definition out first, then explain certain key points, though the justification for each of those points has already been set out somewhere above. This definition is not intended to be general (i.e. applicable to everyone who has ever talked about symbols). Nor does it derive from semiotic theory. Rather, it is a definition that will help me make some sense, in the rest of this dissertation, of the evolution of symbolic communication in the human species.

I propose that a symbol is an ostensive sign that requires a specific kind of interpretant if it is to be understood the first time it is encountered. An interpretant is the effect that perceiving a representamen has on whoever is interpreting it, and in the case of a symbol, that effect is a spontaneous hypothesis about the semiotic ground, i.e. a salience-deciding inference yielding an hypothesis about which aspect of the object the speaker is directing one's attention to in using the representamen. The hypothesis about the semiotic ground is part of a larger context- and relevance-deciding pragmatic inference about speaker intention.

I focus here on the first encounter with the sign. If you were in a Robinson Crusoe-like situation and heard Friday utter two words, one a word in his language and one a nonce word, it would make no difference which is the word and which the invention: at this stage, it is not necessary for a symbol to be conventional in order for you to work it out. On the other hand, if you

and he continued to use the nonce word once you'd worked out its meaning, then it would then become conventional. So 'conventional' is a description of a symbol already in use, but not a crucial aspect of crossing the symbolic threshold.

On encountering an ostensive sign humans spontaneously form an interpretant. In the case of a known word, that will be the appropriate representation; in the case of a novel word, it is an hypothesis. I set out in ch. 3 just what this entails. For animals, interpretant representations may spontaneously activate in response to innate signals, but they typically do not spontaneously form hypotheses about the grounds of novel signals, though a novel signal may attract or direct their attention to some degree. Animal interpretants are typically also peripheral, while humans interpretant representations or hypotheses may engage in further central processes. Chimpanzees have been taught novel symbols, but this typically involves laborious training by a human, and so typically doesn't involve a *spontaneous* hypothesis.

The hypothesis concerns an inference about a ground, but I do not claim that this ground cannot be iconic or indexical; just that it must be inferred. I discussed how these terms are not mutually incompatible (§1.2.4) and also how there is a continuum from purely perceptual icons to icons that require a great deal of cognitive effort to understand (§1.3.1.2). The latter are more symbolic in my terms because they require inference; the former are not because they are more purely perceptual. Figure 1.4 is an example of an iconic ground that requires salience-deciding inference. It does not convey all possible facts about Harrison Ford, such as his height or age. Rather it selectively portrays him as the actor who played Indiana Jones. Presumably seeing the picture activates (in a way not relevant to my discussion just now) representations WHIP and HAT, among others. These in turn activate INDIANA JONES, which activates HARRISON FORD. This process is an example of pragmatic inference, like when Peter interpreted the relevance of Mary's mentioning the Capri restaurant. You can (typically) *see* what a photo resembles and feel quite confident about it, so that doesn't involve an hypothesis, but the same is not true of icons like fig. 1.4. Caricatures may be an intermediate case closer to purely perceptual icons; iconic sign language gestures are an intermediate case closer to non-iconic words.

Finally, the formation of the hypothesis is part of a pragmatic infer-

ence about the speaker's communicative intention. That means it will be context-, salience- and relevance-deciding. It is possible for a linguistic human to scaffold a word learning task for someone else in order to limit context, emphasise salience, or be more transparent about relevance, but there will again be a continuum between highly constrained and highly unconstrained cases. Holding up my wine glass to indicate you don't need to open another bottle provides a comparatively small amount of constraining information because the same gesture was involved in holding up a rabbit above. Holding up my wine glass, pointing at the place where the wine is close to the rim, and shaking my head while glancing at the bottle provides somewhat more constraining information. Saying '*gavagai*' while a rabbit runs past in the distance is less constrained than saying it while holding an illustration of mammalian taxonomy in front of your audience's eyes and touching your finger to the rabbit. Even quite constrained cases need not be deductive, though.

There can still be, to an extent, a distinction between inference about ground and inference about speaker intention. If I know what 'rabbit' means, I must still make an inference about what you're conveying about it on a particular occasion, so I must infer your intention. In a game of Pictionary, understanding the game means knowing that the drawer's intention is to produce a picture that will make you say the word on her cue card, but you must still make an inference about semiotic ground. But both are required at the symbolic threshold.

So the upshot is that evolving the ability to form spontaneous hypotheses about grounds is part of what took our ancestors over the symbolic threshold, and this meant evolving the ability to make salience-, context-, and relevance-deciding inferences about speaker intention. Since inference is a poorly explored part of language evolution, this dissertation will focus on the evolution of this particular kind of inference rather than on speaker intention.

1.5.4 Conclusions

This section has drawn the focus of the debate to hypothesis generation. Having provided a number of reasons drawn from cognitive science (Fitch and Tomasello) and pragmatics (Sperber and Wilson) why symbols origins should be considered inherently inferential, I highlighted a number of fea-

tures of this inference. The main feature is that it is open-ended. It is context-, relevance- and salience-deciding, as is true of pragmatic inference in general, though I argued that Sperber and Wilson's claims about deduction in this regard would not extend to the symbolic threshold. Other features, mentioned but not fully explored yet, include creativity, analogy and insight.

The key task for inference about novel symbols is deciding the ground as part-and-parcel of deciding speaker intention. It is only by generating hypotheses about speaker intention that one comes to understand the symbolic ground, though modern children in carefully scaffolded joint-attentional linguistic environments may have this task dramatically simplified for them. Once a symbol's ground has been established, the role of inference may be downplayed in further encounters with that symbol. These various simplifications have allowed researchers in language evolution to gloss over the nature of hypothesis generation, whereas I propose to examine it in detail.

The following two chapters will explore the relationships between these ideas in more detail. Ch. 2 posits an inferential hierarchy (like Fitch's intentional one) such that we can evaluate to what extent certain groups of animals are capable of complex forms of inference, allowing us to consider what our ancestors may have been capable of.

Ch. 3 looks at the most complex level of inference in more detail, which is where questions about hypothesis generation and evaluation are discussed. That chapter will examine Peirce's theory of abduction, which is his particular take on hypothesis generation. Key terms from this chapter will be related to hypothesis generation, such as insight, creativity and analogy. This will allow for a clearer contrast with induction and for testable predictions about symbol learning in contexts of varying complexity.

Chapter 2

An Inferential Hierarchy

2.1 Introduction

Humans and animals perceive their environment and produce behaviour. Between these two interfaces with the external world, various processes will be going on in the brain or mind. 'Rationality', 'inference' and 'reasoning' are among the terms applied to types of these internal processes. It may be uncontroversial to claim that humans are capable of reasoning or inference, or that animals are capable of rational behaviour, but whether animals display inference is unclear (Allen, 2006). How we evolved these abilities is thus unclear. Further, the distinction between inference and reasoning is currently rather vague, and subdivisions in these nebulous categories tend to be painted with rather broad strokes: people talk of deductive and non-deductive reasoning, or of deductive and inductive reasoning, or of Bayesian induction, but not in terms much more detailed than that.

Since the previous chapter argued that the evolution of symbols is best explained in inferential terms, a description of the symbolic threshold will inherit any lack of clarity found in our theories of inference. This chapter will thus outline an evolutionary hierarchy with rationality at the base, inference as an intermediate term, and reasoning at the top (fig. 2.1). That is, minimal rationality is the simplest, oldest ability, shared with the most other species. Inference is more complex, is a more recent development, and is shared with fewer species. Reasoning is the most complex, the most recent, and is probably limited to humans.

Having set out some basic terms relevant to these distinctions, I will then

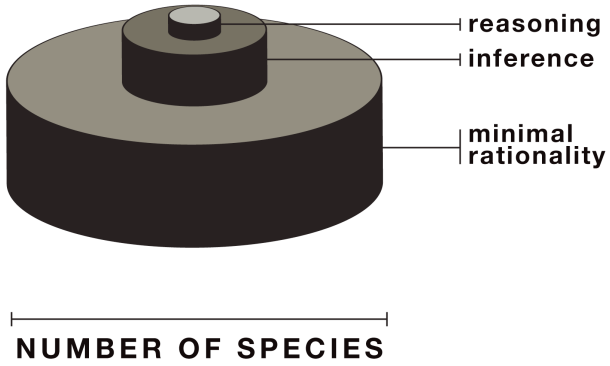


Figure 2.1: An inferential hierarchy ranging from the most basic to the most sophisticated capacity. The level corresponding to reasoning is very narrow (it is limited to just humans) and very shallow (it is comparatively recent). The vast width and depth of the lowest level indicate that it is a great deal more widespread and older. Inference is intermediate.

turn to a closer examination of inference, distinguishing minimal inference from more complex forms, and identifying the level at which symbol learning becomes possible. It will turn out that identifying minimal rationality and reasoning at either extreme of the spectrum is comparatively straightforward, whereas inference seems to be a large gap of the here-be-dragons variety in the middle.

The basic premise here should *not* be read as supporting a saltationist account of evolution: these divisions are not intended to represent sudden, large jumps in ability. Rather, I assume that gradual increases in complexity occurred, and am simply aiming to identify points of interest along that continuum. Nor are these divisions mutually exclusive. A human may be capable of simultaneously bringing all three processes to bear on a given problem or may shift from one process to another on depending on context. That is, the hierarchy is more properly called a 'layered model' (Zlatev, 2008) because each stage does not represent a total replacement of what came before.

Throughout, a major concern will be aligning theoretical distinctions with those supported by empirical evidence: claiming that there is a distinction between rationality and inference is only really sensible if it turns out that this distinction lines up with behavioural or neurological differ-

ences. At times, these links may be suggestive rather than conclusive, but since this is somewhat of a *terra incognita*, my claims are intended to be encouraging first steps, rather than anything stronger. Apart from my overall aims with regard to symbol evolution, one subsidiary aim in this chapter is to identify a stalemate in empirical approaches to animal inference and to propose a way out of this. Another is to show the limits of a computational theory of mind in discussing the evolution of mind. A third is to make good on my criticisms of Sperber and Wilson (1995) from ch. 1.

2.2 Some Key Concepts

2.2.1 Dual-process models

Dual-process theories are a disparate set of approaches that assume the mind simultaneously operates in at least two distinct ways called 'system 1' and 'system 2' (Tversky and Kahneman, 1983). Since a vast range of researchers have been interested in this topic, characterisations of each system have at best family resemblance. Evans (2008) distinguishes four clusters of characterisations in the literature (fig. 2.2) based on attributes of consciousness, evolution, function, and individual differences. Mercier and Sperber (2009) tentatively call system 1 'inference' and system 2 'reasoning', since they focus on the 'consciousness' and 'evolution' clusters: reasoning is reflective, linked to language, and uniquely human. This is a distinction I support.

This does not imply, however, that this is the only dual-process distinction worth attending to, or that all other descriptions of system 1 and 2 processes reduce to this. In fact, other descriptions do not even have to align with this. In what follows I will make a second, orthogonal distinction between two systems based on functional characteristics (associative vs. rule-based or syntactic architectures, and pragmatic descriptions vs. logical norms). In fact, since I am focusing on the evolution of language, and since reasoning requires language, I will not discuss reasoning much in what follows. While I referred to the distinction in Mercier and Sperber (2009) above to constrast inference and reasoning, their distinction will not inform the focus of this chapter, which is an examination of different kinds of inference. When I mention 'dual-systems' in what follows, then, I will be referring to inference and to the just-mentioned functional characterisation rather than that of Mercier and Sperber.

System 1	System 2
Cluster 1 (Consciousness)	
Unconscious (preconscious)	Conscious
Implicit	Explicit
Automatic	Controlled
Low effort	High effort
Rapid	Slow
High capacity	Low capacity
Default process	Inhibitory
Holistic, perceptual	Analytic, reflective
Cluster 2 (Evolution)	
Evolutionarily old	Evolutionarily recent
Evolutionary rationality	Individual rationality
Shared with animals	Uniquely human
Nonverbal	Linked to language
Modular cognition	Fluid intelligence
Cluster 3 (Functional characteristics)	
Associative	Rule based
Domain specific	Domain general
Contextualized	Abstract
Pragmatic	Logical
Parallel	Sequential
Stereotypical	Egalitarian
Cluster 4 (Individual differences)	
Universal	Heritable
Independent of general intelligence	Linked to general intelligence
Independent of working memory	Limited by working memory capacity

Figure 2.2: Clusters of attributes associated with dual systems of thinking (Evans, 2008). I have highlighted in pink the focal attributes which distinguish inference from reasoning. In blue are the attributes that will inform my discussion of minimal rationality and inference. Naturally, highlighted terms may entail others: if reasoning is limited to humans, for instance, it is also evolutionarily recent.

2.2.2 Approaches to rationality and levels of analysis

Kacelnik (2006) distinguishes a number of approaches to rationality: PP-rationality is that which is typical of philosophy and psychology; E-rationality that which is typical of economics and B-rationality of biology. PP-rationality focuses on reasoning and inference and their role in behaviour and is thus concerned with cognitive or internal processes: a behaviour or belief is rational in this sense if it is the result of inference or reasoning. E-rationality is not concerned with internal processes, but rather with the expected utility of a behaviour: an action is rational if it can be expected to maximise utility for an animal in a certain environment, regardless of whatever internal processes led to the behaviour. B-rationality is related to E-rationality but narrower: it replaces the vague notion of utility with biological fitness.

Bayesian inductive inference (with which I will contrast my claims about hypothesis generation) is the default tool of the rational analysis approach (Anderson 1991; Chater and Oaksford 2008), which assumes E-rationality since it explains behaviour in terms of the structure of the environment by providing an analysis of what would be the optimal behaviour by a creature in that environment, given certain goals.

On the other hand, since I am presenting a psychological account of the role of inference in symbol evolution, I will be concerned with PP-rationality. However, Kacelnik admits that this definition is difficult to apply to animals since they don't seem capable of reasoning. Since I distinguish inference and reasoning, I suggest the following as a less anthropocentric definition: PP-rationality focuses on the explanatory role of internal processes linking perception to behaviour and, by extension, on the content associated with those processes.

Given that cognition is such a complex phenomenon to investigate, several researchers have tried to simplify the task by positing a number of levels of analysis (most commonly, three), each focusing on different aspects of the whole, or different degrees of abstraction. The following table is not intended to suggest that the terms in each row are equivalent, for there are differences[1]; nor does it suggest that I am committed to the asumptions or implications they bring with them. Rather, I present these here because some aspect of each aligns roughly with the others, and various subsets may

[1]For instance, the three rightmost columns focus on semanticity at the most abstract level, while Marr's computational level focuses on function, hence the double lines.

be familiar to readers from different backgrounds, so they merely act as an anchor for a few points that follow in the rest of this chapter.

	Marr (1982)	Pylyshyn (1984)	Dennett (1987)	Glass et al. (1979)
Most abstract	computational	semantic	intentional	content
Intermediate	algorithmic	syntactic	design	form
Most conrete	physical			medium

Table 2.1: Levels of Analysis.

The lowest level, the physical, is a matter of concrete implementation. A calculator and human may both be capable of adding two numbers, but a calculator's processes are implemented in electronic circuits, a human's in neurons, and are thus different at the physical level. The level variously called 'algorithmic', 'syntactic', 'design' or 'form' is more abstract. A calculator, unlike a human, represents numbers in binary in order to add them up: one feature of discussions about this level thus concerns the *units* of cognition. Another feature is the *processes* that apply to these units: how a calculator performs addition is different from how a human brain does. In contrast, the highest level is where cognitive processes interact with the world, and the different terms above focus on different aspects of this interaction. Marr's label 'computational' focuses on the inputs and outputs of the cognitive process and the function of that process. Despite algorithmic-level differences between calculators and humans, we nonetheless perform the same computational function when we add up two numbers. In what follows, I'll mostly be concerned with interactions between the computational and algorithmic levels, though occasionally I'll refer to physical differences.

2.2.3 Problematic terms

Some ambiguous or similar-sounding terms have unfortunately cropped up in the literature. There is an ambiguity in the term 'computational', and an important difference between 'rational' and 'rationalist'. These ambiguities and differences are intertwined, however, so this requires a little exposition. It is possible to give a computational-level analysis of a problem and remain agnostic about the kind of processes at the algorithmic level, but it is also common to argue for specific kinds of process. Fodor (2001) outlines two candidates by distinguishing two kinds of mental causation.

(I) Suppose we believe the proposition 'John is bald' and that believing this proposition involves the existence in our minds of a particular mental representation BALD(JOHN), which has the logical form `Fa` in Fodor's notation, where `F` expresses a property and `a` an individual. Take `Ga` to represent the proposition that 'John is in need of a haircut', so `Fa`$\longrightarrow$`¬Ga` corresponds to the conditional 'If John is bald, then he isn't in need of a haircut'. The argument `Fa`; `Fa`$\longrightarrow$`¬Ga` $\therefore$ `¬Ga` is logically valid, but validity here is a formal notion: it has nothing to do with the meaning of baldness, haircuts or who John is, but rather has to do with the syntax of the propositions in which they find themselves. A syntactic process mechanically derives that conclusion from those premises by virtue of their logical form, and if the second premise were `Fa`$\longrightarrow$`Ga`, then the conclusion would be `Ga`, even though it's semantically implausible that, if John is bald, he is in need of a haircut.

In this example, the relevant syntactic properties are local: the validity of the argument depends on the syntax of the propositions in the argument, and not on anything outside that argument, much as the spelling of a word depends only on the arrangement of letters within that word (Fodor, 2001). Processes that rely only on local properties are thus contextually constrained. If some inferences rely on nonlocal properties, then they may be contextually unconstrained. I highlighted the importance of context for pragmatic inference (§1.5.2) and a major thread running throughout this chapter is the evolution of contextually unconstrained inference. In the next chapter, this will be related to hypothesis generation (contextually unconstrained) and evaluation (contextually constrained).

(II) Now suppose in addition that when you think of John, this sometimes causes the representation BALD to become activated, but when I think of John, this always causes BALD to become activated: possibly baldness is less central to your representation of him than it is for me, or my representation JOHN activates BALD more strongly than yours does. This associative relation is not a formal property of these representations: whereas one could substitute any well-formed proposition for `Fa` and `Ga` without affecting the validity of the above argument, the relationship between JOHN and BALD is a property of just that pair. Substituting representation of another individual or another property would result in a different casual relationship. Further, while the validity of the argument doesn't depend on who is evaluating it, this associative relationship is an accidental feature of particular minds.

Based on these two kinds of mental causation, Fodor distinguishes two approaches to the algorithmic level in human cognition. An empiricist psychology maintains that the structure of a thought and the role it plays in cognition are exhaustively described by the associative relationships of **(II)**. A rationalist psychology maintains that the structure of a thought and the role it plays in cognition depend on the formal properties of **(I)**, and that these are not reducible to associative relationships. The terms 'empiricist' and 'rationalist' have alternative meanings in philosophy, but I intend them here in the specific sense of Fodor (2001).

One benefit of a rationalist psychology is that it explains how cognition is truth-preserving (Fodor, 2001): if the premises are true, then truthful conclusions can be ensured by a suitable mechanical process sensitive to formal features. Call this the logical or normative feature of a rationalist psychology. In order to explain how a rationalist psychology achieves this aim, an influential theory in cognitive science assumes that the mind is interestingly like a Turing-style computer. This is the Computational Theory of Mind (CTM Fodor and Pylyshyn, 1988; Pinker, 1997; Fodor, 2001). The term 'computational' is thus ambiguous, referring either to the computational level of analysis or to the CTM thesis that the mind is computational at the algorithmic level. A computational-level analysis needn't suppose a CTM, though. In order to avoid confusion, I'll use 'syntactic' to mean 'computational in the CTM sense'. Other terms for this in the literature include 'rule-based', 'classical', or 'symbolic[2]'.

In this chapter and the next, I will argue for a dual-process model that distinguishes a rationalist psychology (system 2) which makes normative and syntactic assumptions about cognition from an empiricist psychology (system 1) which assumes cognition is associationist (or subsymbolic or non-classical) and does not assume normativity. It simply describes what happens, not what *ought* to happen and is thus called 'pragmatic' or 'descriptive'. The term 'rationalist' thus carries with it specific assumptions, unlike the general term 'rational', but because I will argue something can be minimally rational without being rationalist, this distinction is important. Danks (2008) argues that there is no reason to suppose just one algorithmic level. Instead, a process at a higher level algorithmic level can handle the inputs and outputs for a processes at a lower level algorithmic level. If

[2]'Symbolic' here meant in the private, not public sense (cf. §1.2.2).

that's the case, I don't see why some algorithmic levels can't be empiricist and others rationalist.

2.2.4 Content

Since I have argued that inference was necessary for symbolic language, I am committed to the claim that there are some kinds of inference that do not make use of content or processes that require language. The previous subsection focused on cognitive processes, whereas this subsection focuses on content. Millikan (2006) argues that reasoning requires propositional content. Since inference is simpler than reasoning, an investigation of whether animals have inference thus involves investigating whether animals have something like propositions, but simpler. Let's call these proto-propositions (Hurford, 2007). The question of whether animals have proto-propositions demands a discussion of whether they have proto-concepts.

Hurford states that a necessary (but not sufficient) criterion for deciding whether an animal has a proto-concept of something is that the animal display 'regular and systematic behaviour in connection with that thing' (2007, 16). He posits a representational continuum from proto-concepts to full, linguistic concepts and identifies generality and volitional control as two dimensions along which the complexity of a representation can vary. He reviews a range of evidence, some of which follows below, suggesting that some animals display sufficient generality or volition in their behaviour to be granted proto-concepts.

Zuberbühler et al. (1999) show that Diana monkeys display a degree of generality. Like related vervet monkeys, Diana monkeys give alarm calls in response to predators. A Diana alarm in response to a leopard doesn't sound like a leopard growl, but both indicate the presence of a leopard. Zuberbühler et al. played recordings of predator signals and of monkey alarm calls and found that monkeys became habituated to different signals if they had the same functional referent, despite sounding different. They conclude that this is best explained if behaviour is mediated by a representation of that predator. Hurford claims that because this representation is somewhat general, activated in response to different signals of a predator, it is a proto-concept. Vervet monkeys also have a small degree of control over whether they sound an alarm or not (Cheney and Seyfarth, 1990), so their representations are thus also somewhat volitional.

Concepts can be conjoined to form propositions consisting of a predicate which has a varying number of arguments. Minimally, there must be at least one argument: a subject. Saying 'Socrates' is not, on its own, true or false; neither is saying 'was snub-nosed'. A proposition combining these, 'Socrates was snub-nosed,' is true.

Hurford (2007) thus discusses the possibility that some animals have proto-propositions where, instead of the subject being conceptually represented, an attentional index (a sort of variable standing for whatever is the focus of the animal's attention) attaches to the predicate proto-concept. So if x represents whatever is in the animal's visual focus, then RABBIT(x) is a proto-proposition roughly equivalent to the proposition 'that's a rabbit'. The possibility of a single representation being in some way like a sentence is an idea that has some philosophical pedigree, since Quine (1992) admits the possibility that observational sentences (such as the proposition 'that's a rabbit') could just as well be represented by a single noun 'rabbit!' in certain cases.

I can now turn to look at rationality and inference as intuitive or subpersonal system 1 processes not requiring propositions or linguistic concepts. This step is an important part of investigating symbol origins because our ancestors would have had to have been capable of inference to cross the symbolic threshold. But it is by no means clear that non-human animals can infer anything. If I can provide an objective way of distinguishing minimal rationality (§2.3) and inference (§2.4), then we will be better situated to use information about the inferential ability of our relatives to speculate about the inferential abilities of our pre-linguistic ancestors. The final part of the puzzle will be providing empirically justifiable distinctions between simpler and more complex forms of inference (§2.5), such that we share the former with our relatives, but the latter only with our ancestors. This will explain why chimpanzees can be taught laboriously to use symbols, while humans do so spontaneously.

2.3 Minimal rationality

The lowest level in the inferential hierarchy will be based on the account of minimal rationality in Dretske (2006)[3]. Dretske is interested in contrasting the rational behaviour of some animals with the mechanical behaviour of, for instance, a thermometer or a plant: in the former case, internal representations play an explanatory role; in the latter cases, not. Minimal rationality describes behaviour that is 'not only under the *causal* control of thought, but ... *explained* by the *content* of these thoughts' (2006, 107, emphasis mine).

He provides an analogy to distinguish causation from explanatory content. If I say 'vibrate rapidly' into a microphone, the microphone's diaphragm will vibrate rapidly. This vibration is caused by what I say, but it is not explained by the content of what I say: it would still vibrate if I had said 'do not vibrate rapidly'. Thus, when discussing how an animal can be minimally rational but a thermometer cannot, we are not merely interested in the causal chain from perceptual input to behavioural output, but also with the content of any intermediate internal states and the role of this content in explaining behaviour. Dretske goes on to illustrate this by contrasting comparatively mechanical behaviour in a plant with minimally rational behaviour in a bird.

The scarlet gilia is a plant that changes from red to white in the middle of July each year. Early in the flowering season, humming birds are their chief pollinator and hummingbirds are more attracted to red blooms; later, hawkmoths are the main pollinators and they apparently prefer white. Dretske allows that this counts as B-rationality in the terms outlined above (§2.2.2), but not as minimally rational in his terms. There are internal states of the plant that have a mechanistic relationship with the environment at the relevant time of year. Dretske allows that such internal states represent the environment and that these states cause the change in colour, acting like a biological clock. However, he denies that the *content* of these internal states has anything to do with the colour-changing behaviour and he bases this claim on how we *explain* the colour change. He argues that the explanation for the plant's behaviour is not *its* internal state, but rather, the internal

[3]He contrasts minimal rationality with full rationality, and it seems that full rationality corresponds to my terms 'inference' and 'reasoning', so the discussion here is not about rationality in general.

states of its *ancestors* in the context of some selectional process.

In contrast, Dretske imagines a bird that had eaten a poisonous monarch butterfly, making the bird ill. If the bird then avoids not only monarchs, but also similar looking viceroys (which are not poisonous), then the content of its representations plays some role in explaining this avoidance behaviour. Like the case with the scarlet gilia, the bird's internal state is a cause of its behaviour. Unlike the plant, however, Dretske claims that the content of this representation explains the bird's behaviour: the bird 'avoids something not because it tastes bad (the viceroy doesn't taste bad), but because it "thinks" it tastes bad' (2006, 114). Because the content of the representation has this explanatory value, the behaviour is minimally rational.

While the plant's behaviour, then, was explained by something in its species' phylogeny, the bird's behaviour is explained by something in its own ontogeny. That is, Dretske seems to base the matter of minimal rationality on whether some representation is evolved or learned: the latter behaviour is more flexible and the former more mechanical[4].

However, minimal rationality will not extend to more complex situations. It is designed to distinguish minimally rational from mechanical behaviour, but not to distinguish more sophisticated behaviour from either of these. But what if much of a category is innate, but its application depends on learning? Cheney and Seyfarth (1990) showed that vervet infants initially react to a range of overhead stimuli, including falling leaves, as they would to eagles. They gradually learn, however, to limit this behaviour to eagles. Some aspects of the behaviour are thus innate while others are learnt, so it is unclear where we stand with regard to Dretske's criteria. Further, I reviewed evidence suggesting that the vervets' representations are proto-concepts and thus contentful (§2.2.4). However, it is a proto-concept of a biologically salient category and thus principally explicable by the animals' phylogeny. Intuitively, the sophistication of the behaviour *seems* to be at least as rational as the bird's avoidance of certain butterflies. The criteria for minimal rationality, however, do not decide the matter one way or the other, and 'seems' isn't good enough: we need a more principled way of evaluating the explanatory balance between evolutionary history and representational content. The following section thus offers a definition of 'inference' that

[4]Naturally there is a sense of 'rational' that encompasses both, but that sense is broader than what Dretske is aiming for here.

solves the problem by building on Dretske's definition rather than abandoning it.

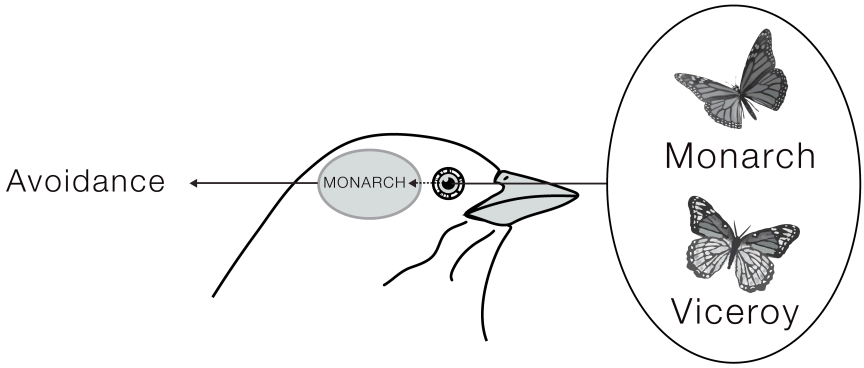


Figure 2.3: Minimal rationality requires nothing more than an empiricist psychology.

Before moving on to inference, though, I want to be clear where we stand relative to the problematic or ambiguous terms in §2.2.3. Assuming that the bird has a representation MONARCH, it's probably the case that perception of both viceroys and monarchs activates the representation which in turn activates avoidance behaviour, and that this all occurs through associative processes (fig. 2.3). Further, if this is descriptively accurate, there is no indication that the bird's cognitive system is truth-preserving (though *we* might make claims about truth from outside that system), because the representation doesn't distinguish these butterfly categories. The causal role of the representation is explained by some of the butterflies (but not others) being poisonous, but we cannot assume normativity in a system relative to categories that are simply not represented by that system. Alternatively and more straightforwardly, if there are no propositions, there is no truth of the matter. Because this process is associationist and doesn't assume normativity, minimal rationality assumes an empiricist, not a rationalist psychology. It is unfortunate that the terms share a morphemic root, but the collocations should help distinguish them in what follows.

This will have implications for inference at the symbolic threshold. Many accounts of sophisticated human cognition focus on similarities between inference and reasoning, and thus frame inference in normative terms[5].

[5]For instance, normativity is assumed by the rational analysis approach (Anderson,

Counter to this trend, I wish to emphasise similarities between (some kinds of) inference and minimal rationality. Since minimal rationality is empiricist and thus non-normative and associationist, I will argue that some kinds of inference are non-normative and associationist.

In particular, I will argue that context-, salience- and relevance-deciding inference needs an empiricist psychology, so it's worth explaining now why it doesn't matter that minimal rationality is empiricist, even though Fodor (2001) is dismissive of the value of an empiricist psychology. He argues that only a rationalist psychology can explain laws of thought (i.e. the way normative processes underlie and explain core features of human cognition such as productivity, systematicity and compositionality[6]). We have no evidence, though, that cognition at the minimally rational level is productive, systematic or compositional, so there is no reason to preclude an empiricist psychology on these grounds: MONARCH(x) is never going to have to enter into a *modus ponens* for this bird.

I think it makes sense of the evolutionary trajectory to see how inference develops out of minimal rationality and to wait until some cognitive process demonstrably needs normativity or compositionality before worrying about the appropriateness an empiricist psychology. Premature comparisons with reasoning risk making procrustean assumptions and setting the bar too high.

2.4 Minimal Inference

In this section, I will first outline and unpack a definition of minimal inference (§2.4.1). This is intended to be just one small step more complex than Dretske's minimal rationality so it will not look as sophisticated as some complex inferences. I will apply the definition of minimal inference to the puzzling case of transitive inference in animals and show how it points the way to resolving a methodological stalemate (§2.4.2). Finally, I discuss some neurological evidence for my claims (§2.4.3).

1991), and Bayesianism is how human cognition is supposed to achieve this (Griffiths et al., 2008). However, this replaces 'truth' with an ecological notion of optimality. As a result, its processes involve probability rather than deductive necessity. I thus extend Fodor's narrow sense of 'normative' to mean 'truth-preserving *or* optimal'.

[6] Chalmers (1990) argues connectionism can provide compositionality. But since I don't need compositionality here, I won't get into it.

2.4.1 Definition

For a cognitive process to be inferential, I propose that

1. It must be minimally rational, but
2. Deciding whether it is minimally rational requires contextual information

By 'contextual information', I mean this: for something to be minimally rational as per point 1, some representational content must play an explanatory role in accounting for behaviour. Contextual information refers to *other* representational content activated at roughly the same time but activated by different stimuli. So if a vervet or Diana monkey has a representation of whether kin are present while its LEOPARD representation is active, then the former counts as contextual information since the stimulus `kin` is distinct from the stimulus `leopard`. Per point 2, if contextual information is needed to decide whether something is rational, then the process is minimally inferential (fig. 2.4). It turns out that it wouldn't be rational for the vervet to attract the predator's attention unless kin are present (Cheney and Seyfarth, 1990).

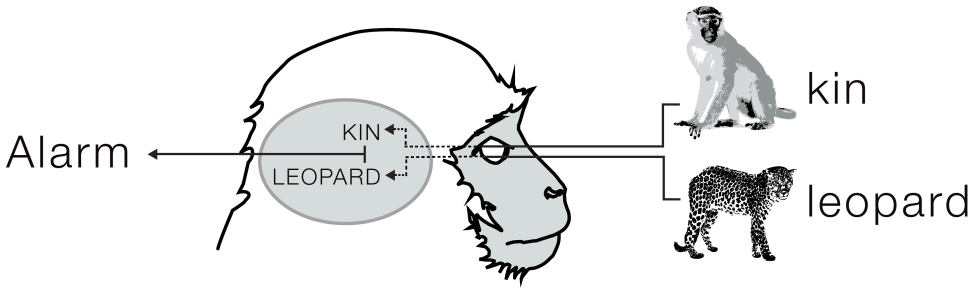


Figure 2.4: Minimal inference requires contextual representational information: if no kin are present, it is not rational for a monkey to draw a predator's attention to itself by sounding an alarm.

The problem previously was that Dretske distinguished the bird's rationality from the scarlet gilia's mechanical behaviour because phylogeny explained the latter while ontogenetically acquired content explained the former. The causal link between vervets' LEOPARD representations and subsequent alarm or avoidance behaviour seems explicable by both. The pro-

posed definition makes a small but important difference in deciding just what is explained by each. Even if we allow that vervet phylogeny (rather than representational content) is the ultimate explanation for why the LEOPARD representation causes alarm call and avoidance behaviour, this is about the existence of the behaviour in the species in general. On its own it does not explain whether or not a particular vervet gives an alarm on a particular occasion. That is (at least partially) explained by whether its KIN representation is active at roughly the same time as its LEOPARD representation[7]. Since deciding whether instances of behaviour are minimally rational requires this contextual information, the process is minimally inferential.

Paraphrasing, a minimally rational process involves a comparatively linear causal chain relating representations to peripheral information (i.e. perceptual input or behavioural output, fig. 2.3). An inferential process involves cognition that additionally relates representations to non-peripheral contextual information: other representations (fig. 2.4). I've already discussed the importance of non-peripheral processes (§1.3.2): this is a fairly basic example of one.

It might be argued that, instead, vervets are sensitive to complex stimuli as in fig. 2.5. This seems more like fig. 2.3. I admit that I have no evidence that kin presence should be considered a distinct representation so I cannot make any claim stronger than this: by the above definition for inference, vervet avoidance behaviour would count as minimally inferential *if* my description is accurate. The next subsection, however, will provide stronger empirical support for a difference between rationality and inference in the case of transitive inference.

2.4.2 The test case: Transitive Inference

I give an outline of transitive inference (TI) and contrast two explanations of animal behaviour that is functionally similar to TI. I examine why current attempts to choose between these explanations struggle to settle the matter definitively, and then show why my definition of inference can do so. The point here is to show that my definition is empirically useful.

If we know that Alice is faster than Barbara and that Barbara is faster

[7]Representations of contextual information do not have to be very sophisticated. I don't claim that a vervet must have a proto-concept of KIN: we lack evidence for this, unlike the LEOPARD proto-concept.

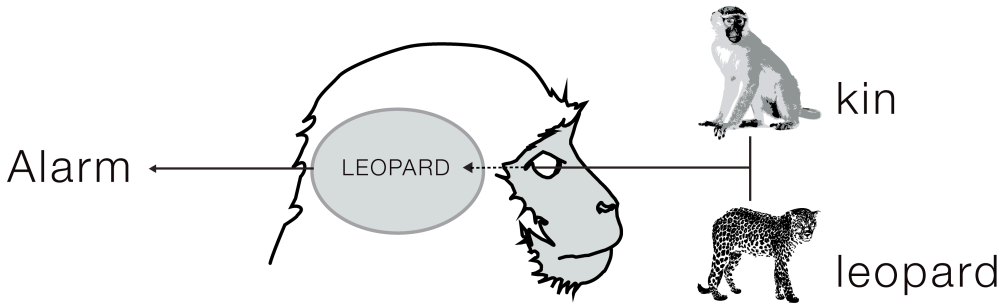


Figure 2.5: A possible interpretation that would be minimally rational, not minimally inferential, since the contextual information is perceptual, not representational.

than Cathy, we can infer that Alice is faster than Cathy despite never having seen them run together because the relationship FASTER-THAN is transitive (in a logical, rather than grammatical sense). If an animal is trained on pairs of arbitrary stimuli `a`, `b`, `c`, such that choosing `a` is rewarded when presented with the pair `a` and `b` while choosing `b` is rewarded when presented with the pair `b` and `c`, then a number of vertebrates, including chimpanzees (Gillan, 1991), monkeys (McGonigle and Chalmers, 1977), rats (Dusek and Eichenbaum, 1997), pigeons (von Fersen et al., 1991) and even fish (Vasconcelos, 2008)[8], are able to choose `a` over `c` despite not having seen them together[9].

So some animals are capable of behaviour that is functionally or descriptively similar to TI, but animal behaviour researchers disagree on which of the following is the best explanation for how a given species actually produces this behaviour (Allen, 2006, though he describes these as loose families of approaches, rather than anything more unified):

Cognitive: The animals explicitly represent the relevant transitive relationships `a>b>c` and cognitive processes use these representations to produce the behaviour

Associative: The animals have no representations of the transitive re-

[8]Though the demonstration for fish involved social dominance relations, not arbitrary stimuli.

[9]The reward values of the series can be represented as `a>b>c` and each training pair as `a+b−` and `b+c−`. Testing `a` against `c` is represented `a?c`.

lationships, and produce the behaviour through simpler associative mechanisms

O'Doherty (2004) claims that, for human and non-human primates, brain regions including the orbitofrontal cortex, amygdala and ventral striatum have neurons that fire in response to reward, firing faster for higher reward. In an associative account of basic TI, stimulus a might come to be associated with high levels of reward-neuron firing, b with medium levels and c with low-levels or none. This is associative in the sense that choosing a over c could be explained only by the higher reward value associated with a, without requiring explicit representation of a relation between a and c.

In order to address this problem, the training series can be extended to a>b>c>d>e. Now, both b and d are rewarded 50% of the time and not rewarded 50% of the time, so an animal's choosing b over d in this situation seems to be better evidence for the cognitive accounts. However, Value Transfer Theory (von Fersen et al., 1991) argues that, because b is seen with highly-rewarded a half the time while d is seen with sometimes-rewarded c, higher reward-neuron firing due to a can transfer to simultaneously perceived b; less transfer happens from c to d.

Such an interpretation is borne out by an experiment in Zentall and Sherburne (1994). Here, pigeons were trained on pairs a+b− and c±d−, but not on b+c−. Neither b nor d were ever rewarded, but a was always rewarded and c was rewarded only half the time. Neither b nor d would have any reward value on their own, but according to Value Transfer Theory, b could derive some reward value from its perceptual association with always-rewarded a while d would derive less reward value from its association with sometimes-rewarded c. When tested with b?d, the pigeons chose b significantly more often. There was no relationship between b and c in training, so no inference about the relationship between b and d is possible. The associative account is thus a potential explanation of TI in pigeons, since TI-like behaviour can be produced without any representation of the series as a whole. However, since the pigeons in Zentall and Sherburne (1994) didn't choose b at test as often as those in a straightforward five-item task did (von Fersen et al., 1991) — 64.6% vs 87.5%, respectively — a cognitive explanation cannot be ruled out entirely.

The length of the series could be increased still further to decide the matter, but training animals on series of arbitrary stimuli is extremely dif-

ficult, so this has its limits. Allen (2006) argues that such limitations in methodology have brought the question of whether TI is associative or cognitive to an impasse. This is problematic for my project since I argue that the evolution of symbols is intimately connected with the evolution of inference. If the best data currently available cannot provide uncontroversial answers about the evolution of inference, given this impasse, then how we crossed the symbolic threshold is obscured. However, I believe my definition of inference, based in turn on Dretske's minimal rationality, offers a way out by making a somewhat counter-intuitive prediction about what should happen at various points along the series `a>b>c>d>e`.

Assume, per an associative account, that `a` is associated with high levels of reward-neuron firing. Choosing `a` is always rational, no matter what other stimuli are present. Choosing `e` is never rational. No similar claim can be made about `b`, though: the rationality of choosing `b` depends on whether it's perceived with `a` or `c`. That is, deciding the rationality of choosing `b` needs contextual information, unlike `a` or `e`. Thus, I would have to claim that choosing `a` and not-choosing `e` is minimally rational, but choosing `b`, `c` or `d` is inferential. Dusek and Eichenbaum (1997) and Greene et al. (2006) acknowledge a distinction between terminal and non-terminal pairs, though they call the former 'associative' and the latter 'inferential' while I call the former 'minimally rational'. The difference is that I base my definition of the latter on the former, while they seem to treat them as distinct phenomena.

If some kind of cognitive account is true, we would thus expect qualitative behavioural or neurological differences between terminal and non-terminal elements of the series according to my definition of inference. If the associative account is true, on the other hand, we would merely expect graded quantitative differences.

2.4.3 Neurological and behavioural evidence in humans and rats

Greene et al. (2006) trained human subjects on a 5-element TI task. Most subjects managed to learn the pair-wise relationships, though pairs involving a terminal element (`a+b−`, `d+e−`) were easier (involving lower error rates and latencies) than inner pairs (`b+c−`, `c+d−`), so some subjects failed to learn one of the latter. Subjects were then tested on novel inference pair `b?d` and novel non-inference pair `a?e`. Subjects were much more successful at the

latter than the former. Those who managed to respond correctly to the former are called BD-performers.

The authors found that the hippocampus was more active during the testing phase for central (i.e. inferential) pairs than for terminal (i.e. rational) ones. Further, hippocampus activity at training was an accurate predictor of who would probably turn out to be BD-performers at test and who wouldn't. There are thus both behavioural and neurological differences between tasks that I have labelled 'inferential' and those that are merely 'minimally rational', suggesting that they are qualitatively different cognitive processes.

Dusek and Eichenbaum (1997) conducted an experiment with rats where, instead of arbitrary visual stimuli, a 5-element TI series involved scents arbitrarily grouped into reward/no-reward pairs such that, for example, `paprika` was rewarded over `coffee` but `coffee` was rewarded over `basil`. Their control group was neurologically intact, while their experimental groups had various types of surgery disconnecting the hippocampus from its surroundings. Both groups succeeded at trained-pair discrimination tasks and at novel terminal-pair (i.e. minimally rational) discrimination tasks, but only the control group succeeded above chance at novel non-terminal-pair (i.e. inferential) discrimination tasks.

So *if* we shift the focus from testing whether the whole series is represented (which is how Allen, 2006, described it above) to testing whether there are cognitive differences between terminal and non-terminal pairs, we may be able to support a cognitive over the associative account of TI for rats. Greene et al. (2006) skirt around this issue by noting with interest that the hippocampus seems to have a dual role, being involved in both contextual tasks and inferential tasks. They go as far as saying that these types of task (contextual and inferential) are related, but stop short of claiming that inferential tasks are just tasks where the representation of contextual information must play a role in explaining behaviour[10], whereas I argue that the one is based on the other, and my definition of inference is thus supported by neurological evidence.

Apart from suggesting one way out of the experimental stalemate iden-

[10]Their definition of inference is this: 'the capacity to make novel decisions on the basis of relevant prior experience, whether by syllogism (e.g., chaining) or by comparative value (e.g., taller, further east)' (Greene et al., 2006, 1157).

tified by Allen (2006), the discussion here is intended to show that context is central to the notion of inference. Traditionally, when it comes to human inference about symbols, the focus has been on features such as ToM, higher-order cognition, joint attention or co-operation (Tomasello, 1999; Hurford, 2007; Fitch, 2010), and I simply wish to add context to the mix. In that case, the nature of this context should be a centrally important way of evaluating inferential complexity and thus differences between human and animal inference, relative to symbolic communication. I had earlier emphasised the distinction between context-bound inference and context-deciding inference in my definition of 'symbol' (§1.5.3), and now turn to look more closely at this distinction.

2.5 Complex Inference

TI is a comparatively simple form of inference, and inference could be made more complex in a number of ways. I first look at a couple of these ways by discussing examples of animal behaviour that seem to require more complex inference (§2.5.1). I then narrow in on the question of whether the context is given prior to inference or decided as part of the inference (§2.5.2).

2.5.1 Examples of complex behaviour

In the context of language evolution and inference, one focus of the discussion of cognitive complexity in Hurford (2007) and the central focus in Penn et al. (2008) is higher-order representations and the relations between them. Some animals can make similarity judgements (Premack and Premack, 1983), responding to stimulus `a` by picking a further `a` instead of `b` (a same/different or S/D task). Chimpanzees can perform a higher-order S/D task (Premack and Premack, 1983): if presented with a pair of stimuli `a/a` they can pick out another pair with the same relationship `b/b` over a pair with a different relationship `c/d`. If presented with differing stimuli `e/f`, they can pick out differing pair `g/h` rather than same pair `i/i`[11].

A study with macaques (Washburn et al., 1997) found that, though they could respond this way to trained pairs, they could not generalise this behaviour to new pairs. D'Amato and Colombo (1985) showed that capuchins

[11] Hurford notes, though, that these chimpanzee subjects, though not language trained, had been previously trained on the concepts SAME and DIFFERENT.

can generalise to a limited extent. Even more pessimistically, Penn et al. (2008) are dismissive of claims of higher-order relational cognition in other animals, claiming that apparent evidence of higher-order S/D behaviour can be explained by non-relational cognition, such as evaluating the degree of informational entropy in same and different pairs. They extend this criticism to other forms of higher-order relational thinking, such as analogy.

So human are exceptionally good at higher-order relations and either animals are less successful at these tasks (the optimistic interpretation) or they do not solve the tasks with anything like higher-order relational cognition (the pessimistic interpretation)[12].

But proficiency at higher-order relationships is not the only difference between humans and other animals, so this is not necessarily the best way of distinguishing human from animal inference with respect to symbol learning. Chimpanzees can be taught higher-order relations and can be taught symbols, but just how they must be trained is itself interesting. I will look briefly at the interpretation in Deacon (1997) of symbol learning experiments in Savage-Rumbaugh and Rumbaugh (1978); Savage-Rumbaugh et al. (1978); and Savage-Rumbaugh et al. (1980). After presenting his discussion of these experiments, I will highlight the role that context plays *in addition to* higher-order relationships. Context has not been a focus in the literature on language evolution.

Savage-Rumbaugh and colleagues performed a series of experiments involving chimpanzees Lana, Austin and Sherman, among others. All the chimpanzees were trained on paired associations between objects or events and lexigrams, which might be described by a human as arbitrary symbols on a keyboard, though a central question in the above articles is whether they were in fact symbolic for the chimps. Fig. 2.6, for instance, is the lexigram representing Sue Savage-Rumbaugh. Because they count as stimuli, I will use `typewriter style` to distinguish lexigrams from their referents.

While Lana's training focused on these paired associations, Sherman and Austin's training was different from hers in two important ways. In Savage-Rumbaugh et al. (1978), Sherman and Austin were trained on communicative tasks with one ape using lexigrams to inform the other of the

[12] I should hope that an optimistic interpretation makes room for the occasional animal genius. One such oddity is the ability of a parrot, Alex, to make sophisticated higher-order judgements, not only about sameness and difference, but also whether it is the colour or shape of a set of objects that is same or different (Pepperberg, 2000).

Figure 2.6: Lexigram representing Sue Savage-Rumbaugh (source: http://en.wikipedia.org/wiki/Sue_Savage-Rumbaugh).

contents of a food-baited closed container, while the second chimp then used the same lexigrams to request the appropriate food from the experimenter, who opened the container and shared out the food if the second chimp was correct.

In Savage-Rumbaugh and Rumbaugh (1978), Sherman and Austin were trained to concatenate lexigrams (including `give`, `pour`, `banana` and `juice`) into verb-object pairs: in training, `give` preceded solid foods and `pour` preceded liquids. However, they were not reliably able to use `give` to request a banana or `pour` to request juice. Despite their training they would sometimes type incorrect strings such as `pour banana` or `give banana juice`, or would just use a previously trained string regardless of the current situation.

They were then trained on particular associations and cross-associations: they were allowed to type `give banana`, `give juice`, `pour banana` and `pour juice`, but only the semantically coherent pairs (e.g. `pour juice`) were rewarded. The others (e.g. `pour banana`) were *explicitly* not rewarded: the foods were in different dispensers and if they typed an incorrect string, a buzzer drew their attention to the appropriate dispenser so they could see it operating but failing to produce the incorrectly requested item. That is, the lack of a reward was made *salient* to them.

After this explicit negative cross-associative training, the chimps were then able to use these pairs correctly, disregarding the other possible ordered pairs. They were also able to produce correct strings to request novel items, whose lexigrams they were familiar with, but which were not used in the above training. While the initial training took a long time, once they had become successful at these tasks, they quickly responded to these novel items

correctly.

The difference this made to the symbolic abilities of Sherman and Austin (as opposed to Lana, who hadn't undergone this training) was demonstrated in Savage-Rumbaugh et al. (1980). All three chimpanzees were trained to put three food items into one box and three tools (i.e. inedible things) into another. Lana was then able to generalise by putting novel food items and tools into the appropriate boxes but she failed at a further task where Sherman and Austin succeeded. The three chimps were trained to press lexigrams `food` and `tool` in response to the original sets of three items and all three succeeded. But while Sherman and Austin were then quickly able to generalise these lexigrams to novel tools and food items, Lana was not.

Deacon's interpretation of these successes and failures (which he admits is speculative) is as follows. He argues that Sherman's and Austin's training with verb-object pairs of lexigrams caused them to reorganise their representations of the events. The explicit positive and negative feedback in associational and cross-associational training led them to discover relationships between representations (call these R-R relationships) in addition to the previously trained associations between representations and the world (R-W relationships). Subsequently, he claims that these R-R relationships came to dominate the R-W associations from earlier training[13]. Once R-R relationships dominate (i.e. once the chimpanzees' understanding of events and lexigrams are recoded in R-R rather than R-W terms), Deacon claims Sherman and Austin are using the lexigrams symbolically. He argues it is because they have crossed the symbolic threshold that they are able to succeed where Lana failed at the novel `food` vs `tool` task described above.

He points out that this restructuring is not due to anything directly perceivable; nor is it explicitly learned. Rather, he says, it is discovered in a moment of insight. He mentions anecdotal evidence from Köhler (1927) of insight problem solving in chimpanzees. For instance, Köhler describes a chimpanzee trying to get at a banana suspended from the ceiling. A number of boxes were available, and the chimpanzee tried stacking them at random places in the room and climbing on them, but apparently didn't see how the boxes would help it achieve its goal. At a later stage, however, an R-R connection was formed (consistent with much insight research, for

[13]He suggests that this step happens because it offers an effective and efficient structure for further learning since it saves effort and storage capacity.

which see §3.5.2) between the chimpanzee's representations of the boxes and its banana-retrieving goal, at which point it stacked the boxes up in the appropriate place, climbed them, and retrieved its reward.

Returning to the features that prompted this detour (higher-order relational cognition and context), Deacon describes Sherman and Austin's formation of R-R relationships as involving higher-order inference: they solved the problem by forming an explicit connection between representation GIVE and representation BANANA and by disassociating GIVE from JUICE. This feature makes their behaviour similar to the higher-order S/D task described above. Penn et al. (2008) explicitly frame the difference between human and animal cognition in terms of such higher-order R-R relationships.

But another feature underlying animal success in inference is the extent to which experimenters constrain the context in training or make relevant information salient. In Savage-Rumbaugh and Rumbaugh (1978), for instance, a buzzer was used to draw the chimps' attention to a dispenser failing to produce an item requested with an incorrect string. The negative effects of their actions were thus made salient to them. Further, the lexigrams were positioned on a 3×8 matrix, but only the subset of these 24 relevant to a particular task were backlit in training. Indeed, in earlier training of the R-W relationships, just one lexigram was backlit for a round of training, then two, then three, etc. The experimenters note '[t]he chimpanzees do not necessarily attend to the dimensions of the environment which the experimenter deems salient' (1978, 289), so this constraint plays an important role in the chimpanzee symbol learning, whereas I argued that human symbol learning requires salience-deciding inference (§1.5.3).

Inference can involve higher-order relationships, but just which relationships turn out to be relevant (i.e. the context) is something that can be constrained or made salient by experimenters or by evolution. Sherman and Austin were unable to solve the R-R task until the relevant relationships were made salient (or the context was constrained) by experimenters. On the other hand, evolution made KIN part of the relevant context of LEOPARD for vervet monkeys. While the pitch of a male toad's call is salient to a female toad, the pitch of a leopard's growl isn't salient to a vervet monkey. In other words, evolution has solved the problem of relevance in vervet alarm calls. Penn et al. (2008) similarly emphasise that evolution constrains salience. While it seems that corvids such as rooks are capable of flexible

behaviour in novel situations, they argue that:

> Rooks, like other nonhuman animals, appear to solve tool-use problems based on evolved, domain-specific expectations about what perceptual features are likely to be most salient in a given context and a general ability to reason about the causal relation between observable contingencies in a flexible, goal-directed but task-specific fashion. (Penn et al., 2008, 119)

In situations such as symbolic communication, on the other hand, the relevant context is often left comparatively unconstrained or lacking in salience. Since I argued that unconstrained contexts are typical of pragmatic inference in §1.5.2, and given that context is built into my definition of inference above, I propose that minimal inference involves highly constrained contexts while more complex forms of inference involve comparatively unconstrained contexts. As mentioned in §2.1, I am not arguing for a saltationist view, though, so I assume there is a continuum between highly constrained and comparatively unconstrained inferences (fig. 2.7).

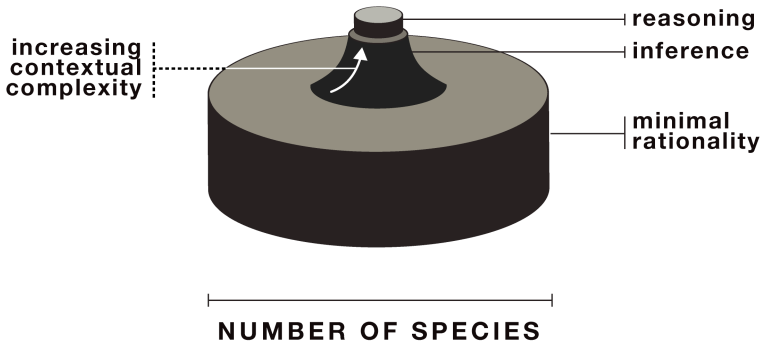


Figure 2.7: The inferential hierarchy updated to reflect the fact that comparatively more species are capable of contextually simple inference and very few capable of contextually complex inference (perhaps, just our pre-linguistic ancestors). Compared to minimal rationality, very few species are capable of inference at all.

2.5.2 Contextually unconstrained inference

This is going to be a difficult section: it's not a question that anybody has addressed directly with regard to the evolution of cognition, as far as I can tell. I do provide empirical support for the basic claim here in chs. 5 and 6, but the discussion for now will be more a matter of sketching out the corners of the ballpark. First, I will compare our starting point (TI) with our end-point (pragmatic inference) to spell out how context is constrained in the former and unconstrained in the latter (§2.5.2.1). Secondly, I will support my claims by looking at a range of experimental designs and animal behaviours, focusing on the extent to which constraint makes a problem easier and lack of constraint makes it harder, suggesting that the latter inferences are more complex (§2.5.2.2).

However, since I have admitted that this is why I set up the definition of inference in the first place (though I did provide support for it at the time), I have to provide independent reasons for making such a claim. So the third tactic will be to look (briefly) at the frame problem, which underscores the computational problem of unconstrained contexts, but which hasn't (again, as far as I know) been used to distinguish between stages in the evolution of inference (§2.5.2.3). Just to repeat: I claimed in ch. 1 that context, salience and relevance are all interrelated, so I will occasional switch from one to the other.

2.5.2.1 Starting points and end points: transitive and pragmatic inference

In the TI examples discussed above, choosing non-terminal elements (`b`, `c` and `d`) at test depended on contextual information, but this contextual constraint was provided by the training: `b` was only paired with `a` and `c`. Choosing `b` over `d` at test involves expanding this context by one step, determined at training by presentation of `c` with `d`. So context is limited in a number of ways here:

1. Size of series: 5
2. Maximum distance between non-terminal elements: 3
3. Maximum elements associated with each element in training: 2

4. Increase in context size by advancing one algorithmic step: 1

Regardless which of these is the best measure of context size, they are all quite small, so this is quite simple contextually. I would like to focus briefly on (4), though. In order to provide a clear comparison with pragmatic inference, let's treat `b` and `d` like utterances in a dialogue, as between Peter and Mary in ch. 1. Peter says `b` and then wants to understand how Mary's response `d` is relevant to `b`. According to Relevance Theory, the interpretive module (IM) first derives implications from the initial utterance `b`, which would load two assumptions (A+B− and B+C−) into its memory.

The next step is for the IM to retrieve assumptions related to either A or C, but just which one depends is retrieved first depends on accessibility. I have no idea just how that would be evaluated in this case, but let's say that whichever was seen most recently with `b` in training is the most accessible, since they would both have been seen an equal number of times. So it could be either, but let's imagine A turns out to be more accessible. If the IM unpacks assumptions connected to A, it finds nothing of relevance since stimulus `a` is terminal and representation A only includes redundant information A+B−, which is already in working memory.

This subroutine terminates at this point and returns to the previous step to retrieve assumptions from the only other alternative: C. Unpacking this involves retrieving whichever is most accessible of the two assumptions associated with C: B+C− or C+D−. The former is redundant and does not count as a cognitive benefit; the latter achieves relevance, and the process terminates[14].

This highlights two important differences between transitive and pragmatic inference. First, the TI context only increases linearly: at each step, the IM has two contextual assumptions that could potentially be added to its memory, but one of these is redundant, so only one new assumption is added to the IM's memory. The process repeats, adding one new assumption at each step, until it achieves relevance or reaches a terminal element.

To compare this with pragmatic inference, assume for the sake of simplicity that, in an imaginary language, each retrieved assumption uniformly

[14]I am not suggesting that this is how anyone solves a TI task, but I think one positive feature is that this doesn't require a single representation of the entire series `a>b>c>d>e`. I was never particularly attached to that aspect of the cognitive account anyway: my definition of inference did not require such a unified representation for a cognitive account to be true.

allows access to α further non-redundant assumptions[15]. So for instance, if $\alpha = 3$, assumptions associated with B include B+C′−, B+C″− and B+C‴−; assumptions for C′ include C′+D′−, C′+D″− and C′+D‴− (as well as redundant B+C′−). Then, if s represents the number of algorithmic steps taken and C the amount by which the context size increases at time step s, then $C = \alpha^s$. The TI context increases linearly, then, because in that case $\alpha = 1$, while pragmatic inference involves an exponential expansion, which is clearly more complex, and in the long-run may even be computationally intractable (Blokpoel et al., 2011).

The second important difference is that TI has a stopping mechanism *apart from* the achievement of relevance: terminal elements. If reached, these will terminate any subsearch, returning to the previous step to search for other, less accessible assumptions. In the TI example above, this happened when it turned out that A didn't achieve relevance. It thus does not suffer from Fodor's Hamlet problem (§1.5.2.2). Further, because of these terminal elements, $s \leq 3$, which could potentially keep even an exponentially increasing series constrained, but no such check exists in pragmatic inference.

True, C is the *maximum* size of the increase in context. If the interpreter's representations happened to be arranged such that, at each step, the most accessible assumption was the one guaranteed to lead towards relevance, then the problem becomes greatly simplified. However, this is not guaranteed, and even if it were, the IM as described by Sperber and Wilson (1995) wouldn't know this ahead of time: it is the nature of pragmatic inference to be context-deciding rather than context-constrained. Anyway, I review evidence in §2.6 below and present experimental evidence in ch. 6 showing that there is in fact a cognitive difference between situations where the relevant interpretation just happens to be the most accessible, and those where relevance is less predictable.

So, though I have shown that inference is needed for symbol evolution and that some animals are capable of transitive inference, we cannot simply claim that some animals are capable of inference and leave it there. My point here is that the degree of contextual constraint is a major difference

[15]This is meant to parallel how OSSOBUCO above could lead to retrieval of assumptions such as OSSOBUCO IS FROM MILAN, OSSOBUCO IS MADE FROM VEAL, OSSOBUCO IS A MAIN-COURSE (ch. 1 fig. 1.9). If these are the only three assumptions, then $\alpha = 3$.

between simpler and more complex forms of inference, in addition to higher-order relational cognition. For the remainder of this section, then, I will look in more detail at just how and to what degree some other animal and human inferences are constrained in this way (§2.5.2.2), and whether a lack of such constraint makes the problem more complex computationally (§2.5.2.3).

2.5.2.2 Context constrained by experimental design

In the first-order S/D tasks, subjects are required to evaluate relationships between one sample and just two visually distinct possible matches, one of which was perceptually identical to the target and the other perceptually quite different. As with TI, then, experimental design is one way of limiting the context. In this case it not only constrains the context to two alternatives, but also makes the relevant features perceptually salient.

Compare the S/D task with a more open-ended similarity task. Would you say that George Clooney and Chuck Norris are similar? The answer, surely, is, 'Well, sort of, I guess. It depends, but not really.' If the context includes Alice Walker and Toni Morrison, then I imagine you would consider Clooney and Norris quite similar in that they are both white male actors. On the other hand, if the context was Jean-Claude Van Damme and Steven Seagal, you might be less inclined to consider the first pair similar (they are all white male actors, but Clooney is the only one who doesn't star in martial arts films).

It probably wouldn't occur to you, unless you were forced to sit down and write extensive lists, that AMERICAN is not a relevant feature (they are all American) or that INTELLECTUAL could be: I think I would be more likely to represent award-winning writers as being intellectual, compared to martial-arts film stars. In the TI case, there was only one relevant feature: the REWARDED-MORE-THAN property. In the S/D task, there was only the question of perceptual similarity. But here, just which feature counts as relevant is itself context-dependant. This detour is intended to illustrate the fact that context size is not merely a matter of counting entities: just which features count as relevant is left rather open-ended here, while experimental design tends to make these salient to some degree.

Back to the question of experimental design, then. Savage-Rumbaugh and Rumbaugh (1978) used a buzzer to draw chimpanzees' attention to a vending machine failing to produce a reward. They also backlit a small sub-

set of possible lexigrams so that specific pairs of just four keys (`give`, `pour`, `banana` and `juice`) could be trained. In semiotic terms, both representamens and objects were made salient by the experiment design. That the chimpanzees managed to form higher-order relations between the relevant representations is impressive, but just which representations were the relevant ones was decided and indicated by the experiment design. In this case, the problem turned out to be unsolvable by chimps without this limitation, so contextual constraint may underlie those few cases were animals achieve higher-order relational cognition.

Moving to a non-linguistic context, Roberts et al. (2000) manipulate visual salience in tests of inference. Their central question is the extent to which inferential processes are domain-free or domain-specific, but their experiment involves comparing success at inference problems that are logically equivalent, but couched in differing visual stimuli. I will assume here that lower success rates are an indicator of inferential complexity.

Their experiment is a response to Richardson (1991), who compared low rates of success at comparatively abstract tasks called Raven's Progressive Matrices (fig. 2.8) with higher rates of success at less abstract, socio-cognitive[16] versions of the same tasks (fig. 2.9). Richardson (1991) had found that children performed better at the socio-cognitive tasks, despite their logical equivalence to the abstract original. He concluded, therefore, that the children had access to domain-specific schema-supported forms of inference in the socio-cognitive tests, while these were lacking in the abstract ones.

Roberts et al. (2000) argued, however, that Richardson's tests are not simply domain-free and domain-specific versions of the same task. Rather, Richardson's socio-cognitive condition made certain features of the task more salient in two ways: it manipulated visual salience and provided verbal direction of attention. The instructions in the socio-cognitive tasks, for instance, involved a commentary that framed the pictures with a recognised schema such as, in this case, people departing. The original pictures, on the other hand, involved novel figures superimposed on one another, so that it was not necessarily clear just what counted as an element. Further, no identifying schema was prompted by the instructions in this condition.

Consequently, Roberts et al. (2000) replicated these tests with a further variation: an abstract task, but with a particular feature made salient

[16] 'Socio-cognitive' is Richardson's term for the more real-life example in fig. 2.9.

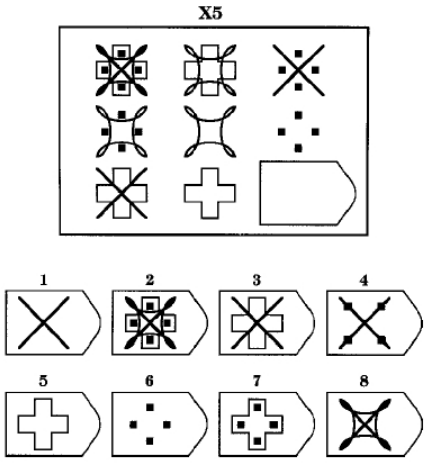


Figure 2.8: Raven's task E5. The 3-by-3 matrix at the top is the test: fill in the lower right gap; the eight options below are possible answers. In this case, the third element of a row or column is what remains after the features in the second element are subtracted from those in the first so the answer is 1 (Roberts et al., 2000).

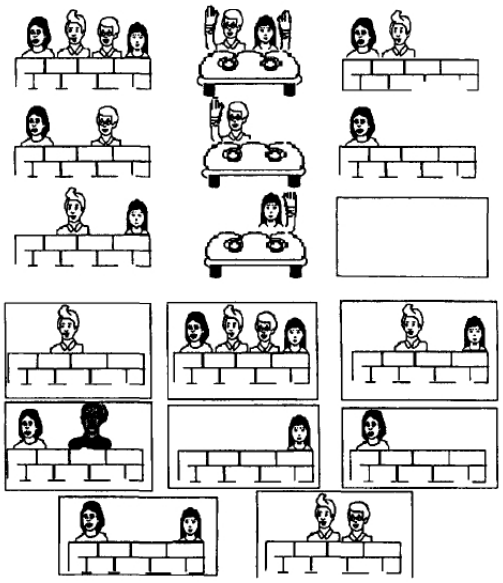


Figure 2.9: Socio-cognitive adaptation of Raven's task E5: the problem here is framed in terms of people leaving and others remaining behind (Richardson, 1991).

(fig. 2.10). They found that children did as well at the salient abstract tasks as they did at the socio-cognitive one, meaning that the results in Richardson (1991) were not due to domain-specific inference. More importantly for my purposes, their experiment demonstrates that manipulating visual salience in experimental design (making it clearer which features are relevant) has an effect on the complexity of the inference. All these tasks involve higher-order relational cognition, so that cannot be the sole criterion for inferential complexity.

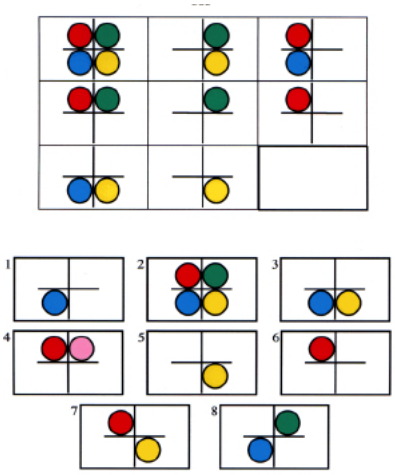


Figure 2.10: Abstract, salient version of task E5: the only difference between elements is colour (Roberts et al., 2000).

Perceptual salience is not the only common way in which experiment design constrains context, though. The following example does not address the question of how complex an inferential problem is, but demonstrates a way of constraining context non-perceptualy.

Lee and Holyoak (2008) investigated the relationship between causal inference and analogy by presenting subjects with a source (a description of an animal with three properties and a possible effect caused by or prevented by those properties) and a target (a description of another animal sharing all or some of the same properties). The experiment tested subjects' ability to make causal inferences about the target based on analogies between source and target. Unlike the open-ended Clooney-Norris example, here the relevant features involved in cause and effect relationships are explicitly given.

A source animal might be described as having enzyme aliesterase, neurotransmitter tyrosine, hormone TSH and exceptional hearing. The participants are also told that aliesterase and tyrosine cause exceptional hearing while TSH prevents it. A target animal might be described as having aliesterase and TSH, leaving subjects to infer how likely it is that the target has exceptional hearing. The analogy part of this experiment is a bit like asking, 'To what extent are George Clooney and Chuck Norris similar, in as far as they are both white, male, actors?' The subsequent, causal-inference part of the experiment is thus limited to a small set of features made relevant by the experimental design. This experiment is typical of a large family of what might be called inductive accounts, and other examples will be explored in ch. 3 where I highlight the distinction between unconstrained hypothesis generation and hypothesis evaluation.

But to address the question of complexity and to return to language, let's consider claims by Medina et al. (2011), who contrast contexts for realistic language learning with approaches typical of inductive experiments (fig. 2.11 *a* and *b*, respectively). They highlight the fact that realistic language learning involves contextual uncertainty while experimental approaches often involve small contexts with objects salient against neutral backgrounds (the shoe in fig. 2.11 *b* is hardly visible in *a*: it is directly in front of the baby). They stress that larger contexts (or hypothesis spaces) place immense burdens on cognition and show experimentally that subjects are less successful at word learning in larger contexts, noting that statistical models tend to be successful only because they involve artificially constrained contexts. This supports my central claim here that context is a major factor in evaluating inferential complexity.

Medina et al. speculate that 'learners may have implicit means for distinguishing between more and less useful contexts, discarding some input without its entering into the search for meaning' (Medina et al., 2011, 2). That is, context-deciding inference may play a role in generating a reduced hypothesis space before statistical hypothesis evaluation mechanisms can function effectively, which again turns our focus to hypothesis generation. I offer a concrete proposal concerning this speculation in ch. 3.

Figure 2.11: *a* — realistic language learning task; *b* — typical inductive language learning task (Medina et al., 2011).

2.5.2.3 Computation

Since I purposefully built context into my definition of inference, I should also provide some independent justification for the claim that context matters. It turns out it matters hugely to computational theories of mind. The Frame Problem (FP) is actually a set of problems with, at best, family resemblance. It originated as a problem in AI but has since expanded to include various philosophical and cognitive mysteries. This is not the place to give a history or full description, so just a rough outline will have to do.

The original FP was a computational problem: how to formally specify what is and what is not affected by a change to some part of a description of a system of entities and events (McCarthy and Hayes, 1969). One might assume that moving an object does not change its colour, but as soon as one attempts to formalise this assumption, one runs into the problem of having to formalise all assumptions about the various features not changed by movement. If there are many assumptions, or many features that might

change, this quickly becomes a difficult computation. And what happens if the object is moved into a bucket of paint? Exceptions to the assumptions would then have to be formally represented, adding up to an awful lot of representation. The original problem is a matter of how such assumptions are represented and what formal logics handle them, then.

But the FP was taken over by philosophers as indicative of deeper issues in cognition more generally. Dennett (1978), for instance, phrased it in terms of how an organism comes to update (some) beliefs it has about the world so that its representations remain accurate descriptions of the world. That is, how are these processes truth-preserving? Or what degree of belief in something would be optimally rational? In particular, how does cognition determine which beliefs should be updated when new information is learned, without having to check the status of all beliefs in memory? This is where we run into the Hamlet problem (when to stop thinking, Fodor, 1987). It is also where we run into features invoked in attempts to avoid or simplify the FP. These include relevance (Shanahan, 2009; Xu and Wang, 2012). If the system included specifications of just which beliefs are relevant to an update, then the FP would go away. But this would provoke a regress: how would the system know what's relevant? Relevance turns out to be context-dependant (Shanahan, 2009).

My central claim in this section is that degree of contextual constraint distinguishes minimal inference from more complex sorts, such as the pragmatic kind needed to cross the symbolic threshold. The central claim of *this* subsection, then, is simple: the fact that contextually unconstrained inference potentially engages the FP means that it is more complex than minimal inference. Chiappe and Kukla (1996) apply these Fodorian worries to pragmatic inference, arguing that it runs into the FP because it is context-deciding. So a secondary point in this subsection is to explore the source of the disagreement, and suggest a particular kind of compromise between these Fodorian worries and Sperber and Wilson's version of pragmatic inference.

Recall the distinction between a rationalist and empiricist psychology (§2.2.3, §2.3). A rationalist psychology maintains that cognition is normative (either truth-preserving, or at least able to estimate what degree of belief in a proposition would be optimally rational) and syntactic (sensitive to formal or logical features of a representation, not its content). An

empiricist psychology maintains that cognition is associationist, but makes no assumption of normativity. The problem arises because Sperber and Wilson's description of pragmatic inference, if taken as an algorithmic-level account, blends features from both a rationalist and empiricist psychology in a particular way that engages the Frame Problem. I mentioned the possibility of dual-process models (§2.2.3), but the intention was to allow that both processes operate in a complementary fashion, not that a single hybrid process combine aspects of both. The solution I'll end up proposing is that Sperber and Wilson should abandon the idea that pragmatic inference is deductive.

The FP is a matter for a rationalist psychology because it concerns normativity. I grant that many cognitive processes are normative, but Sperber and Wilson (1996) explicitly state that pragmatic inference is not one of them: its conclusions are not justified or valid, but must rather be judged by their success or efficiency. Accessibility plays an important role in pragmatic inference: the more accessible an assumption, the sooner it is retrieved. Relevance is not directly represented in their account: it is the outcome of the interpretive process, not a value to be computed by that process. Relevance in their terms involves maximising cognitive benefit while minimising cost. Accessibility is a matter of processing cost, which means it is built into the foundations of relevance. Accessibility is a matter of an empiricist psychology, so the same must be true of relevance in so far as it rests on accessibility. Though Fodor considers it a major failing to have recourse to an empiricist psychology, I gave reasons why most of his objections do not extend to minimal rationality and thus to inference in my terms (§2.3, see also ch. 3 below). It would seem that there is no problem, then: claiming that relevance theory runs afoul of the FP simply misunderstands the nature of relevance theory (Sperber and Wilson, 1996).

However, there is a hole in this defence of pragmatic inference. Sperber and Wilson allow it to be non-deductive overall, and do mention that there are some mysterious, creative kinds of inference out there, but they still want the core of interpretation to be deductive. Deduction is a prototypically syntactic process. While they reject the normative assumption of a rationalist psychology, they thus retain the syntactic assumption. Depending on just how much explanatory burden is shouldered by deduction, this may still allow the FP to sneak in. My criticism of the ossobuco example in

§1.5.2.2 centred on just this point. Their description of *overall* relevance can decide whether an utterance was going to be attended to or not (i.e. whether cognitive resources were going to be spent on interpreting an utterance or not), so it can control whether interpretation as a whole proceeds or stops. But then they shift to deduction and describe the actual process of retrieving assumptions as happening in a formally rigorous, sequential, mechanical fashion. Allott (2013) highlights the mechanical nature of hypothesis generation in their account and stresses that they assume a CTM. In §1.5.2.2 I pointed out that this is what ran into the Hamlet problem: it couldn't control the algorithmic flow of that process, and thus couldn't explain why exploration of assumptions associated with MILAN could ever be abandoned to return to less accessibly assumptions associated with OSSOBUCO.

If deduction bears any explanatory weight in pragmatic inference at the algorithmic level (that is, if the deductive process is what explains how Peter ended up understanding Mary) then pragmatic inference still succumbs to the FP. If on the other hand, Sperber and Wilson accept that their deductive account is a computational-level, simplified model of the process and that there must be some non-deductive inference at the algorithmic level[17], then pragmatic inference can avoid the FP. I'll go on to argue that this inferential process should take seriously the commitment to an empiricist psychology brought in by basing relevance on accessibility. In other words, rather than a non-normative yet syntactic process, I'll argue that relevance is handled by a non-normative associationist process, and thus an empiricist psychology. This will make sense of a number of terms introduced previously, such as insight or creativity.

As previously suggested, Sperber and Wilson's tactic of claiming that our interlocutors are helpful and that they try to be relevant *may* work to avoid the FP in certain cases where context is artificially constrained and may occasionally work by chance in unconstrained contexts, but this won't help at the symbolic threshold, where all attempts to direct attention (or make features of the object salient, or otherwise constrain context) are not conventional and are thus themselves matters to be inferred. Sperber and Wilson (2002) eventually abandon the idea of a global (i.e. non-modular)

[17]They make vague allowances for the existence of these in general, but manage to avoid them entirely in all worked examples of interpretation. I'm simply saying that non-deductive inference should be in the spotlight, or at least share the spotlight, particularly at the symbolic threshold.

interpretive process (which it had been in Sperber and Wilson, 1995) and propose a submodule within a Theory of Mind module that carries out pragmatic inference. This move was partly motivated by the above Fodorian problems since modules are one way of constricting context (Fodor, 2001), but again this won't help with evolution. A module, the purpose of which is to interpret language, cannot explain how a pre-linguistic species managed to infer the meaning of the first symbols, given that these required pragmatic inference (§1.5.2.3).

Fodor (2001) is sceptical about the whole idea that appealing to modules can solve the FP because some processes are inherently global. His tactic is to say that such contextually unconstrained problems simply cannot be computed, that they are thus not explicable by a CTM, and that he doesn't know what to do about all this since he doesn't like the alternative (an empiricist psychology). Which at least is honest. Anyhow, my response to the disagreement between Fodor and Sperber and Wilson has suggested that pragmatic inference should be rendered an empiricist process, though other cognitive processes can still be rationalist. Minimal rationality is empiricist, reasoning is rationalist (probably, but reasoning falls outside of the scope of this dissertation), but inference is mixed. Just how this mixed approach is supposed to work depends on the relationship between abduction and induction in the following chapter.

For now, though, I should point out that such a mixed design is not unusual for those cognitive scientists interested in the origins of human intelligence or in comparisons between human and non-human intelligence. Harnad (1990) argues that it is needed to solve the symbol grounding problem[18]. Penn et al. (2008) argue that a purely connectionist account cannot explain humans' advanced higher-order relational cognition, but that a purely syntactic account cannot account for animals' failure to make systematic inferences, and that a mixed system is thus needed[19].

The main conclusion here was that a distinction between contextually constrained and unconstrained inference is justified by a consideration of the Frame Problem. A secondary conclusion was that, regardless of whether one

[18]This is a matter of private, not public symbols, though, which is why I haven't explored it here.

[19]Though not equivalent to 'rationalist' and 'empiricist', their terms 'computational' and 'connectionist' split the playing field along similar lines; what's missing is explicit reference to normativity or the lack thereof.

leans more towards Fodor or to Sperber and Wilson, since hypothesis generation in pragmatic inference (and thus symbol origins) is context-deciding, it seems to be either very, very difficult computationally or, instead of being syntatically computational at all, could be a matter for an empiricist psychology, more concerned with causal relationships between representations. Sperber and Wilson claim that they escape the FP because relevance is not computed, that accessibility plays a major role, and that interpretation is not normative, so it seems self-defeating to nonetheless insist on deduction. The next chapter looks more closely at why some context-deciding inference needn't assume normativity; the following section looks at evidence for context-deciding processes in human brains.

2.6 Neurological and behavioural evidence

Questions of computation can be quite abstract, whereas my central question needs, at least potentially, to be relatable to the evolution of actual brains. This section thus reviews evidence showing that there are observable behavioural and neurological differences in how humans process constrained and unconstained inferences, especially in pragmatic interpretation. The upshot is that humans have two basic interpretational processes, one dominating in context-constrained cases and the other in context-deciding cases. Both operate in parallel, and the context-constrained case shares some similarities with Sperber and Wilson's deductive approach, while the other is what will allow for insight and creativity.

The experiments discused below are principally concerned with hemispheric differences in processing meaning (not inference generally) and these differences are describable in a number of ways, including distinctions between fine and coarse coding; inference at the sentence and discourse level; context evaluation; figurative language; and graded salience. I claim that these are related to context size, and I support this with empirical evidence in part II. Most of what follows assumes some kind of spreading activation between contentful representations, which is one reason I highlighted the role of an empiricist psychology in the previous section.

2.6.1 Fine vs. coarse coding

I begin with the distinction between coarse and fine coding which underlies much of what follows. Simplifying somewhat, when processing words, the left hemisphere (LH) shows rapid and focused activation of a very narrow range of associated representations or meanings while the right hemisphere (RH) involves a slower, weaker, broader and more diffuse spread of activation (Beeman et al., 1994; Bowden and Jung-Beeman, 1998; Jung-Beeman, 2005). Take the ambiguous word 'bank', where the dominant meaning is a financial institution and the subordinate meaning is the side of a river. If a subject is primed with such a word, spreading activation will facilitate quicker or more accurate processing of associated target stimuli `money` and `river` than unrelated words, such as `farm`. Typically this facilitation is tested behaviourally by lexical decision tasks (where subjects press different keys to indicate whether a stimulus is a word in their language or not); go/no-go tasks (where subjects press a key if some criteria is met, and do nothing otherwise); or naming tasks (where subjects simply read the word displayed). It is also investigated neurologically with various methods of measuring brain activity.

Differing degrees of priming facilitation show that LH processes are more likely to involve sustained activation for the dominant meaning, while RH processes involve more sustained activation of both dominant and subordinate meanings (Jung-Beeman, 2005). However, talk of dominant and subordinate meanings suggests a clear-cut dichotomy, while the matter is actually more graded, so it is better to talk of comparatively narrow/fine vs. comparatively broad/coarse semantic fields, where 'narrow' suggests not only a smaller number of associated activations, but also activation of more closely related representations; 'broad' suggests activation of a larger number of more distantly related representations. So LH processing of words is typically a matter of fine coding while RH processing is typically a matter of coarse coding (Jung-Beeman, 2005).

The relevance of this fine/coarse distinction to the current discussion about contextual constraint is as follows. The LH usually dominates in processing linguistic meaning, but its fine coding means that the LH is better at processing representations with highly constrained relationships. When the relationships between representations are less constrained, the RH is more likely to shoulder an increased burden, comparatively. A loosening

of contextual constraint should thus lead to observation of increasing levels of RH involvement. This jump from 'fine vs. coarse coding' to 'contextually constrained vs. unconstrained' needs a little unpacking, though.

2.6.2 Semantic coding and inferential context

There are two main ways in which fine coding can be related to constrained contexts, and coarse coding to unconstrained contexts.

(1) Fine coding in the LH provides a highly constrained inferential context in that limited spreading activation in response to a word constrains the set of candidate representations that might be involved in inferences about the meaning of subsequent words. Coarse coding provides a comparatively unconstrained inferential context in that activation spreads to a much wider range of representations. There are at least two ways of construing this: for a fine context, we could claim that only a small set of representations are available (or, figuratively, 'visible') to LH processes and that other representations are unavailable or invisible. Alternatively, we could claim that the LH dominates in situations where only a small set of information is predicted to be very relevant. The corresponding construals for the RH are that either a wide range of representations are available for or are visible to the inferential processes, or that the RH dominates when there is not much information about which representations are predicted to be relevant. The RH is 'less discriminant' (Faust and Kahana, 2002, 893).

(2) We could distinguish degrees of inferential contextual constraint in terms of the scale of the structure to be interpreted. That is, interpreting a word involves a smaller-scale context than a sentence, which involves a smaller-scale context than a group of sentences, in turn smaller than an entire narrative structure. The larger the structure, the more information one must sift through in order to find what is relevant. The relevance of the fine/coarse distinction is this: the larger the structure, the less likely it is that the LH's small set of closely related representations is sufficient for providing coherence, and the more likely that distantly related representations must be integrated in the RH.

We thus have the following, related points in need of empirical support. I examine these in turn below.

1. (a) A small set of representations visible to LH inference; a wide

range visible to RH inference

(b) LH performs better where small set of highly predicted representations are sufficient for understanding; RH performs better where levels of prediction unknown or varied

2. RH increasingly important as structures grow larger

2.6.3 Hemispheric differences in inference

In support of point 1 (a) above, I examine evidence for the claim that the RH's broader semantic network succeeds better in contextually unconstrained situations than the narrow network of the LH. I begin with looking at evidence for broad coding, then turn to discuss whether the coding is particularly semantic.

Beeman (1993) compared the ability of RH-damaged (RHD) patients and normal subjects to draw causal inferences implied by contextual information in short texts. For example, given a vignette such as: 'John waded in the water, not knowing there was lots of glass nearby. Then John called for help, and the lifeguard came running,' one might link his cry to contextual information about glass and reach the coherence-creating inference that John called for help because he had cut his foot on the glass. Compared to normal subjects, RHD patients performed badly at answering questions about inferred information, though they succeeded at questions about explicit information. So RHD patients struggle with inferences that require weighing up the contribution of contextual information.

Abstracting away from the narrative setting and using only neurologically normal participants, Beeman et al. (1994) showed subjects sets of three priming stimuli followed by a laterally presented target word that the subjects had to name. Summation primes were made up of three words only distantly related to the target: 'foot/cry/glass' were distantly related to 'cut', as per the above vignette. Direct primes (e.g. 'none/scissors/whether') were made up of one word ('scissors') directly related to the target ('cut'), surrounded by two unrelated words. After direct priming, subjects responded better to targets presented to the right visual field/left hemisphere (RVF/LH), while left visual field/right hemisphere (LVF/RH) targets were facilitated by summation priming.

Their interpretation of all this is shown in fig. 2.12. LH activation of

FOOT only spreads to very closely associated representations TOES, SOCK, HEEL, while RH activation of FOOT also includes 12 INCHES and PAY (presumably in the context 'foot the bill'). The context in the former is tightly constrained; in the latter it is broader. In the sets of primes in fig. 2.13, the constrained LH activation of the three summation primes does not activate CUT, while the overlap afforded by broader activation in the RH does, even though the individual contribution of each stimulus is very weak. These results support point 1 (a) above: the set of contextual representations available to the LH is smaller than those in the RH and small-context constrained inferences are thus cognitively distinct from larger-context unconstrained inferences.

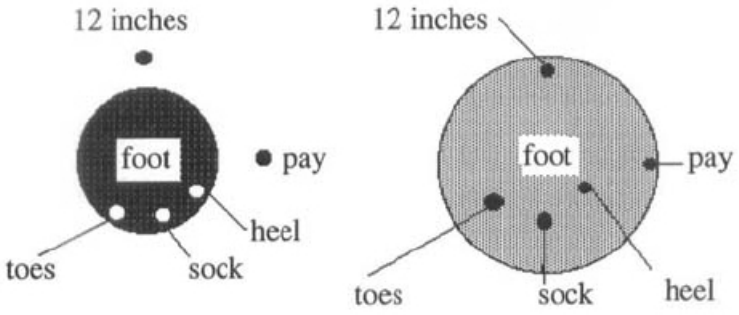


Figure 2.12: Small, focused semantic field activated by FOOT in the LH and broad, diffuse field in the RH (Beeman et al., 1994).

Jung-Beeman (2005) distinguishes various areas within each hemisphere that contribute to these results[20]. He identifies three main processes and brain areas implicated in those processes (fig. 2.14). Semantic activation is when activation of one representation spreads to related representations (as in fig. 2.12), principally associated with the posterior Middle/Superior Temporal Gyri (pMTG/pSTG). Semantic integration is when multiple such activations converge on a potential target (as in fig. 2.13), mainly in the anterior Middle/Superior Temporal Gyri (aMTG/aSTG). So semantic activation focuses on the word level, while semantic integration focuses on the message level. Semantic selection chooses among competing activations, bringing one

[20]I only mention this because some of these regions will turn out to be important in what follows in this section and in the next chapter. You don't need to buy his story completely in order to appreciate later evidence that focuses on these regions.

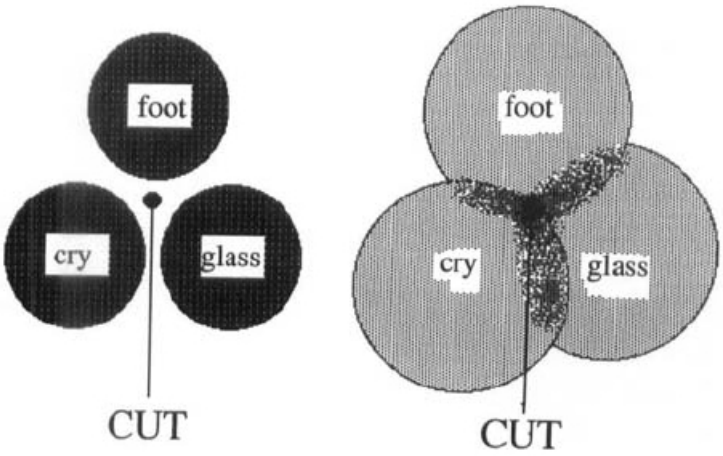


Figure 2.13: Non-overlapping fields in the LH and overlapping fields of summation primes in the RH (Beeman et al., 1994).

to consciousness or initiating behavioural output, typically associated with the Inferior Frontal Gyrus (IFG).

There are also physical-level differences underlying these differences in processing. Jung-Beeman (2005) claims that the RH has a greater proportion of white matter than the LH, promoting connections between neurons. He also points out that pyramidal neurons in the RH have longer dendrite branches than those in the LH, meaning that each neuron is connected to a broader range of inputs, allowing more overlap between connections in the RH (fig. 2.15).

I turn now to the question of whether the RH representational network differs from the left in encoding semantic information. In a split-visual-field priming task, Chiarello et al. (1990) tested two kinds of relationships: associations (one word typically elicits the other as a response, such as 'bee' and 'honey') and semantically similar pairs (both words belong to the same category, such as 'dagger' and 'rifle'). They also tested combination associated+semantic pairs ('bread' and 'butter'). They found that semantic-only pairs facilitated responses to lexical decision and naming tasks in the LVF/RH, but not in the RVF/LH. Semantic+associated primes facilitated responses in both hemispheres.

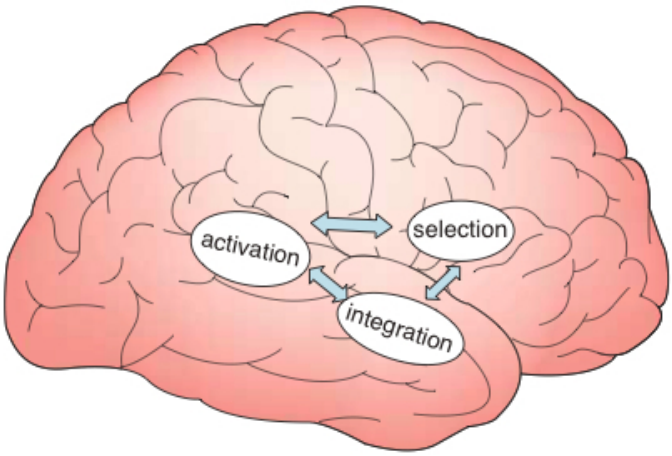


Figure 2.14: Location of brain areas identified above: in this case, the RH (Jung-Beeman, 2005).

Grose-Fifer and Deacon (2004) analysed event related potentials (ERPs), measurements of electrical activity across the scalp in relation to specific events such as stimulus presentation. A wave called the N400 (so called because it is a Negative deflection, typically peaking around 400 milliseconds after stimulus onset) is an indicator semantic relatedness: if two stimuli `s1` and `s2` are observed in that order and are semantically related, the N400 for `s2` decreases in amplitude[21]. Grose-Fifer and Deacon presented subjects with pairs of stimuli that were either unrelated, or shared relatively few features ('sofa','vase'), or shared many features ('bookcase', 'cabinet'). They found attenuated N400 signals (indicating semantic relatedness) for pairs with high feature overlap in the RH only, and no significant attenuation for low feature overlap in either hemisphere[22].

In another ERP study, Kiefer et al. (1998) found that indirectly related

[21] Differences in N400 response are also implicated in congruity in world knowledge, as in examples discussed by Menenti et al. (2009) below.

[22] For the sake of completeness, Grose-Fifer and Deacon conceive of the LH/RH difference here as follows: the LH involves local representation, by which they mean that concepts are represented as entire units; the RH involves distributed representation, by which they mean that concepts are represented as sets of features which may be activated independently. It seems that Jung-Beeman (2005) doesn't consider this entirely incompatible.

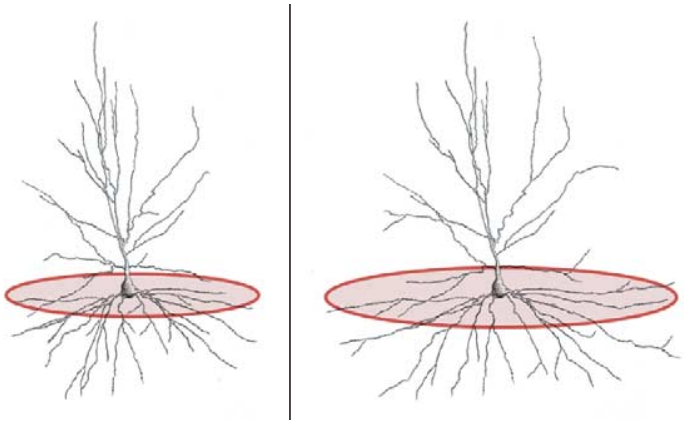


Figure 2.15: Pyramidal neurons in LH and RH language areas respectively, showing broader branching in the RH (Jung-Beeman, 2005).

words ('lemon' and 'sweet' are indirectly related through 'sour') showed reduced N400 in the RH only. Semantic activation in the RH can thus spread from one representation to its associates, and then spread from the second representation to *its* associates in turn.

So semantic information — including category membership, shared features, indirect relations and even antonyms (Beeman et al., 1994) — seems to be more typical of RH processing, while associative information seems to span both hemispheres to different degrees. Zlatev (2008) suggests that lateralisation played a role in the evolution of signification, and that the RH is principally implicated in the sign's content (as opposed to its form). The results reviewed above are compatible with that claim.

An interesting complication in this picture is creativity. Atchley et al. (1999) administered a test of verbal creativity to divide subjects into low, medium and high creativity groups. They found that low and medium creative subjects displayed the above pattern, whereby subordinate meanings are primed only in the RH, while high creatives had subordinate meanings primed in both hemispheres. One particular measure of creativity, then, involves coarse or RH-style coding extending to the LH.

2.6.4 Predictability

In support of point 1 (b) above, I examine evidence for metaphor interpretation based on what is called the Graded Salience Hypothesis (GSH, Giora, 1997). I should point out, though, that 'salience' in this body of literature means something more like 'conventionality' than how I've been using it so far, but that's fine given that I'm interested what happens before there are linguistic conventions.

According to Mashal et al. (2005), previous studies of metaphor had focused on dimensions such as literal vs. figurative meanings and had found conflicting results. Consequently, they base their study on the GSH, where the fact that a conventional metaphor is coded linguistically predicts that it is processed like a literal expression, despite involving figurative meaning. That is, a conventional metaphor such as 'bright student' has salient (strong or easily accessible) associations between the two words; the second is reasonably predictable given the first. Conversely, a non-salient or novel metaphor such as 'pearl tears' has no linguistically coded or conventional association between the words, so interpretation involves a creative process of finding novel semantic connections. Richards (1950) argues that metaphor interpretation involves finding the ground, or set of shared properties, between the topic 'student/tears' and vehicle 'bright/pearl'. His usage suggests something very much like the Peircean ground I discussed in §1.2.3.2: we must make a judgement of relevance in establishing the respect in which something is bright or like a pearl. While the ground is conventional in salient metaphors, it must be inferred for novel ones, so this is very much like the process of interpretation I argued is necessary for pragmatic inference of a novel symbol (§1.5.2.3).

Mashal et al. (2005, 2007) generated four lists of word pairs: literal ('broken vase'), conventional metaphor ('bright student'), novel metaphor ('crystal river') and unrelated ('boot laundry'). Twenty judges rated these pairs as literal, metaphorically plausible, or unrelated, and 75% agreement was needed for assignment to a category. Ten further judges rated the metaphors on a five-point familiarity scale, where those scored less than 3 (average 1.98) were considered novel and those more than 3 conventional (average 4.67). An fMRI study was conducted on still further subjects, who were shown word pairs and were asked to decide silently whether the pair was literal, metaphorical or unrelated (but not what kind of metaphor).

Mashal et al. (2005) conducted a principal components analysis to discover functional networks involved in processing stimuli, while Mashal et al. (2007) investigated regions of interest independently of such networks. Both found that the RH, particularly the temporal lobe, played a prominent role in interpreting novel metaphors, as opposed to classical LH language areas for conventional metaphors and literal expressions[23]. So contextually constrained, predictable relationships can be distinguished from less constrained, novel relationships in terms of differing hemisphere involvement.

Still on point 1 (b), Faust and Kravetz (1998) show that the LH benefits more from contextual constraint than the RH, where by 'constraint' they mean cloze probability, i.e. predictability that a word follows, based on the previous parts of the sentence. They had twenty judges complete unfinished sentences by providing three possible endings, and used these results to divide sentences into groups according to level of constraint: high, medium and low. If a possible completion was given by more than 90% of judges, for instance, this completed sentence was considered highly constrained. In a split-visual-field task, a further group of subjects was presented with incomplete sentences centrally, followed by lateral presentation of a target in a lexical decision task. Analysis of accuracy and latency by constraint group show that the LH is more sensitive to sentence constraint (read 'predictability') than the RH. While both hemispheres responded better to high constraint than medium or low, the difference between these conditions was smaller in the RH than in the LH. So the LH is more sensitive to constraint and performs best at highly constrained contexts, while the RH is less sensitive to context. But this doesn't address what the RH may be better at.

Kircher et al. (2001) conducted an fMRI experiment where subjects had to read a sentence, or choose between two potential completions of the sentence, or generate a possible completion. All sentences had low close probability. During the GENERATION condition, Kircher et al. found significantly more activity in the RH (particularly the temporal cortex, including the STG) than in the baseline READING condition, and no significant difference in the LH. They conclude,

> The prominent engagement of the right lateral temporal cortex

[23] Other implicated regions for novel metaphors include the IFG-RH and MFG-RH, cf. fig. 2.14.

> during the GENERATION conditions may reflect the processing of linguistic context, and particularly the activation of multiple meanings in the course of producing an appropriate completion. (Kircher et al., 2001, 798)

This claim about the RH processing context sounds at odds with the claim in Faust and Kravetz (1998) that the RH is less sensitive to context, but they seem to mean different things. The interpretation offered by point 1 (b) above is that Faust and Kravetz show that the LH excels at context-dependent processes, while Kircher et al. show that the RH performs more of a role in processing what the context is. Assuming that a sentence provides the context for inferences about words in that sentence, the task in Faust and Kravetz and the DECISION task in Kircher et al. involve given contexts and processes sensitive to levels of predictability within those contexts, while the GENERATION task in Kircher et al. involves hypothesising what the message or context might be in the first place.

There are two sources of support for this interpretation. Firstly, Federmeier and Kutas (1999) conduct an ERP study with manipulations of semantic relatedness and contextual constraint (again, a matter of how predictable a sentence completion was given the rest of the sentence). They conclude that:

> The left hemisphere is more sensitive to contextual constraint because constraint specifically reflects the extent to which context information allows specific predictions to be made. Because only the left hemisphere seems to be generating expectations, only its processing reflects a strong influence of constraint. In fact, the right hemisphere seems to outperform the left precisely under conditions where prediction is difficult. (Federmeier and Kutas, 1999, 387)

A second source of support comes from a divided-visual-field lexical priming experiment by Faust and Chiarello (1998). Priming sentences ending in ambiguous words such as 'second' were presented centrally, followed by lateral presentation of a target word for a lexical decision task. The sentences could be consistent with the ambiguous word's dominant meaning ('He could not wait for even a second') or subordinate meaning ('She stood in line

and was second'). The target word could relate to the dominant meaning ('time'), subordinate meaning ('number') or be unrelated ('sound').

They found RVF/LH facilitation only for a target word coherent with the priming sentence: after a dominant-consistent sentence, only the dominant target was facilitated; after a subordinate-consistent sentence, only the subordinate target was facilitated. In contrast, for the LVF/RH, both dominant and subordinate meanings were primed, regardless of sentence context. So the LH selects the most relevant meaning of the ambiguous word *within* a given context, while the RH operates *across* these context boundaries. This is consistent with the previous study, but it adds the key terms 'relevance' and 'salience':

> The unique capabilities of each hemisphere are drawn upon as needed based upon alterations in information salience and relevance that modulate the use of multiple processors distributed throughout the brain. (Faust and Chiarello, 1998, 833)

There is thus a range of support for claim 1 (b): the LH dominates in situations where meaning is predictable from context, while the RH dominates in unconstrained contexts.

2.6.5 Scale

The previous experiments examine only sentence-level effects, while the RH's ability to process information across sentence contexts suggests it may be useful in processing larger discourse structures, which moves us on to point 2. The following two experiments show firstly that coarse coding allows activation of a wide range of representations at the narrative level (Virtue et al., 2006), not just the sentence level, and secondly that this activation plays a role in RH coherence-creating inferences (Menenti et al., 2009).

Virtue et al. (2006) presented subjects with vignettes such as:

> After the rugby match, Justin's friends teased him for not knowing the rules. He gathered around his friends and joked about beating them next time. In order to look macho, Justin grabbed a beer from the cooler.

The vignettes were completed either by strongly constraining sentences ('His friends were soon covered in beer') or weakly constraining ones ('His

friends were soon cheering him on'). The former is strongly constraining in that it promotes a causal inference that strongly predicts the target word (which in this case was 'spray') while the weaker ending does not. Response to laterally presented targets was facilitated in both hemispheres by strongly constrained vignettes, while weakly constrained vignettes facilitated response to the target in the LVF/RH only. Coarse coding thus has an effect at the level of narrative structure, not just within a sentence.

The narrative level involves conceptual structures not present within any individual sentence. Integration of world knowledge to cohere with narrative-level representations may thus be distinct from integration of world knowledge to cohere with sentence-level representations. We would thus expect hemispheric processing differences when world knowledge coheres with or fails to cohere with contexts on different scales.

Contrast 'the train is sour' and 'the train is on time'. The anomaly in the first depends on linguistic meaning, while the second depends on non-linguistic information: it is untrue if you happen to know that the train is late. Menenti et al. (2009) review previous research showing that cases like the latter, involving world knowledge integration, are accompanied by increased activation in the inferior frontal gyrus (IFG, see fig. 2.14). They investigate processing of such world-knowledge anomalies in narrative context. The sentence 'the elephant flies', for instance, involves an anomaly. But when preceded by the contextual sentences 'the circus is is travelling by airplane,' or 'Dumbo is a fantasy animal,' this local context promotes an interpretation that reduces the anomalous effect, such that we would expect decreased activation in the IFG compared to a neutral context.

In an fMRI experiment, Menenti et al. showed participants vignettes consisting of four sentences each. The first sentence was invariant across condition and introduced the topic. The next two sentences either introduced a world-knowledge anomaly (local context) or not (neutral context). The fourth sentence had a critical word that differed between conditions, either cohering with the neutral or local context. So, for example:

Neutral context: Carl Barks wrote many Donald Duck stories and invented Duckburg. In his early sketches we see Huey, Dewey and Louie as young well-behaved boys with hats and scarves. They often go out to help old ladies.

Local context: Carl Barks wrote many Donald Duck stories and invented Duckburg. In his early sketches we see Huey, Dewey and Louie as young bad boys with striped sweaters and masks. They often go out to rob old ladies.

Critical sentence: Donald Duck's nephews are boy scouts/thieves and very smart.

One prediction is that the IFG should show increased activation at the critical sentence when preceded by local contexts (world-knowledge anomalies) compared to neutral contexts. A second prediction is that, since the RH is better at discourse-level processing, it should be better at integrating local context information into an inference about anomalous critical sentence interpretation and thus the IFG-RH should display less activation at the critical point ('thieves') than the IFG-LH after local contexts. These predictions are borne out by the fMRI data. So RH broad coding is of benefit at the narrative level.

In an fMRI study, Xu et al. (2005) adapted nine of Aesop's fables to match them for features such as word frequency and syntactic complexity. In three conditions, subjects were presented either with individual words from the fables, or with sentences made up of these words, or with entire fables made up of these sentences, so each subject observed three fables per condition. Subjects were instructed to read the stimuli silently. Brain activation for each of these was compared against a baseline (reading random letter strings). Classic LH perisylvian language areas were observed to be active throughout the experimental conditions, but RH activation increased with context size, reaching a maximum at the narrative level. Within that level, RH activity increased as the narrative progressed, becoming significantly more active at the end of the narrative than at the beginning, suggesting the RH is responsible for forming a representation of the narrative as a whole, further supporting point 2 above.

Failure at narrative-level inference is typical of RHD patients. For example, Kaplan et al. (1990) presented RHD and normal patients with vignettes with manipulations of two features:

1. The literal truth or falsity of one character's claims about another character's ability

2. The relationship between characters.

Literally false comments invited alternative inferences, but these inferences depended on the characters' relationships. For instance, if one character complimented another's golf technique, but it was obvious that the second character was a terrible golfer, then the comment is literally false. If the two are good friends, one might infer that the first character was being kind and encouraging; if they are not friends, one might infer that the first character was being sarcastic. The authors found that RHD patients were unable to use this contextual information to make inferences about speaker meaning.

This range of experiments, in supporting points 1 and 2 above, then, shows that differential hemispheric processing supports a distinction between contextually constrained and comparatively unconstrained inference (at least when it comes to meaning) since a small set of representations are made salient by LH processes, while RH processing evaluates a wider set, and does so in situations where a relevant interpretation is found over the course of multiple sentences, which is one of the things pragmatic inference needs to do.

2.6.6 RH processes and pragmatics

It only remains, then, to compare such claims with pragmatic interpretation as described by Sperber and Wilson (1995). While they treat pragmatic interpretation as though it were one kind of process, these claims suggest that we should distinguish at least two kinds, or more weakly, at least two aspects. I admit this may not matter to their account of inference *if* we were to treat it as a computation-level account, but the algorithmic-level difference will turn out to be important for symbol evolution. Consequently, I will not be focusing overtly on their computation-level deductive analysis, but rather on their reliance on the explanatory role of accessibility, which suggests an empiricist psychology as discussed in §2.5.2.3. By addressing my criticisms from §1.5.2.2, I will highlight ways in which symbolic-threshold inferences rest more heavily on non-deductive, creative inferences, which Sperber and Wilson admit operate in the background.

Their account involves retrieving assumptions related to utterances in the memory store of the interpretive device. These assumptions are re-

trieved in order of accessibility until relevance is achieved. Sperber and Wilson do not distinguish between highly accessible and less accessible assumptions apart from the criterion of relevance, and state explicitly that the process proceeds step by step. The various approaches above, though, suggest that LH and RH interpretive processes operate simultaneously, and that LH processes search more accessible assumptions and do so within a given context, while RH processes additionally search weakly accessible assumptions and do so context-independently, but since the RH processes are slower and weaker, they contribute most when less accessible assumptions turn out to be relevant or when coherence (for which, read 'relevance') depends on integrating representations across contexts or above the sentence level.

This can be seen clearly by comparing their account with the Beeman (1993) vignette above, where John cried out because he had cut his feet on the glass. The following discussion should be considered a parallel to my worries in §1.5.2.2 about the accessibility of representations related to OSSOBUCO. Let's assume that we are at the point of making an inference about why John cried out. If contextual assumptions had included activation of the representation SCISSORS, then activation of CUT would be quick and strong in the classic language areas of the LH. As it is, though, CUT was less accessible in this vignette and inferring that he had cut himself thus required the overlap of spreading activation in the RH, as in fig. 2.13. The conclusion in Beeman is based on 32 such scenarios, so it is not merely a peculiarity of this particular example. It might be countered that the LH would have got there eventually, if only the RH hadn't found the answer first. After all, RHD patients are still able to find *some* interpretation in the above discussion of Kaplan et al. (1990). These patients' interpretation is not the intended one, but I agree with Sperber and Wilson's foundational point that relevance is not a matter of validity or fail-safe intention-matching.

But the worry is more pressing given my second concern: though they reject a code model, it is conventional coding that provides initial access to their interpretive process (§1.5.2.3). They included novel gesture under the same umbrella as linguistic interpretation, but the work on the GSH (Mashal et al., 2005, 2007, comparing conventional — i.e. coded — and novel metaphors) showed that the more novel the metaphor, the more RH the interpretive process. It thus seems plausible that novel or unconven-

tional signals require more RH processing than (and are thus cognitively different from) conventional language. These experiments proved this point for a novel collocation of units as opposed to novel units (though there is a similarity: a ground must be inferred in both cases), so the pressing concern that falls out of all this is that we need to establish whether, or to what extent, a novel sign involves the same cognitive difference.

I have mentioned on several occasions that hypothesis generation must therefore be our focus, so the problem for language evolution now becomes establishing empirically whether, or in what circumstances, hypothesis generation in the face of a novel sign is a context-deciding inference of the type I have been discussing. If it is, it would be unsurprising that animals are generally incapable of such a thing, but that chimpanzees, for example, are capable of learning a symbol when the context is heavily constrained by experimenters. This is not to say, though, that the complexity of such inference is the only reason other animals don't manage to learn symbols, but this is a feature of the problem that has not been addressed in the literature, and it is an important gap to fill. The next chapter thus finishes off the theoretical section of this dissertation by looking at abduction, which is Peirce's account of creative or novel hypothesis generation; at insight, which Peirce claims is a feature of abduction and which is typically a STG-RH phenomenon; and at induction, which has been the focus of most work in this area but is contextually constrained.

2.7 Conclusions

In this chapter I have argued for an inferential hierarchy. It begins with minimal rationality, which is distinct from normative rationality in that it is concerned with the explanatory role of content, not with truth or optimality. The addition of contextual information to this provides us with minimally inferential cognition and I argued that this makes sense of animal behaviours such as TI. Basing the one on the other means that some kinds of inference might not be truth-preserving or optimal. Further, degree of contextual constraint should thus be a major dimension of inferential complexity in evolutionary terms in addition to higher-order relationships, which have been the main focus in the limited body of research in this area.

I then gave evidence supporting a cognitive distinction between com-

paratively constrained and comparatively unconstrained inferences. While much research has focused on constrained contexts, I argued here and in the previous chapter that we need to look beyond those for symbol origins. The following chapter thus turns to examine contextually unconstrained inference in more detail, where I will evaluate the relative contributions made by abduction and induction and argue that the former played a dominant role at the symbolic threshold.

To avoid potential misunderstanding: I claimed that contextually unconstrained inference is more complex than contextually constrained inference, and I argued that the LH dominates in constrained cases, while the RH dominates in unconstrained cases. This does not mean that I believe the LH processes evolved before RH processes, or that our LH processes are more similar to animal cognition than our RH processes. Rather, the claim is that evolution solved the relevance problem in animal inference, making certain representations salient in the context of others, whereas humans are comparatively sophisticated because we evolved two strategies that deal with the relevance problem: one for situations where meaning is comparatively predictable from context and the other for situations where meaning is comparatively unpredictable. In both cases, meanings must be inferred, unlike animal signals which are largely innate.

A question for further research, based on the previous subsection, is whether and to what extent my inferential hierarchy aligns with the mimesis hierarchy posited by Zlatev (2008). Both are layered models concerning the evolution of linguistic meaning, but Zlatev focuses on the sign, while I focus on the cognitive processes interpreting the sign. I already highlighted a parallel between Zlatev's claim that hemispheric differences might correspond to a distinction between signifier and signified in the sign function and my claim that semanticity interacts with the fine/coarse coding distinction (§2.6.3). I also drew a line from Zlatev's claim about triadic mimesis through inference about speaker intention to the role of relevance and thus context in inference (§1.4.2.3). Whether there are further connections or differences remains to be seen.

Chapter 3

Abduction, Induction and Insight

3.1 Introduction

So far, I've argued that context-deciding inference (specifically, in hypothesis generation) was a necessary part of our evolving across the symbolic threshold, and that context-deciding inference is more complex than and cognitively distinct from contextually constrained inference. I also highlighted a few relevant key terms: creativity, insight, analogy, hypothesis generation, hypothesis evaluation, deduction, and induction. The last two are types of inference; the rest are either features of inference, or of cognitive processes more generally. But some of these don't seem to match up: in particular, creativity and insight are not typical features of either induction or deduction, and it's not clear to what extent analogy is related to either. Further, as I will argue in this chapter, hypothesis generation does not fall within the boundaries of induction.

Consequently, I describe and evaluate a less well known and poorly understood form of inference called abduction, which is Peirce's label for hypothesis generation, but which he claimed is creative in that it is the only form of inference capable of producing new ideas, and which operates insightfully or by analogy, thus incorporating the above key terms that were out of place in deduction or induction.

First I discuss how *reasoning* has been preoccupied only with deduction and induction but argue that *inference* needn't be limited to these

two options (§3.2). I then present Peirce's account of abduction or hypothesis generation (§3.3) as a type of inference distinct from both of these. Thereafter, I make explicit comparisons between abduction and induction to show that hypothesis generation is not, in general, amenable to inductive explanation (§3.4). Finally, comparisons between abduction and creative mechanisms such as insight and analogy will provide testable predictions for the previous claim (§3.5).

3.2 Background

It is commonly assumed that there are two types of reasoning: deduction and induction (Burks, 1946; Goel and Dolan, 2000; Thagard, 2007). Deductive conclusions follow necessarily from their premises, while the influence of Mill (1843) has made it standard practice to call everything non-necessary 'induction'. The following is thus typical of descriptions of induction: it encompasses 'all inferential processes that expand knowledge in the face of uncertainty' (Holland et al., 1986).

Both are rational in a high-level, normative sense (as opposed to a minimally rational non-normative sense) as invoked by a rationalist psychology (Fodor, 2001) or rational analysis (Anderson, 1990). The former is concerned with truth-preserving processes (thus, primarily, deduction); the latter with optimality computations based on probability (thus, primarily, induction). This means that reasoning is concerned with standards of justification, or the extent to which beliefs are warranted. Because reasoning is conscious, explicit or reflective, these standards of justification are potentially conscious, explicit or reflective: we have the potential to ask ourselves whether a certain conclusion is warranted, and on what grounds.

However, since I have distinguised reasoning and inference, the claim that deduction and induction are the only two kinds of reasoning needn't mean that they are the only two kinds of inference[1]. Inference differs from reasoning in two crucial ways that make room for a third kind of inference: **(1)**, reasoning is propositional while inference needn't be. **(2)**, reasoning involves rationalist (normative, syntactic) strategies. I argued for the possibility of empiricist (non-normative, associative) inferences.

[1]Since Peirce doesn't distinguish reasoning and inference in the way I do, I will translate his claims about reasoning to claims about inference where appropriate.

Concerning **(1)**, both truth (for deduction) and degree of probable belief (for induction) require propositions or at least proto-propositions. You cannot believe that SOCRATES or think it's likely that SOCRATES, but you can believe or think it likely that MORTAL(SOCRATES). On the other hand, I argued that animal inference needn't be propositional: a vervet doesn't need to represent CURRENTLY-PRESENT(KIN) or KIN(x) to sound an alarm in the context of a leopard and some kin[2]. But since their behaviour is explained by mental content, though doing so requires context, their behaviour is minimally inferential in my terms.

Concerning **(2)**, minimal rationality involves an empiricist psychology (§2.3). Inference evolved out of minimal rationality, so some aspects of inference may be empiricist (§2.6). An empiricist psychology is not sensitive to logical form, but rather to causal relations between representations: representation DODO probably causes activation of DEAD, but the premise $\forall(x)(\text{DODO}(x) \longrightarrow \text{DEAD}(x))$ is not reducible to this causal relationship, and the validity of arguments containing this premise is thus not evaluated in terms of this causal relationship (Fodor, 2001).

So if deduction and induction are typically propositional and normative and if some inferences are not propositional or normative, then it is possible that some inferences are not deductive or inductive. Non-deductive inferences are ampliative, meaning that they increase our knowledge of the world by going beyond necessary conclusions, but rather than equating ampliative inference with induction (as Mill, 1843, does), a Peircean approach is to distinguish abduction and induction as different kinds of ampliative inference (fig. 3.1). Kemp and Jern (2014) equate ampliative reasoning with a broad sense of 'induction', but allow a distinction between abduction and a narrow sense of 'induction'.

3.3 Abduction

As mentioned previously, it is difficult to arrive at a clear, unified account of what Peirce thought about anything. It is typical, then, to refer to stages of development in his thought. In this section, I contrast his early views on abduction with his later views. The early account allows for more explicit

[2]Since I think this vervet representation isn't propositional, there's no sense in trying to decide whether KIN would be a predicate or an argument, hence both options above. As in the previous chapter, x represents an attentional index.

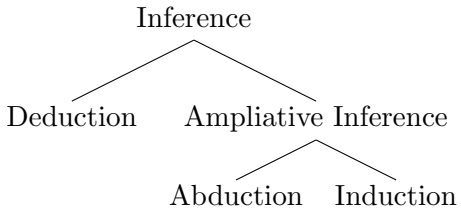


Figure 3.1: Three basic kinds of inference.

comparisons with induction and deduction; the later one is more general in that it doesn't assume propositions

3.3.1 Peirce's early views

Peirce initially offered a syllogistic account of the difference between these three types (CP 2.623, 1878): the same propositions could be arranged in three ways to yield different argument structures. In each case, imagine a room with a table in it, a bag of beans on the table and separate pile of beans also on the table. The phrase 'these beans' refers to the beans on the table.

(3.1) All beans from this bag are white
These beans are from this bag
∴ These beans are white

(3.2) These beans are from this bag
These beans are white
∴ All the beans from this bag are white

(3.3) All the beans from this bag are white
These beans are white
∴ These beans are from this bag

In (3.1), you know that all beans from the bag are white but can't see the colour of the beans in the pile. If you know that the beans in the pile came from the bag, you can deduce that the beans in the pile must be white: the conclusion follows necessarily. In (3.2), you're taking beans out of the

bag one by one and putting them in the pile. Initially you don't know what colour the beans in the bag are, but because all the beans you then place in the pile are white, you generalise to conclude that the beans remaining in the bag are probably also white. This is ampliative and inductive: you've reached a conclusion about something you haven't observed and thus don't know (there could be a black bean in the bag that you just happen not to have picked), so you generalise from a previously observed sample to a population. In (3.3), you know that the beans in the bag are white, and are wondering where the white beans on the table came from. You abductively hypothesise the possibility that they came from the bag. Again, this is ampliative because, like the inductive inference, it is potentially untrue. The claim that all ampliative inference is inductive obscures crucial differences between (3.2) and (3.3).

For instance, consider what sorts of things might make (3.2) or (3.3) stronger or weaker as inferences. In (3.2), if you have only taken two beans out of the bag, the conclusion is weaker than it would be if you'd taken 50 beans out. In (3.3) on the other hand, these quantities wouldn't make much difference: if you see two beans on the table or 50, you might still wonder where they came from, and you know all the beans in the bag are white, so statistical properties of these proportions don't enter into it.

Further differences relate to context and creativity and are thus relevant to symbol evolution. Context affects argument strength differently in (3.2) than in (3.3): the context is constrained by the premises of (3.2), but not in (3.3). In (3.3), you'd be less likely to hypothesise that the pile came from the bag if you noticed a large mound of white beans in the corner of the room. But how far does this context extend? What if you saw a pile of beans in the corridor outside the room? Or a man leaving the room with a bag of beans that had a hole in it? Or if the room was in a bean-processing plant? This is a context-deciding inference. In (3.2), in contrast, the context of the argument is limited to categories appearing in the premises of the argument: the pile and the bag. A pile of beans in the corner of the room will not affect the statistical evaluation of how likely it is that all the beans in the bag are white, given that a certain percentage of them are white. In other words, induction involves local information, abduction global information (Fodor 2001. I introduced these terms in §2.2.3 above).

Further, (3.3) goes beyond its premises more creatively than (3.2): (3.2) only introduces new *formal* information while (3.3) introduces new *content* information (CP 6.531, 1901, though he doesn't use these terms). Consider the inductive ampliative inference made by someone who observes a series of red apples, and generalises to conclude that all apples are red: $\forall x(\textsc{apple}(x) \longrightarrow \textsc{red}(x))$. This conclusion contains new formal or syntactic information (the universal quantifier $\forall$), but no new content (i.e. semantic/conceptual) information since the concepts APPLE and RED are already part of this person's representation of the world. Since these particular concepts are contained in these particular premises, the inference is contextually constrained.

On the other hand, consider the abductive ampliative inference made by someone who observes a series of apples falling, and hypothesises that something is causing everything to fall: $\exists y \forall x(\textsc{cause}(y, \textsc{fall}(x)))$. Naturally *we* already know that this 'something' is gravity, but the existence of this something, y, is a new addition to this persons' representation of the world: this is novel content information, unlike the red-apple example.

While y is new, let's suppose that his mind already contains the concept CAUSE. In that case, CAUSE wouldn't be new content information relative to his representation of the world. Nonetheless, CAUSE is not contained in the premises here, so it is new relative to this inference. The abduction thus involves retrieving information from somewhere in his set of representations of the world, but the premises themselves do not contain information that would constrain that search, so this is a context-deciding inference, just like pragmatic inference as described by Sperber and Wilson (1995). Example (3.3) is similar, in that the beans' coming from the bag wasn't observed, and is therefore new content information. To learn more about y, he would have to decide that the colour of the apples isn't a relevant feature, but that their mass is, so further progress would require relevance-deciding inference.

So the basic idea is that induction (in early Peirce) is about generalisation, while abduction is about coming up with an explanation (Burks, 1946). Further, creativity, context and probability are three differences between these types of ampliative inference. Before moving on to Peirce's mature account, a couple of points need clarification.

Firstly, abduction is guessing. But it is informed, not random guessing. If you were shown three buckets and told that food was under one of them,

you'd make a guess and pick one at random. There'd be no information to go on and this guess wouldn't be abductive.

Secondly, since these simple forms of induction and deduction are constrained by their premises, they can be minimally inferential. The transitive inference example (§2.4.2) was a low-level form of deductive inference, for instance, and many animals are able to generalise across stimuli. Even this simple abduction, however, is contextually unconstrained, so it is only likely to be useful at comparatively advanced levels of inference.

3.3.2 Peirce's mature views

Peirce eventually rejected the syllogistic approach as being too constrained (Campos, 2011). He also wanted to integrate the three types of inference into his general theory of inquiry: how humans manage to learn about the world. He thus eventually defined these types of inference according to what role they play in inquiry. Abduction generates hypotheses; deduction derives the necessary consequences of those hypotheses to provide testable predictions; induction then tests these to decide which hypothesis should be believed, and the cycle can be repeated indefinitely until it produces stable habits of belief[3]. This may sound like a theory of scientific method, but Peirce maintained that it also characterises patterns of thought in our daily lives.

Deduction remains unchanged in this new framework, though it offers a different take on induction: previously 'induction' either meant all ampliative inference, or just generalisation from observed cases to others. Now, induction is hypothesis evaluation, which is how Bayesian approaches to cognition see it. These are three quite different senses of 'induction' and it is sensible to bear the differences in mind[4].

In the earlier framework, abduction offered an explanation of an observation in the premises, but here it is made explicit that

> [a]bduction is the process of forming an explanatory hypothesis.

[3] I say 'stable habits of belief' rather than 'truth' because induction provides probability-based reasons for why it may be rational to believe something, but does not claim to reach incontrovertible truths.

[4] Tenenbaum et al. (2006) for instance, skip from the generalisation sense to the Bayesian sense in the space of a paragraph. On its own, this is not a serious problem. But if one intends to mean something useful by a claim such as 'humans are uniquely able to induce a solution to a particular problem', such a move is best avoided.

> It is the only logical operation which introduces any new idea; for induction does nothing but determine a value, and deduction merely evolves the necessary consequences of a pure hypothesis. Deduction proves that something *must* be; Induction shows that something *actually is* operative; Abduction merely suggests that something *may be*. (CP 5.171, 1903)

At this stage, Peirce's schematic characterisation of abduction is as follows (CP 5.118, 1903):

(3.4) The surprising fact, C, is observed.
But if A were true, C would be a matter of course.

Hence there is reason to suspect that A is true.

We have already seen something with the above form in Tomasello's example of a joint-attentional scene in an Hungarian train station (§1.5.1.2): A was a guess at the meaning of a novel word; C was the ticket seller's utterance of that word in that context. Despite not using this term at all, Tomasello was talking about abduction. A similar argument pattern occurred in §1.5.2.3 while I was arguing that crossing the symbolic threshold was an instance of pragmatic inference. I'll give a few more examples, then use them to unpack features of abduction. In each case, people (X or Y) come up with an abductive hypothesis that explains a feature of their environment.

(3.5) While playing charades, X shouts out a series of increasingly desperate guesses about what the novel gesture means.

(3.6) X is waiting for two friends, F1 who is always punctual and F2 who is always late. Both F1 and F2 are 10 minutes late. X wonders if F1 was held up at work, but doesn't wonder about F2.

(3.7) X and Y are medical students who have learned that symptoms S1 and S2 are both caused by diseases D1 and D2. X only recalls D1 and offers that diagnosis. Y recalls both D1 and D2 and knows that D2 causes symptom S3, which the patient has. He draws X's attention to S3. X then recalls D2 and offers this diagnosis.

(3.8) X lives in a culture where disease is thought to be caused by god's ill-will. But having noticed a relationship between degree of cleanliness and mortality, X comes up with the idea that something in the dirt causes diseases and calls that something a germ.

(3.9) X notices that the pavement is wet and guesses it has rained. He goes out and takes his umbrella. Upon leaving the house, he notices that the sky is blue. He abandons the rain hypothesis, and wonders if someone had turned the sprinklers on. Later than night, he spontaneously recalls having been told about a burst pipe on the street, and updates his hypothesis accordingly.

(3.10) X and Y spot something vaguely triangular in the water. X thinks it might be a shark and won't go in the water. Y thinks it might be a shark or may be a piece of driftwood and goes swimming.

(3.11) X is trying to diffuse a bomb. He has already examined the bomb mechanism and had a hunch that he should cut the red wire, but hasn't confirmed that yet. There is 1 second remaining on a bomb's timer, so X goes with his hunch and cuts the red wire.

(3.12) X is a stereotypical teenager who enjoys lying in his room contemplating why life and his parents are so unfair.

(3.13) X and Y are scientists. A report of data comes in calling into doubt a theory of theirs. X assumes the report is reliable and starts working on an updated theory; Y starts looking for problems in the data.

(3.14) One of the planets is not orbiting precisely as predicted by Newtonian mechanics. X hypothesises a previously unobserved planet to account for it; Y hypothesises that Newtonian mechanics is wrong.

Four points follow to clarify just what abduction is, given that hypothesis generation is poorly understood (Gettys and Fisher, 1979; Fodor, 2001; Dougherty and Hunter, 2003; Navarro and Perfors, 2011). In §3.4 I will make explicit contrasts between abduction and induction[5], and in §3.5 I

[5]I mostly ignore deduction in what follows, since word learning is a matter of ampliative, i.e. non-deductive, inference.

make comparisons between abduction, insight and analogy. What follows here, then, is just a description of abduction in order to allow such comparisons. These points are not intended to be strong arguments that abduction as described here is a psychological reality: that can wait until the empirical chapters.

3.3.3 The hypothesis is mere conjecture

The hypothesis A is meant to be 'entertained interrogatively' (CP 6.524, 1901) or as 'mere conjecture' (CP 8.209, 1905): the conclusion is not that one should believe that A; rather, one would conjecture or wonder if it's possible that A. This is a conjecture in the sense that 'P, but we conjecture that $\neg P$' is not a logical contradiction, unlike 'P, but we believe that $\neg P$' (Gabbay and Woods, 2006, 207). 'Suspect' in (3.4) implies this weaker type of epistemic attitude, whereas if one inferred A to be true, one would be committing the fallacy of affirming the consequent. In (3.5), X may shout out a number of guesses, but the whole point of the game is that X is trying to *discover* the truth of each, and only in this very weak sense is X suggesting that any hypothesis might be true. So 'whereas deduction is truth-preserving and induction is probability-enhancing, abduction is ignorance-preserving' (Gabbay and Woods, 2006, 192).

The abductive conjecture is accepted not as true or even as probably true, but merely as a provisional guide for future investigation (Plutynski, 2011; Gabbay and Woods, 2006), in that it points out things that might be worth investigating if one's resources allow and if one's goals include reaching a stable belief about the matter. These resources and goals constitute the economics of inquiry (CP 5.600, 1903). In (3.11), the bomb expert's resources (here, time) limit further investigation while in (3.12), the teenager can afford to speculate endlessly. The teenager's goals may not include reaching a stable belief on the matter, while the bomb expert would much prefer to be sure before cutting anything. In (3.5), a stable belief is reached quite easily: the gesturer simply tells X if he's right. In general, though, one would have to proceed with deduction and induction to test the matter for oneself; or new information may come along. This may lead to retraction of the hypothesis and the generation of a new hypothesis, as in (3.7). Alternatively, it might be the case that one's memory already contains the information, but that one only makes the connection much later, as in (3.9).

If the content of the hypothesis explains subsequent behaviour, it is minimally rational, at least (cf. §2.3). In (3.10), X might think it as unlikely to be a shark as Y does. But if X has a phobia of sharks, the mere activation of SHARK in the hypothesis would trigger a strong fear reaction (an emotional interpretant) in X. Even if this effect is out of all proportion to the probability of the hypothesis, its content still plays an explanatory role in X's subsequent behaviour: he is behaving that way, not because it is a shark, or because he believes it is a shark, but because he worries or conjectures that it might possibly be. Relative to the economics of inquiry, in (3.13), the actions of X are partly explained by his hypothesis that a new theory is needed while the actions of Y are partly explained by his hypothesis that the report may contain errors. Neither may yet have reached any kind of stable belief on the matter: Y might say 'Since checking the report is less time-consuming than rehauling our theory entirely, I'm beginning with that hypothesis despite thinking a mistake unlikely.'

Finally, abduction participates in a cycle of inquiry, but since it is the conjectural origin of that cycle, it is the cheapest part of the process in terms of the economics of inquiry. It is only deduction and induction that need be directly concerned with the truth or optimality of those behaviours in a particular environment. If we believe that a particular sabre-toothed tiger trap is efficient and it turns out to be a dud, we might be killed by the tiger in question, or will have wasted time, energy, and materials in its construction, so we have every reason to decide whether our beliefs are warranted. But the act of imagining new traps is unlikely to get us killed by the tiger, and we can imagine a host of new traps at comparatively little expense. Popper (1972) argues that a uniquely human step in evolution is our ability to let our hypotheses die in our stead.

3.3.4 Conjecture is best considered in the context of discovery

There have been two focuses for studies of abduction, borrowing terms from Reichenbach (1938) and Popper (1968): in the context of discovery, we have to account for how hypothesis A appears in the conditional premise in (3.4) at all, or how it is that one's representational system comes up with A in the first place; in the context of justification, we investigate whether there is some context in which (3.4) would offer a reason to *believe* that

A (Josephson, 2000), or whether some modification to (3.4) might warrant such belief (Kapitan, 1992).

The context of discovery is sometimes called creative or Hansonian abduction, while the context of justification is called Harmanian abduction, or Inference to the Best Explanation (IBE, Lipton, 2004). IBE is 'not so much an inference *to* the best explanation as an inference *that* the best explanation is true' (Paavola, 2006, 8). Such approaches try to strengthen the abductive conclusion, usually by adding further requirements. For instance, if A is the only potential explanation, then (to some degree), belief in A might be warranted. Alternatively, if A is simpler, more economical or more elegant than other hypotheses, or explains a wider range of apparently unrelated facts, or makes more detailed predictions, then belief might be warranted. Such features contribute to an explanation's 'loveliness' (Lipton, 2004), a gloss for whatever features make an explanation the best. The idea is that particularly lovely explanations are worthy of belief. There are a number of problems with such an approach, though.

Kapitan (1992) argues that if additions to (3.4) succeed in making it warranted, then that would only make abduction dependent on deduction and induction, in which case it wouldn't be a distinct form of inference. Campos (2011) argues that IBE is just deduction and induction, and that abduction is thus better considered in the context of discovery. Paavola (2006) takes the weaker position that IBE blurs the distinction between abduction and induction. But induction already has powerful, well supported techniques for dealing with warranting: Bayesianism. On a practical level, then, the IBE approach wouldn't add anything to this. Though a reliance on Bayesianism would reduce abduction to induction in the context of justification, it would still be a distinct type of inference in the context of discovery.

The problem underlying all of this is that Harmanian approaches attempt a rational or normative reading of abduction, but hypothesis generation is inherently non-normative (Gettys and Fisher, 1979): one can creatively imagine any number of sabre-tooth tiger traps, but truth or optimality only become an issue when someone actually decides to build and use one. Plutynski (2011) argues that Peirce didn't intend an IBE reading of abduction, and Sperber and Wilson (1995) similarly claim that pragmatic inference is not a matter of justification. The creative process itself is somewhat haphazard:

> It appears to me that the clearest statement we can make of the logical situation ... is to say that men have a certain Insight, not strong enough to be oftener right than wrong, but strong enough not to be overwhelmingly more often wrong than right ... The relative frequency with which it is right is on the whole the most wonderful thing in our constitution. (CP 5.173, 1903)

We *could* attempt to account for the logic of abduction by showing how an hypothesis logically guides further action. But no specific further action is determined by the adoption of a hypothesis, so this is impossible. Any potential action will be relative to the economics of inquiry and these can vary widely across contexts. So abduction is only logical to the extent that it participates in unpredictable cycles of pragmatic inquiry, and it is misguided to evaluate the logic of abduction *per se*.

We are thus left with the context of discovery: where the hypothesis comes from. I am interested in how our ancestors managed to make any guesses at all about meaning, compared to chimpanzees who have to be trained so exhaustively in constrained contexts. This is not to say that abduction is irrational. Rather, in the next subsection, I will argue that it is rational in a minimal, not normative sense. The remainder of this subsection, though, will first explore the notion of 'discovery'.

As mentioned above (§3.3.1), abduction has the potential to add new content to our representation of the world. 'Potential' means it doesn't always do this, though, so we must distinguish creative from selective abduction (Schurz, 2008). Selective abduction involves retrieval of an hypothesis from memory; creative abduction involves creation of an hypothesis not contained in memory. The medical student Y in (3.7) performed selective abduction; the genius in (3.8), on the other hand, introduced a novel concept GERM to his representation of the world. It wasn't enough to correlate dirt with disease (which would have been a generalisation of the type called 'inductive' in early Peirce); rather something unobserved in the dirt was inferred to exist and to have some properties which would explain the disease[6].

[6]Though (3.8) is a fictional example, Paavola (2006) provides a detailed account of the actual discovery of germ theory in the 19th century by Ignaz Semmelweis, to show that he performed a creative abduction, not merely a generalising induction or Harmanian abduction.

There is a graded continuum between creative and selective abduction. In (3.5), X is guessing the meaning of a novel gesture. The concepts corresponding to the guesses he makes already exist in his memory, so this is not purely creative abduction. But these concepts are not connected to or cued by the novel gesture: he is not remembering what this sign means, so this is not purely selective either. He must make a new link between an existing representation and a novel sign, so it is partly creative.

(3.14) offers examples of creative abduction taken from the history of science (Rosenberg, 1974). The orbit of Uranus turned out to differ from that predicted by Newtonian theory. Adams and Leverrier, rather than hypothesising that Newtonian theory should be abandoned, hypothesised the existence of a new planet that would explain the perturbations. The hypothesis was borne out when Neptune was observed. Later, perturbations were observed in the orbit of Mercury. Scientists again hypothesised the existence of a new planet: Vulcan. No such planet was found, though, so the orbit of Mercury remained unexplained. Much later, Einstein hypothesised the possibility of curved spacetime which would explain Mercury's orbit. While Adams and Leverrier added a new representation, Einstein posited an entirely new representational system or theory. Both are creative abductions, though Einstein's discovery required rather more insight.

I've mentioned the economics of inquiry in relation to deduction and induction, but Peirce (CP 2.776, 1902) points out that it also includes discovery: it is usually uneconomical to think up all (or even many) possible hypotheses that might explain something. Creative abduction, then, is not just grabbing at random straws, since this would be uneconomical: it would take a long time to reach a good hypothesis by random methods, and humans seem to have a knack for reaching good hypotheses very quickly.

> The human mind's having such a power of guessing right that before very many hypotheses shall have been tried, intelligent guessing may be expected to lead us to the one which will support all tests, leaving the vast majority of possible hypotheses unexamined[7]. (CP 6.530, 1901)

Having argued in this subsection that we cannot talk of the logic of abduction, though, I still have to account for how abduction can still be

[7] Compare this quotation with the first of the two options offered by Hurford in §1.3.1.3.

(minimally) rational rather than random.

3.3.5 Discovery is psychological, not logical

The argument pattern in (3.4) allows for an explicitly cognitive approach, as opposed to a purely logical or normative one. It is about how we think, not just about how we ought to think or what we are warranted in thinking; Peirce sometimes calls the former 'psychology' and the latter 'logic' (Burks, 1946) and admits that abduction is 'very little hampered by logical rules' (CP 5.188, 1903)[8]. Similarly, outside the Peircean tradition, Popper (1968) argued that discovery is not a matter of logic, but of psychology instead. This is a reasonably common view in the philosophy of science (Aliseda, 2004). Arguments in favour of a psychological approach to hypothesis generation follow here.

Firstly, schema (3.4) speaks of surprise. This is an explicitly psychological response to an event that is unexpected, i.e. incompatible with an individual's world knowledge or some previously held hypothesis, so it is context-dependent. In (3.6), X is suprised at one friend but not the other, and thus has more reason to theorise about the lateness of the former than the latter. It may have been surprising to someone pre-Galileo that a cannonball wouldn't fall faster than a feather in a vacuum, but this wouldn't be surprising to a modern physicist. In (3.5), X expects to see and is thus not surprised by novel gestures, but in this case X is required by the rules of the game to hypothesise about gestures that are not explainable given what X currently knows. It is, on the whole, a rather human desire to want to guess at all.

Secondly, the notion of explanation here is also psychological, not logical. Adapting an example from Thagard (2007), if a friend is grumpy and the hypothesised explanation is that they are stressed, then the premise according to schema (3.4) would be, 'If they are stressed, their grumpiness would be a matter of course'. But this explanation is not the logical conditional, 'If they are stressed, then they are grumpy.' One could adopt the hypothesis and still allow that the friend might be stressed without being grumpy. In the case of a material conditional, this would contravene *modus tollens*. On the other hand, a coroner can assent to the truth of the material conditional

[8]Despite Peirce's calling it 'psychology', abduction is far more likely to be studied by AI than by psychology (Thagard, 2007). This is an incongruity I hope to mend.

'if one is human, one will die' without finding 'he was human' a satisfactory explanation of someone's death (Josephson, 2000). The hypothesis is an explanation to the extent that it achieves some level of psychological satisfaction: one simply stops wondering why the friend is grumpy on this occasion if one is satisfied with the explanation, and one needn't feel pressed to find out whether the friend can be stressed but not grumpy[9].

So explanation here is rather subjective, varying according to the individual's beliefs and knowledge in addition to their goals and resources, as is typical in the economics of inquiry. Science aims at objective, logical explanations, but as Sperber and Wilson (1995) remarked, scientific progress is not necessarily the best analogy for the very fast, highly fallible inferences that humans make on a daily basis, including pragmatic inference. Sperber and Wilson thus reject this open-ended form of inference to focus on deduction. I differ from Sperber and Wilson by claiming that pragmatic inference and scientific reasoning both involve abductive hypothesis generation. The difference is that scientific reasoning proceeds more logically, and thus more slowly and dependably, *after* hypothesis generation.

Thirdly and finally, abduction allows for quite a broad view of representation: the hypothesis needn't be a proposition, but could be an image, a concept, a fact, a rule, or a theory. If you see a scratch on your car in the parking lot and spontaneously have an image of a shopping trolley scraping past then this is an hypothesis (Thagard, 2007). Logic is a matter of propositions only; this example is a matter for psychology.

An area of confusion in Peirce's writing is that he sometimes describes abduction as inference and sometimes as instinct. However, Peirce doesn't distinguish inference and reasoning as I have, so by 'inference' here he means beliefs reached by deliberate or conscious reflection (CP 2.182, c. 1902), i.e. my 'reasoning'. Peirce's use of the term 'instinct', then, simply means everything that is not consciously or deliberately reasoned: it is not limited to innate behaviour, and can be adapted by learning (Paavola, 2005). So Peirce's claim that abduction is instinctive is compatible with my sense of 'inferential'.

Peirce presents this in an explicitly evolutionary framework: 'all human

[9]Cf. Wason's card selection test (Wason, 1968) where people deviate from rational application of *modus tollens* depending on context: the context is a matter of just what is to be explained.

knowledge, up to the highest flights of science, is but the development of our inborn animal instincts' (CP/, 2.754, 1883), so my inferential hierarchy is Peircean in spirit. In particular, Peirce's claim seems to be that there is an affinity between our ideas and nature (CP 2.776, 1902). That is, our representations and the structures that they participate in are unlikely to misrepresent the world entirely. If they did, our survival would be unlikely. Abduction may be capable of flights of imagination, but if its mechanisms are rooted in representations and structures that have an affinity with nature, then it may be constrained by them, as I will argue below.

In sum, this subsection has argued that research into abduction is better carried out in a psychological framework (how we think), rather than a normative one (how we ought to think). The next subsection looks at what sort of psychological process this might be.

3.3.6 The psychological processes of discovery involve plausibility, not probability

Abduction is not a random guess. It is a process operating over information of some kind. The description above may seem to suggest (and it is commonly assumed in cognitive science) that the relevant information is probabilistic, but Peirce claims that abduction a matter of plausibility, not probability (CP 2.103, c. 1902; 2.662, 1910). These concepts are sometimes considered distinct (e.g. Gettys and Fisher, 1979; Bylander et al., 1991), but are more commonly conflated (e.g. Griffiths and Tenenbaum, 2009). Since I will argue (following Peirce) that probability cannot explain creative abduction, I need *something* other than probability to go on. This section thus offers a proposal for what plausibility might be, such that it differs from probability. It is just a proposal, however, because cognitive science has not studied plausibility as extensively as probability. Fortunately, Peirce mentions two particular mechanisms which play an abductive role: insight (CP 5.181, 1903) and analogy (CP 7.218, 1901). These are less vague than plausibility and are studied in some detail in psychology, so these will eventually provide testable predictions.

Simplifying somewhat and taking a subjectivist reading of **probability**[10], a Bayesian approach to new data evaluates a prior probability (the

[10]That is, by 'probability' meaning someone's degree of belief in an hypothesis (Chater et al., 2006). This is opposed to a frequentist reading, which concerns statistical relation-

probability of the hypothesis prior to seeing the data) and a likelihood (how probable that data would be, assuming the hypothesis) to calculate the posterior probability (how probable the hypothesis is, given that specific data). Bayes' rule makes the posterior probability proportional to the product of the prior probability and the likelihood (Griffiths et al., 2010).

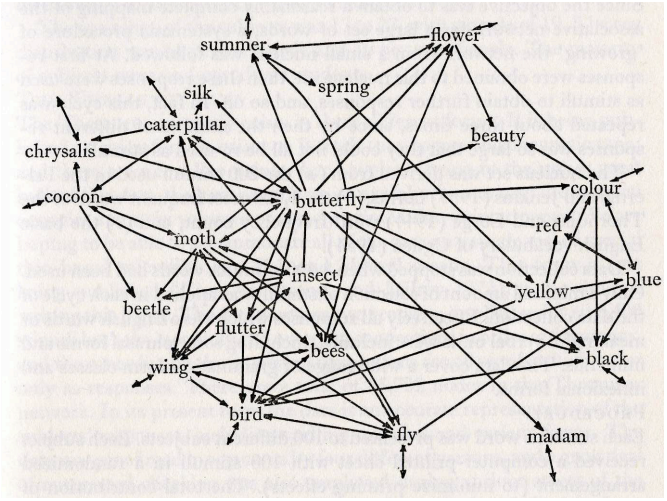
The general idea concerning **plausibility** is that it has something to do with over-all coherence in relation to background knowledge (CP 7.220, 1901). To provide something more concrete, I ground plausibility in a semantic network (fig. 3.2), a complex, layered, multi-dimensional, structured web of representations where one node causes activation of others associated with it according to the strength or weight of the connection between them. Negative strengths or weights provide inhibitory relationships. A comparison with fig. 1.1 in §1.2.3.1 suggests that nodes in these networks function as representational interpretants (Eco, 1978).

The proposal here is intended to underlie the notion of accessibility in Sperber and Wilson (1995). But in their model, the associates of a concept are accessed one-by-one, while a node here simultaneously activates all of its associates with differing strengths. Sperber and Wilson don't explicitly allow for inhibitory relationships, while I do. Holyoak and Thagard (1995) argue that something just like this underlies analogy, and it is also similar to a Bayes net (Griffiths and Tenenbaum, 2009). Bayes nets typically (but not always) express causal relationships (Griffiths et al., 2008), while I intend this to include semantic relationships in multiple dimensions.

Bayesianism typically assumes that the structure of the net and the strength of the connections between nodes are in turn derived from previous data by probabilistic principles (for instance, Griffiths and Tenenbaum, 2009). In that case, probability would be the ultimate explanation for the network. I assume that strength and structure may *partly* be derived from probabilistic information, but that this isn't the whole story since salience, relevance and context interact with probabilistic information in complex ways not yet fully understood.

An example will help distinguish plausibility- from probability-based hypothesis generation. Imagine that a friend told you they had a new pet and that you hypothesised that it was a dog. A subjective probability-based account would argue that the ultimate explanation for why you generated this

ships in the world.



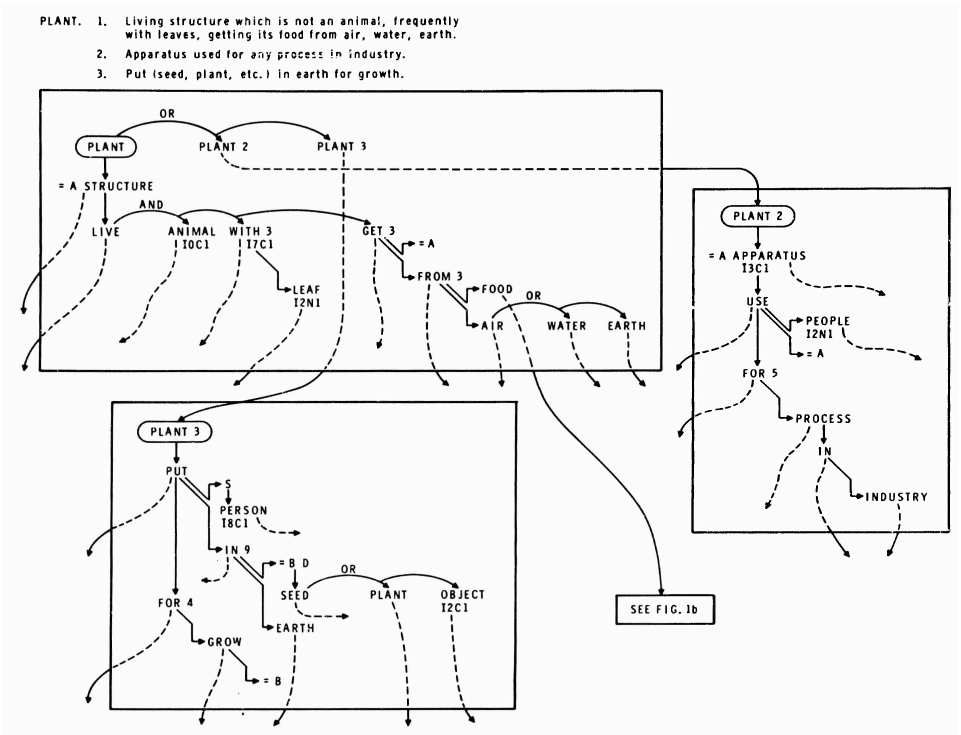


Figure 3.2: Two ways in which information might potentially be represented in a semantic web (though neither happens to show connection weights). The first shows word associations centred on BUTTERFLY (Kiss et al., 1973). The second shows world knowledge associated with PLANT (Quillian, 1968). They simply represent different dimensions of our semantic knowledge.

hypothesis first is that your prior probability for dogs being pets is higher than for other animals or, taking a frequentist view, that dogs are the most common pet in your experience. A plausibility-based account would claim that the representation most accessible from PET in your representational network was DOG.

On a plausibility-based account, the accessibility of DOG from PET isn't *necessarily* rationally derived from probabilistic data. When asked to give the first word that comes to mind when they hear 'pet', people are much more likely to say 'dog' than 'cat' (Kiss et al., 1973; Nelson et al., 1998). Perhaps dogs are more prototypical pets. However, there are approximately equal number of dogs and cats in Britain (Pet Food Manufacturers Association, 2013), where the data in Kiss et al. was collected. Unless most people's experience of pets is severely biased such that they encounter dogs much more often than cats despite the numbers being equal, this would mean that the weights in their semantic network don't necessarily reflect probabilities and thus aren't solely explained by them. I'm not claiming that probability has *no* role to play here, but I'm currently arguing that abduction is distinct because it involves plausibility, and that plausibility doesn't *reduce* to probability; nor is it an algorithmic-level approximation of probability alone. A large scale study comparing similar word associations and statistics would determine whether the above accessibility/probability disjunction is the norm or an aberration.

As a further illustration of the plausibilty/probability disjunction, there are bound to be facts that you recall with difficulty, but are certain about once recalled[11], as well as things you're uncertain about but recall with ease. Switching to a frequentist view, there are bound to be irrelevant things you've experienced frequently that you struggle to recall and salient things you've experienced once that you will never forget[12].

This discussion has been quite informal so far, but there is evidence supporting a plausibility account. Morewedge and Kahneman (2010) review

[11] Until I looked it up just now, I would have struggled to name the capital of South Sudan, even though I've read about it recently. But if instead I'd been shown a list of 20 African capitals, I'd have been able to pick out the right answer with absolute confidence, so the information is in my brain and has a high prior, but is relatively inaccessible.

[12] These are empirical questions, but they lie beyond the limits of this dissertation. This is not a serious gap in my argument, since I derive testable predictions from Peirce's claims that abduction rests on analogy and insight. I will argue that these, like plausibility, rest on representational structures.

experiments based on subjective judgements of probability, and argue that accessibility in representational structures underlies commonly observed deviations from optimal or rational judgement. One such deviation is much like the above pet/dog example and is normally called 'the representativeness bias' in the relevant literature (Tversky and Kahneman, 1983). After reading a text that describes Linda as concerned with social justice issues (among other things), participants rate it more likely that she is a feminist bank teller than that she is a bank teller, even though, rationally, the probability of two events occurring cannot be higher than the probability of one of them. The description of Linda makes her appear similar to participants' representation of a typical feminist. Morewedge and Kahneman explicitly frame the matter in dual-process terms, such that system 1 processes 'generate impressions and tentative judgments, which might be accepted, blocked, or corrected by controlled [system 2] processes' (Morewedge and Kahneman, 2010, 435). They seem, then, to characterise system 1 as associative and non-normative and the mention of 'tentative' sounds compatible with abductive conjecture.

Recall the vignette about John wading in water not knowing there was glass nearby, then crying for help (Beeman, 1993). The hypothesis was that he'd cut his foot. A basic probability-based account would have to explain how it is that an interpreter's representative system contained the hypothesis *and* how the interpreter came to represent or estimate both a prior probability for that hypothesis *and* the likelihood that he would cry out if he cut his foot. Not knowing John, I have no idea whether he tends to cry when bleeding, though I think it plausible. A plausibility-based account rests on activation spreading through a semantic network, as described in §2.6. Neurological evidence supports the plausibility account. Firstly, RHD patients struggled to generate the plausible hypothesis about why John cried out (Beeman, 1993), but these RH temporal areas are not implicated in processing inductive probability, which is associated more with hippocampal and LH frontal activation (Goel et al., 1997; Goel and Dolan, 2004). They are, however, implicated in the sort of semantic and associative relationships under consideration (Jung-Beeman et al., 2004; Bowden et al., 2005). Secondly, Beeman et al. (1994) showed RH facilitation for the target `cut` given primes `foot/glass/cry`. These correspond to general concepts, not propositions, so they (and the weights of their connections) are distinct from

whatever probabilities might play a role in a basic Bayesian account, since this would concern degree of belief in a proposition[13].

As a final way of distinguishing plausibility and probability, consider novel metaphors. I reviewed evidence (§2.6.4) showing that these involve context-deciding inferences over semantic structures and are thus of interest here. Compare 'the investors were squirrels collecting nuts' with 'the investors were trams' (Bottini et al., 1994). The first is plausible, the second nonsensical, but probability has nothing to do with the fact: investors are not squirrels and they are not trams. They are not even closely related to squirrels, so the simile 'investors are like squirrels' is quite unlike the literal statement 'rats are like squirrels'. The latter straightforwardly supports inductive generalisation from 'squirrels have novel property X' to 'rats have novel property X'. The former cannot support a similar generalisation prior to relevance-, salience- or context-deciding inferences about the ground of the metaphor. For instance, the metaphor may prompt the inference that 'investors are acquisitive' but not 'investors have bushy tails'. Having bushy tails is a salient property of squirrels, but is not salient in this particular context. Bottini et al. (1994) show in a neuroimaging study that participants' evaluation of plausibility in metaphors implicates RH areas like those active in context-deciding inference (§2.6), and Bookheimer (2002) explicitly interprets the results of Bottini et al. as a matter of context.

In sum, rather than evaluating the probability that something belongs to a certain category or has a certain feature, a cognitive system might evaluate its similarity to a prototype in semantic memory or its centrality in the structured representation of that category; rather than evaluate the probability of an event, a system might respond with whatever representation is most accessible given a particular cue. These plausibility features of semantic representation do not reduce to probability.

The discussion above has focused on distinguishing plausibility from probability in selective abduction. In selective cases, I admit that plausibility may not sound all that different in practice from probability. But even if it turned out that plausibility sometimes reduces to probability, that reduction wouldn't work outside of contextually constrained, selective cases. That is, it wouldn't extend to creative cases. Inductive accounts tend to stick to contextually constrained cases, though, so it's unsurprising if they some-

[13] I examine more sophisticated Bayesian approaches below.

times think hypothesis generation isn't much of a problem to explain. But for the symbolic threshold, we are interested in contextually unconstrained, creative abduction, so this difference is important.

Peirce's reasons for rejecting a probabilistic account of creative abduction are quite straightforward: if a problem is entirely new, one is ignorant of the hypothesis in question. Or, in the case of a partly creative abduction as in (3.5), one is ignorant of the relationship between the relevant hypothesis and the explanandum. So in the context of discovery, since one's memory lacks either a representation of the hypothesis, or of the relationship between hypothesis and explanandum, the representational system lacks either a prior probability or a likelihood function[14]. Basic forms of induction are thus unable to explain such cases.

But there are more sophisticated forms of induction that propose ways of getting around the problem of no priors, no likelihoods, or no hypotheses. These have to posit additional machinery that are much like the representational structures I've been discussing, so in the next section I examine that machinery in light of contextually unconstrained, novel hypotheses.

3.3.7 Conclusions

I've argued that abduction is conjecture in the context of discovery, and thus requires a non-normative psychological account, rather than by a logical account that looks for optimality or truth. I made some proposals for how plausibility might underlie such an account, but this rests on semantic associations rather than syntactic processes. That is, abduction is an empiricist, not rationalist process. In the next section, I turn to examine whether the rationalist processes posited by Bayesianism can ever explain hypothesis generation in unconstrained, novel contexts. That is, Bayesian accounts are appropriate to the context of justification, and are misapplied to hypothesis generation, properly considered in the context of discovery.

[14]Though Peirce didn't phrase things in these Bayesian terms, Psillos (2009) argues that this is what he meant, and that Peirce's phrase 'inverse probability' translates to 'likelihood', while 'antecedent probability' is 'prior probability'.

3.4 Abduction can't be reduced to induction

Crossing the symbolic threshold requires novel context-deciding inference. I suggested in the previous section that induction in its basic form isn't able to deal with novel contexts. Here, I consider more sophisticated Bayesian tactics for dealing with novelty and complex contexts. I conclude that even sophisticated induction is still contextually constrained and ill-equipped for creativity. The conclusion is not that induction is flawed or inappropriate; just that inductive accounts are incomplete, that this is most obvious in creative, context-deciding situations, and that abduction provides a small set of hypotheses as the constrained context for inductive evaluation, thereby allowing it to be psychologically realistic or computationally tractable. In other words, the central point of this section is to argue that hypothesis generation is not fully explicable in inductive terms, though it might seem to reduce to induction in artificially constrained cases such as experiments.

3.4.1 Types of Bayesianism and Bayesian assumptions

I must be specific about just what kind of Bayesianism I am engaging with when I say that abduction complements induction, and which I'm avoiding outright or think incompatible. There are two important dimensions here. The first dimension distinguishes enlightened from fundamentalist Bayesians (Jones and Love, 2011). The second distinguishes methodological from theoretical Bayesians (Bowers and Davis, 2012a). **(1)** Fundamentalist Bayesians are concerned only with the computational level and make no commitments to psychological mechanisms of any kind. Englightened Bayesians focus on the computational level, but are interested in what computational models tells us about human cognition at the algorithmic level. **(2)** Methodological Bayesians are enlightened Bayesians who think that computational models of Bayesian processes inform us about human cognition, without claiming that human cognition actually carries out Bayesian calculations. Theoretical Bayesians are enlightened Bayesians who think that the mind either performs or approximates Bayesian calculations, which requires representations or approximations of priors and likelihoods.

My claims in this section (that probabilistic accounts of hypothesis generation are either psychologically unrealistic or computationally intractable) are intended to engage only enlightened methodological Bayesians. Funda-

mentalist Bayesians are uninterested in the question of psychological processes, and theoretical Bayesians assume that the relevant processes are probabilistic. I recognise that there is an element of straw man (or even Bogeyman) in this picture, and leading Bayesians deny the existence of everything other than enlightened methodological Bayesianism (Chater et al., 2011; Griffiths et al., 2012). Nonetheless, Jones and Love (2011) and Bowers and Davis (2012b) quote from the literature to show either that these Bogeymen exist, or at least that Bayesians often talk as though they exist. Regardless of their existence or imputed character, I think it important to stake out the playing field clearly, and the above can serve as warning flags indicating that one is about to go out of bounds. They can usefully serve that function even if nobody explicitly intends to do so.

I think there is a particular temptation to sneak out of bounds when it comes to hypothesis generation. If one assumes that inference is either deductive or inductive (and I've mentioned that this is a widespread assumption), then since hypothesis generation is not deductive, it would follow that it must be inductive. Unless, of course, one allows for ampliative inferences that are not inductive, as methodological Bayesians must allow. However, very few explicitly do so (one example is Kemp and Jern, 2014); others tend to talk as though everything ampliative is inductive. Similarly, there might not actually be any real fundamentalist Bayesians. But if it turns out that some computational process is intractable but nonetheless performed effortlessly by humans, and if one wants to account for the evolution of this process, one should inquire into just how algorithmic processes handle the situation and whether there might be several types of algorithmic process at work, unless of course one is a fundamentalist.

With all that in place, we can now set out some basic assumptions. Recall that, for Peirce, induction is 'for testing hypotheses already in hand' (CP 7.217, c. 1901). A related point underlies Bayesian induction in that it assumes structured hypothesis spaces that are well defined or 'known, enumerated and exhaustive' (Gettys and Fisher, 1979). That is, '[a] probabilistic model starts with a formal characterization of an inductive problem, specifying the hypotheses under consideration, the relation between these hypotheses and observable data [i.e. a likelihood function], and the prior probability of each hypothesis' (Griffiths et al., 2010, 358). So the relevant hypotheses are already known and in place before induction begins its work

(Xu and Tenenbaum, 2007; Holyoak and Cheng, 2011).

While Bayesian accounts are concerned with relative strengths of those hypotheses already contained in the hypothesis space, 'they do not provide criteria for evaluating the hypothesis space itself' (Carroll and Kemp, 2013, 287). Secondly, Chater and Oaksford (2008) claim that calculating posterior probability is descriptively simple, but that assigning priors and likelihoods in the first place is the difficult part of a Bayesian approach. Both these points are emphasised by Jones and Love: '[t]he prior distribution is the well-known and oft-critized lack of constraint in most Bayesian models ... However, a much more serious source of indeterminacy comes from the choice of the hypothesis set itself' (2011, 178); 'the hypothesis space is where the interesting psychology lies in most Bayesian models' (2011, 219). Given all this, some Bayesians admit that hypothesis generation is still poorly understood from an inductive point of view (e.g. Bonawitz and Griffiths, 2010; Navarro and Perfors, 2011). However, even these tend to assume that it is somehow still inductive.

Given that a well defined hypothesis space is a basic assumption of induction, it seems that Bayesianism alone can thus never provide a psychological account of how new hypotheses are discovered. But it has nonetheless made some valiant efforts, which I must thus examine in coming subsections, where I'll be arguing for cognitive mechanisms that involve neither representations nor approximations of probability. The above boundary markers will be important in what follows, given that it is standard practice for cognitive Bayesians all of stripes to assume rational analysis, while treating abduction as rational is to misconstrue the nature of discovery: the minimal unit that includes abduction and is rational is the entire process of inquiry (§3.3.4).

3.4.2 Constrained and unconstrained word-learning tasks

Xu and Tenenbaum (2007) investigate word-learning by presenting participants with a picture together with a novel label. The picture could be interpreted by categories at various levels of generality: the label accompanying a picture of a red pepper could mean 'red pepper', 'pepper' or 'vegetable'. Before the word-learning task, however, the experimenters showed the participants all 24 items in the test. Participants could thus be pretty certain just what counted as a possible hypothesis by the time they had to learn novel words, and it was not a large set (*pace*, Quine). This task is

thus constrained by the experimental methodology. The authors admit that the main work is done by specifying priors and likelihoods in their model, and that, from a psychological point of view, hypotheses could be derived from similarity judgements (to be discussed further when I come to look at analogy in §3.5.1).

Compare the above with a less constrained word-guessing task such as charades. Let's say that a particular game is limited to English nouns. Like the above constrained example, players must generate hypotheses about the meaning of a novel signal. To do so, either **(1)** guessers must evaluate a vast hypothesis space of all English nouns, or **(2)** they can first make inferences to generate a small, focal set of hypotheses to evaluate.

If **(1)** were true, then by an inductive account, participants would have to evaluate the posterior probability of all nouns in their memory. However, both within Bayesianism and outside, it is generally recognised that for such complex, unconstrained cases (so typical of real-world problem solving), exact calculations of posterior probabilities are either computationally intractable or, if tractable, would take much longer than is normal for human problem solving (Chater and Oaksford, 2008; Griffiths et al., 2010; Kwisthout et al., 2011; Brighton and Gigerenzer, 2012). Without mentioning Bayesianism, Sperber and Wilson (1995) make a similar point when they argue that pragmatics is context-deciding (cf. §1.5.2.1). These are computational-level worries, but there is also experimental evidence demonstrating that evaluating probabilities in large spaces is psychologically unrealistic.

Smith et al. (2011) investigated word-learning in situations where participants must track probabilities across contexts. Context size could vary. They found that participants were unable to track probabilities accurately once the context contained more than eight entities. It is unlikely, then, that the charades problem is solved by evaluating posterior probabilities for all English nouns (or, indeed, for many). Other experiments demonstrate that humans do not actively evaluate large sets of hypotheses: typically, we generate a comparatively small, focal set of *relevant* or *salient* hypotheses on the fly (Gettys and Fisher, 1979; Johnson and Krems, 2001; Dougherty and Hunter, 2003; Thomas et al., 2008). If the generated focal set is well defined, induction avoids having to evaluate posterior probabilities across vast spaces.

It is thus computationally intractable *and* psychologically unrealistic for humans to inductively evaluate vast hypothesis spaces: the evidence shows **(1)** to be false, leaving **(2)**. Evaluation of a vast range of hypotheses is possible in theory, but in practice some process produces a focal set for evaluation. The vast range of possible hypotheses may exist in the semantic memory of the agent in question (as in a game of charades), or the hypothesis might not exist at all in any psychologically realistic sense (before Einstein thought up curved spacetime to explain the anomalous orbit of Mercury, it could not have realistically been the case that this idea existed in anyone's memory). Hypothesis generation either selects or creates hypotheses for a focal set, placing them in working memory. The conclusion here is a two-step model: generation of the focal set (by selection or creation), then evaluation of hypotheses in working memory (Gettys and Fisher, 1979; Dougherty and Hunter, 2003; Cherubini et al., 2005; Gabbay and Woods, 2006; Thomas et al., 2008; Bonawitz and Griffiths, 2010; Navarro and Perfors, 2011).

Consequently, a potential issue with all inductive experiments (such as Xu and Tenenbaum, 2007, above) is that they begin with the second step: by giving the participants the relevant set of hypotheses, they constrain the context of the inference. This may not be problematic for methodological Bayesian purposes, but it does mean that the results of such experiments are not necessarily informative in accounting for symbol origins. I argued that the symbolic threshold requires an account of how humans generate hypotheses in unconstrained contexts. Therefore, an inductive account of word-learning in a constrained context cannot be used to argue for an inductive account of symbol origins. I've discussed how Medina et al. (2011) raised the same issue in child world learning (§2.5.2.2).

3.4.3 Two-step models: single- or dual-process?

The plausibility of two-step models (first hypothesis generation, then evaluation) raises the question of whether the two steps are cognitively similar at the algorithmic level. Although the first step does not involve computation of posterior probabilities, hypotheses could still be selected for inclusion in the focal set according to inductive features (such as priors or likelihoods), in which case the two steps would be quite similar. If this fails, it could be that some or other probabilistic algorithm succeeds.

In the charades example, assuming uniform priors won't help, since that

wouldn't recommend any hypothesis over another for inclusion in the focal set, and the context would remain all possible nouns. Neither will assuming that priors can be derived from word frequencies: words used in a game of charades are typically infrequent, so working from more frequent to less frequent meanings would still involve evaluating a vast space.

Bonawitz and Griffiths (2010) provide experimental support for the basic idea that hypothesis generation is not simply a matter of priors or likelihoods. They presented participants with a variant of a transitive inference task. Participants saw pairs of cubes marked with letters approaching each other. When the cubes met, one would light up according (unbeknownst to the participants) to an arbitrary ordering of those letters. Before the inference task, one group of participants was primed with a vignette about teachers observing children playing a game and trying to predict who would win. The vignette stated that teachers realised that each time the taller child won: tallness is a transitive relationship. The other group was given a neutral vignette which didn't prompt a transitive interpretation. The primed group were more likely to solve the cube problem, but both groups rated the answer equally probable when given it. The results of the experiment would be surprising if participants represented the relevant priors and likelihoods and if hypotheses were generated for evaluation because they were probable by either count. A natural interpretation, I think, is that humans simply don't explore a vast hypothesis space, selecting hypotheses for inclusion in the focal set according to priors and likelihoods. Bonawitz and Griffiths seem to assume that generation still somehow involves approximations of probability, but they note that larger hypotheses spaces will compromise the success of these approximations, which raises the question of how humans still manage to be quite good at this sort of thing in open-ended contexts.

In fact, there are serious issues with trying to give any kind of probabilistic account of how the focal set is generated. It turns out that our evaluations of probability are sometimes relative to a focal set. For instance, Dougherty and Hunter (2003) show that the size of the focal set (determined in part by individual differences in working memory) influences judgements of probability. Sprenger et al. (2011) show that dividing attention also has an effect on probability judgements, and claim that memory encoding and retrieval (i.e. accessibility) play an explanatory role here. Hayes and Newell (2009) demonstrate, by manipulating the relative salience of focal hypotheses, that

different inductive inferences are drawn depending on the set. Since subjective probability is determined relative to a focal set, it becomes doubtful whether it can fully explain the set's constitution. Tversky and Kahneman (1974) show that participants often neglect priors and sample sizes in their probability judgements. I discussed why sample size is irrelevant for abduction, unlike induction (§3.3.1).

Kaplan and Simon (1990) tested participants on a problem called the Mutilated Checkerboard[15]. Solving the problem either involves an exhaustive search through a vast problem space, or involves noticing salient features of the problem and forming a representation of them. Some representations were relevant, others irrelevant. A relevant representation allowed comparatively quick solution times. They found that participants typically began with an exhaustive search, but that since that '[s]ubjects are not equipped with generators for searching the space of "all possible representations"' (Kaplan and Simon, 1990, 403), they became discouraged with this approach and switched to the other. The authors describe this switch in strategy in terms of salience, relevance, representation and insight. Because the strategic switch resulted in a significant shortening of the solution time, I take it that these features are implicated in producing a focal hypothesis set. I eventually aim to show that they are.

There are also a few computational models divided (roughly) along the sorts of lines I'm proposing. Johnson and Krems (2001) argue that since context-dependent processes compete with context-independent ones, a complete model of abduction requires both connectionist (system 1) and syntactic (system 2) processes[16]. Thomas et al. (2008) examine hypothesis generation and evaluation in medical diagnosis, roughly the sort of situation described in example (3.7) above. The model does not evaluate the posterior probability of all possible hypotheses held in the semantic memory of the agents, but just evaluates a small set held in working memory. Hypothesis

[15]Participants were asked to decide (and prove informally) whether it's possible to cover all squares on a checkerboard with dominoes that each cover two squares (horizontally or vertically adjacent) after the top left and bottom right corner of the board have been removed, leaving an even number of squares. Solution involved noticing that the removed pieces were the same colour, but that all possible domino positions covered different colours. So it is impossible to cover all squares with dominoes.

[16]They consider abduction in the IBE sense, though, which assumes a degree of normativity and blurs the lines with induction (§3.3.4). It is uncontroversial to characterise induction as rationalist (and thus system 2), so my focus here is on whether abduction is distinct from induction by including system 1 processes.

generation is what selects a subset of the hypotheses in semantic memory for activation in working memory, so it is a two-step model. Hypothesis generation in the model is a combination of accessibility and similarity, not explicitly encoded probability, so it is a dual-process account, too. Hélie and Sun (2010) argue that, in order to model human performance in tests of creativity, insight and inference, one needs dual-process architecture with redundant representations: both distributed, associative representations and local probabilistic representations.

These experiments and models suggest that hypothesis generation is not solely a matter of inductive probability, firstly since probabilities are judged relative to a focal set, not independently of it; and secondly since generation is describable in non-inductive terms. It is thus highly plausible that something other than probability might play an explanatory role in how the focal set is generated in the first place. Based on my discussion of contextually unconstrained inference in ch. 2 and abduction in §3.3, and the evidence here, candidates for such a role include insight, analogy, salience, relevance and accessibility. These are all hallmarks of abduction in unconstrained contexts.

That is, I am proposing both a two-step model (generation then evaluation) *and* a dual-process model (system 1 is empiricist and system 2 is rationalist). These two parameters are logically independent: Bonawitz and Griffiths (2010), for instance, assume a two-step model without assuming dual processes. What I am arguing, then, is that the parameters are not independent in practice, given the problem of novelty in unconstrained contexts. 'System 1' and 'system 2' are used in variety of senses by different authors (Evans, 2008), but a common distinction involves associationist vs. syntactic implementations. My characterisation of them as empiricist and rationalist, respectively, adds normativity to the mix: abduction is non-normative because it is conjectural, which is why typical normative arguments against system 1 accounts (for instance, Fodor, 2001; Griffiths et al., 2010) don't apply. Abduction doesn't need to be normative, because that's what induction is for. It is common to characterise system 1 processes as modular and system 2 as global (Evans, 2008; Mercier and Sperber, 2009), but since abduction is contextually unconstrained, it is in fact global (Fodor, 2001, cf. §3.3.1 above).

In sum, I predict that there should be measurable differences in cognitive

processing between constrained and unconstrained contexts. Abduction and induction both play essential roles in unconstrained contexts, since these require a two-step process. But in constrained contexts (such as inductive word-learning experiments) the role of abduction is reduced because a focal set of hypotheses is simply handed to participants. Because I've argued for a dual-process account of the two-step process, I predict that these cognitive differences due to contextual constraint should involve terms highlighted above, such as insight, salience, relevance and accessibility. These predictions are tested in part II.

3.4.4 Missing information: the problem of no priors

The previous section focused on the problem of context-size. But hypotheses can be unconstrained in other ways, too. In charades, the gestures are novel. Even though someone trying to interpret the gesture has a representation of the eventual answer stored in their memory, this doesn't mean they are able to make the link between that particular representation and that particular gesture (unless they always link all possible nouns to all possible gestures, which is psychologically unrealistic). Psychologically, there simply is no likelihood function. At the symbolic threshold, our ancestors would have lacked priors, likelihoods, whole hypotheses, and even entire theories of meaning.

Before I look at how Bayesianism deals with novelty, I'll outline how the lack of likelihoods and priors in novel cases makes a principled difference to the problem. Binmore (2009) and Brighton and Gigerenzer (2012) argue that Bayesian methods are only appropriate in small worlds. A well defined hypothesis space is a small world: its elements are known, as are the associated priors and likelihoods. If there is missing information, such as uncertainty about which hypotheses are in the set, or unknown probabilities, it becomes a large-world problem. They argue that large worlds increase uncertainty in both variance and bias, and that methods of reducing one tend to increase the other. Large worlds (and thus novelty) thus undermine the assumptions and (thus the validity) of a rational analysis approach. Bayesianism often ignores this limitation, though (Gigerenzer and Sturm, 2012). I argued that abduction is not a rationalist process, so it is coherent that a two-step dual-process account should treat a large-world problem as abductive and a small-world problem as inductive.

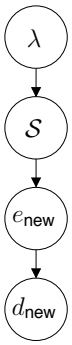


Figure 3.3: A hierarchical Bayesian model (Griffiths et al., 2008).

This worry notwithstanding, Chater and Oaksford (2008) suggest that one way of constraining probabilistic models (for instance, when lacking priors or likelihoods) is to make representation, rather than probability, foundational. In that case, probabilities can be inferred from representational structures. Hierarchical Bayesian models offer a plausible way of doing this: representations may fall into higher-level categories or be decomposable into lower-level features, so missing information at one level might be derivable from representations at other levels in a hierarchical structure.

For example, Griffiths et al. (2008) discuss how priors can be derived from background knowledge in a case of novel property induction. Learners discover that members of a category have a novel property (e.g. that gorillas carry enzyme X132) and must then decide how far this property extends (e.g. whether chimpanzees carry X132). Assuming the hierarchical model in fig. 3.3, they are trying to infer e_{new}, the extension of the property; they have observed a subset of d_{new}, the data that might result from that extension.

In this example, e_{new} is a binary vector representing the extension of the property. Let's assume that learners have hypothesised that the element of vector e_{new} corresponding to gorillas is 1 since d_{new} includes observations of gorillas with that property; it would be 0 otherwise. The task is to infer values for other elements of the vector for which there are no observations (such as the element corresponding to chimpanzees). Given that the property is new, they do not initially represent a prior for chimpanzees having X132.

The point of a hierarchical Bayesian model is that, since there are no priors for elements in e_{new}, priors can be derived from higher-level information in $\mathcal{S}$, which is the structured background knowledge relevant to this problem. In this case, $\mathcal{S}$ is a tree structure representing taxonomic information: chimpanzees are closely related to gorillas, and seals are less closely related (fig. 3.4). Priors for elements in e_{new} can be derived from the structured representation in $\mathcal{S}$: the closer an animal is to gorillas in the taxonomic tree, the higher the inferred prior for the element of e_{new} corresponding to that animal: chimpanzees would have a higher derived prior for e_{new} than seals.

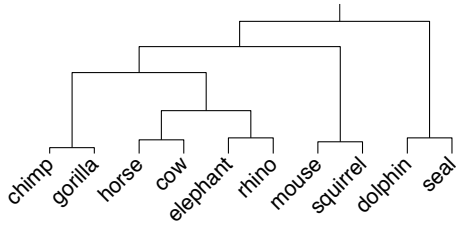


Figure 3.4: Representational structure in $\mathcal{S}$ (Griffiths et al., 2008).

Since $\mathcal{S}$ is itself learnt, its parameters are not fixed and must be derived from a higher level. So 'the hyperparameter λ specifies a prior distribution over a hypothesis space of structured representations' (Griffiths et al., 2008, 26). We cannot have an infinite regress of derived probabilities, so at some point we need a level with parameters that are fixed in advance, hence λ.

It is possible to include additional levels between $\mathcal{S}$ and λ, though. In the above example, $\mathcal{S}$ is a taxonomic tree, but Griffiths et al. (2008, 2010) and Tenenbaum et al. (2006) point out that other problems might need chains, rings, sets of clusters, linear continua, Euclidean spaces, domain-specific theories (such as Newtonian physics), schemas, causal Bayes nets or various other kinds of representation. Fig. 3.5 shows suitable structures for, respectively, a novel property problem of the sort discussed above; a problem where the property is a disease and the structure reflects the intuitive theory that diseases are much more likely to pass from prey to predator than the other way around; a linear representation animals ranked by weight. In fig. 3.6, $\mathcal{F}$ allows us to infer which kind of $\mathcal{S}$ best fits a given set of data.

In sum, when faced with a lack of priors, hierarchical Bayesian models

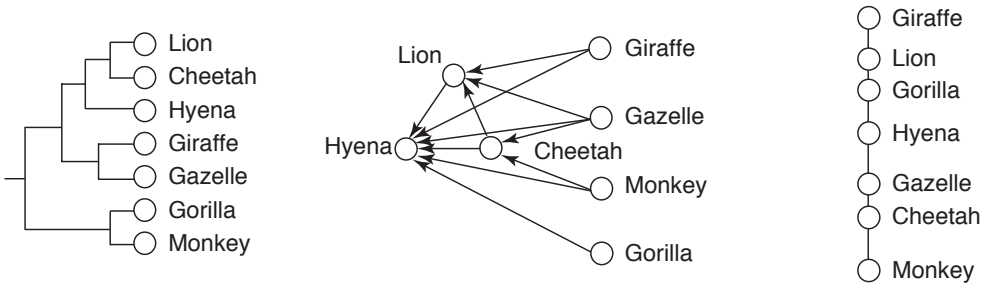


Figure 3.5: A range of representational structures (Tenenbaum et al., 2006).

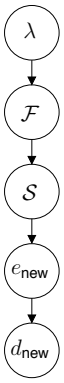


Figure 3.6: A more complex hierarchical model (Griffiths et al., 2008).

ground probability in background knowledge such as a taxonomic tree. The structure in $\mathcal{S}$ is not a representation of probability *per se*, but probabilities can be derived from it. This is a tactic I wholeheartedly support, given that I argued that such representations underlie plausibility (§3.3.6). I will, however, mention some important differences below between a purely Bayesian approach and one adding abduction as a prior step because, now that we've pushed the question of novelty back to representations, we must consider how those representations are learnt or discovered.

3.4.5 Where do representational structures come from? A problem of relevance

I agree with Bayesianism that hypothesis generation can be based on representational structures. Where we differ is on the question of where these structures come from. Griffiths et al. (2008) derive them from previous Bayesian processes, such that the ultimate explanation of each structure is probabilistic. I argue that this is not always the case. If they are not produced by a Bayesian process, but play a role in hypothesis generation, then hypothesis generation isn't reducible to induction.

Griffiths et al. (2008) argue that learners in the above example would have observed familiar properties across those same categories prior to observing novel property X132, and that the structure in $\mathcal{S}$ is thus learnt from that data. Initially, $\mathcal{S}$ was an hypothesis space of all possible trees over the animals seen in fig. 3.4. This space is searched to find the tree with the highest posterior probability relative to that data[17].

I gave a few examples in §3.3.6 of how accessibility in representational structures might not reduce to probability. Here, though, I'll try a different tack. Since they do not explicitly assume a two-step process here[18], Griffiths et al. (2008) do not distinguish the full, rich, vast, unbounded representational structures stored in semantic memory ($\mathcal{S}_s$), from the severely simplified, pruned structures stored in working memory ($\mathcal{S}_w$). However, two-step models require that, for the purposes of a particular task, a manageable subset of one's semantic memory is extracted and put in working memory as a basis for induction, so fig. 3.5 and fig. 3.7 represent $\mathcal{S}_w$. It is psychologically unrealistic that any $\mathcal{S}_w$ might pre-exist a novel inductive task, because then all possible subsets would have to pre-exist.

Because Bayesians conflate these two representations (or at least, construct models that fail to distinguish them), but given that a two-step model insists on a difference, their account of novelty conflates explanations of how $\mathcal{S}_s$ is learnt and explanations of how *this* particular $\mathcal{S}_w$ comes to be in

[17] Naturally, the number of possible trees is vast, but the authors point out that the calculations can be approximated with Markov chain Monte Carlo methods. Whether humans actually do so is an open question, though Griffiths et al. (2012) profess not to be theoretical Bayesians, which means they should think not. I don't intend to argue.

[18] One of these authors, Griffiths, argues for a two-step model in Bonawitz and Griffiths (2010). The latter paper didn't involve a hierarchical model, though, so representational structure didn't enter the discussion there.

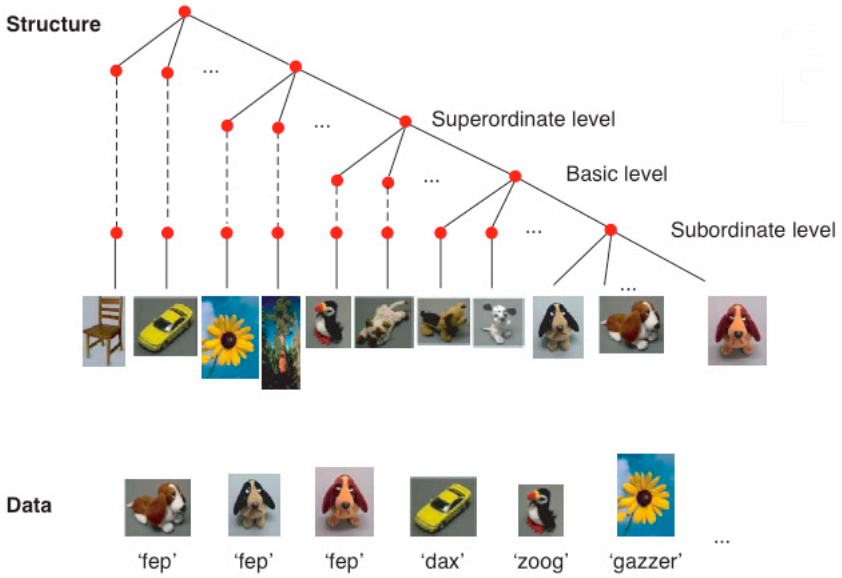


Figure 3.7: The structured representation ($\mathcal{S}_w$) that an inductive account supposes in a word-learning task (Tenenbaum et al., 2006).

working memory when dealing with *this* particular inference[19]. The latter introduces a problem of relevance which inductive accounts usually manage to avoid, either because $\mathcal{S}_w$ is derivable from a higher-level theory $\mathcal{F}$, or because $\mathcal{S}_w$ is artificially constrained by experiment or model designers. For example, when Edmond Halley was trying to work out whether three observations were of the same comet or three different comets, Newtonian theory dictated which properties are relevant for comet orbits: the mass of comets is relevant, their colour is not. The taxonomic tree in fig. 3.7 was constrained by showing the experimental participants all the stimuli before training. In either case, there has usually been a degree of abstraction prior to the inductive inference. The question now is whether it is possible to give a inductive account (either psychologically realistic or computationally tractable) of how humans manage to perform relevance-deciding processes in unconstrained contexts.

[19]This problem cropped up earlier, while discussing Lewis's suggestion of a scarecrow in quicksand (cf. §1.4.1.3)

Shafto et al. (2005, 2008), working within a Bayesian paradigm, acknowledge the basic problem here. They point out that a cat has lots of features (it climbs trees, eats mice, has whiskers, is feline). If we find out that mice have a certain disease, it's the mouse-eating feature of cats that turns out to be relevant to whether cats have the disease, not the having-whiskers feature. To investigate this context-sensitive problem, a hierarchical Bayesian model would need to distinguish relevant from irrelevant features. They construct two models, each with a different representational structure in $\mathcal{S}$: one involves a taxonomic tree, the other a Bayes net (much like the first two structures in fig. 3.5). They find that the tree structure predicts human performance on a novel-biological-property problem and the net structure does so on a novel-disease problem. So different theories guide inferences about different properties, and such inferences are thus context-sensitive. However, they designed the models such that the causal structure applies to reasoning about disease, because diseases are more likely to pass from prey to predator than the other way round. That is, their models are sensitive to relevance, but are not relevance-deciding.

Though not directly addressing the problem of relevance, Kemp et al. (2004) describe a model that can itself infer which of two representational structures (taxonomic tree or linear continuum) best suits certain data. The two types of theory were built into $\mathcal{F}$, so the model is solving the relevance problem in $\mathcal{S}$ by choosing from a built-in list at a higher level in the Bayesian hierarchy: the problem is still artificially constrained, just at a higher level. Kemp et al. (2007) acknowledge that inductive learning in novel contexts would be impossible without such theories (or overhypotheses) and that some theories are innate, but they assume that all others are ultimately explainable by yet more abstract hyper-theories in a hierarchical Bayesian model[20].

Inductive explanations of novelty have thus been pushed back from the representational to the theory level, and I will investigate theories in the next subsection. For now, though, there are formal, computational reasons for thinking that the problem of relevance imposes a principled limit for Bayesian approaches. Kwisthout (2012) offers a computational complex-

[20]They mention another option: theories could be inferred by analogy with previous cases. Tenenbaum et al. (2006) acknowledge, however, that analogy is not yet explained by induction. I show in §3.5.1 that analogy is nonetheless a matter of representational structure, in which case not all representational structure is explained by induction.

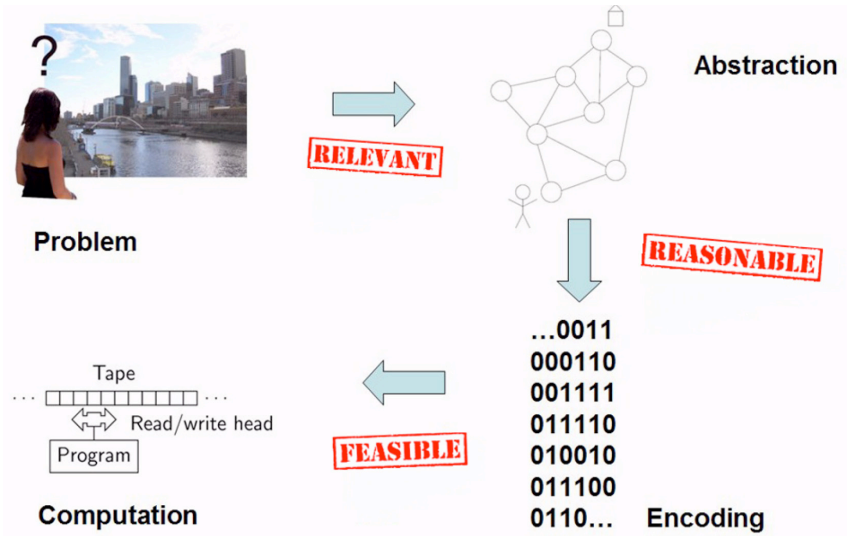


Figure 3.8: When faced with a problem, full, rich, unbounded reality is first abstracted to produce an input to a computational problem (Kwisthout, 2012).

ity analysis of the general problem I'm looking at here. He mentions that computational complexity is usually discussed in the context of deriving solutions, i.e. whether it is computationally tractable to derive a solution from certain inputs given a formal syntactic representation of the problem (the arrow marked 'feasible' in fig. 3.8). But this formal representation presumes that a relevance-deciding process of abstraction has already taken place: the input in these computational models is not full, rich, vast, unbounded reality. If, for instance, one's goal is to get to one's hotel quickly, then the problem of achieving that goal is computed given certain inputs such as position, distance and time (as Kwisthout points out, if it's 3am one might have to take a taxi, rather than public transport). The pattern of paving stones where one is standing or the amount of light currently emitted by the Pleiades are not relevant, so they are not part of the input to or representation of the problem. If one's goal is to evaluate Euler's Seven Bridges of Königsberg problem, then distance and time are not in fact relevant, since the problem depends only on topology.

'[A] computational complexity analysis typically assumes that a relevant abstraction of the problem is readily available' (Kwisthout, 2012, 20), but these complexity analyses typically don't take the abstraction process itself

into account. The state of reality at any given moment would need to be described by a massive number of dimensions or properties, but hardly any of these will turn out to be relevant to a given problem. Kwisthout thus investigates computational complexity in the context of abstraction (the arrow marked 'relevant' in fig. 3.8).

Kwisthout's formal analysis shows that, in general, the abstraction problem is intractable. So the complexity of a problem is not just a matter of solving it, but also of representing it. This result applies to my discussion above about deriving $\mathcal{S}_w$ from $\mathcal{S}_s$: even if a Bayesian account explains some structural aspects of $\mathcal{S}_s$, it does not explain extraction of $\mathcal{S}_w$ from $\mathcal{S}_s$. If inductive accounts ignore the implications of a two-step model, they risk glossing over this problem.

Kwisthout notes, however, that the abstraction problem can be tractable in specific cases if people have subjective expectations of relevance. These are derived from prior knowledge, among other things. We know what's relevant in getting to our hotel in part because we've had to get places before. Halley knew what was relevant in computing comet orbits because he had studied Newtonian theory. Sometimes we will expect few dimensions to be relevant, sometimes many, but when we have no prior experience with a problem, we have no idea how many dimensions are relevant and the problem remains intractable. We might have experience with a roughly similar problem (cf. the suggestion of Kemp et al. above), but making this connection involves analogy, for which see §3.5.1 below.

There are easy-to-abstract and hard-to-abstract problems, and inferring the meaning of novel symbols is one of the latter because symbols involve relevance-deciding inferences about grounds (§1.5.3). At the symbolic threshold, our ancestors would have no experience with problems of this type, so would have had no expectations about which features of a novel sign were relevant. Kwisthout (2012) explicitly links his claims to relevance in the Sperber-and-Wilson sense, and to the Frame Problem. I discussed the link between these in §2.5.2.3, so Kwisthout's formal analysis supports the less formal claims I made there. He also links relevance to insight, for which see §3.5.2 below.

In conclusion, Bayesian hierarchical models propose to solve the problem of novelty. However, they pose a problem of relevance. Potential solutions to this problem include theory-derived constraints or analogy. The symbolic

threshold still presents an extremely serious problem for inductive accounts because neither of these tactics could have worked and because, as I now turn to argue, it involved novel theories.

3.4.6 Extreme cases of novelty: the problem of no theories

Halley was able to mechanically derive an hypothesis space from Newtonian theory since the theory decides what properties are relevant (and thus what observations should be made, or what the relevant data would be), and it gives laws relating those properties (Griffiths and Tenenbaum, 2009). But the transition from Newtonian to Einsteinean physics is an entirely different story (§3.3.4): the orbits of Uranus and Mercury deviated from Newtonian theory. Hypothesising a new planet only fixed the first problem; the second remained a mystery until Einstein thought up relativity and curved space-time. Einstein's hypothesis involved creatively going beyond Newtonian theory since it was not mechanically derivable from that theory in the same way that Halley's was (Rosenberg, 1974). Rosenberg explicitly describes theory change as abductive.

Einstein's creative leap produced an entirely new way of representing the universe, but simpler cases involve new concepts rather than entire theories. This example is taken from Paavola (2006) and Weisberg (2009). Ignaz Semmelweis, while trying to explain unusually high mortality rates in a particular maternity clinic in Vienna, considered and rejected a range of hypotheses. Fortuitously, a friend of his pricked his finger while conducting an autopsy and died with similar symptoms. Semmelweis noticed that in this particular clinic, doctors would move from the autopsy room to the maternity ward. He thus posited some 'cadaveric matter' that caused the disease, the origins of germ theory. Paavola (2006) and Weisberg (2009) both describe this process as abductive. Paavola highlights the role of similarity which allowed Semmelweis to make a connection between two previously unconnected things; Weisberg argues that the process cannot have been inductive.

The previous subsections looked at how Bayesianism deals with representations by deriving probabilities from a theory, whereas the present examples involve a novel successor theory or novel concepts at the theory level. Until these hypotheses were created and disseminated among the scientific or medical community, nobody's representational systems could have included

information about them, so nobody could have represented priors for them, so theory-based induction cannot account for the creation of these hypotheses. Neither could anyone have estimated the probability, since estimation requires representation of the thing to be estimated. In short, probability does not explain how the idea was created by Einstein's or Semmelweis' mind, nor does anything else at the theory level.

> Major scientic breakthroughs provide striking examples, introducing wholly new concepts and theories that no one could have had a prior degree of belief in. ... Even just day to day experience provides hypotheses for which we do not have prior degrees of belief. I am right now wondering why I feel fatigued despite having drunk four cups of coffee. I think it most likely that the regular and decaffeinated pots have gotten mixed up, so that I have been drinking decaffeinated coffee all morning, but I had no prior degree of belief in that hypothesis when I walked into the cafe. (Weisberg, 2009, 133)

A hierarchical Bayesian response to this criticism would be that, just as a prior for property X132 could be derived from a higher-level theory, so could a prior for the theory of curved spacetime be derived from a yet-more-abstract hyper-theory. The hyper-theory would contain all possible physics theories and a prior distribution for them. Indeed, this is just what Griffiths and Tenenbaum (2009) suggest for problems of this type, though their example involves a constrained set of just two entities, one property, and two possible theories linking them. Glymour notes that '[Bayesians] are singularly unembarrassed by the rarity of explicit probabilistic arguments in the history of science' (1981, 67) and I have the feeling that spelling out entire theory changes in this way would sound just as unappealing to enlightened methodological Bayesian ears as it does to mine.

Anyway, we would then need to know where the principles, categories and probabilities of the hyper-theory come from, so a hyper-hyper-theory must be posited to account for that. Griffiths and Tenenbaum admit that 'at this point, concerns about an infinite recursion, providing no ultimate solution to the question of how people learn causal relationships, seem justified' (2009, 708). They're not too worried by this, though, since they seem to slip into a somewhat fundamentalist mood at this point. They later make

two concessions to the demands of psychological reality (though these don't address the psychological reality of hyper-theories themselves): **(1)** they allow that hyper-theories eventually bottom out into basic assumptions about causation. Griffiths and Tenenbaum don't spell this out, but Kemp et al. (2007) explicitly claim that these could include innate general theories of causation. **(2)** Associative mechanisms *could* play algorithmic roles *if* those mechanisms serve as approximations for these increasingly abstract probabilities, allowing probability to bear the explanatory burden.

It is unclear just how a general, abstract theory of causation could ever explain how Einstein had such a momentous insight, but there is a more pressing problem with applying **(1)** to the symbolic threshold: it may or may not be the case that humans have an innate idea of causation, but meaning is not causation. Prior to the symbolic threshold, there would have been no innate theory of symbolic meaning in any psychologically realistic sense.

Further, while physics or medical theories seek (to varying degrees) to represent reality, meaning is a purely cognitive relationship, so it is doubly inappropriate to posit hyper-theories that derive theories of signification. It is not the case that some australopithecine with an abstract hyper-theory about theories of signification was hanging around the savannah thinking to itself, 'Gee, I guess '*gavagai*' could mean rabbit, but I don't know which theory of signification is most appropriate here, so I'll put symbolic communication on hold until my species solves that problem'. His species was incapable of meaning anything at all by '*gavagai*' at that stage, so it seems odd to explain how they developed that theory by positing a hyper-theory of meaning.

Concession **(2)** fares no better:

> Intractable Bayesian computations are not generally tractably approximable. This is not to say, of course, that cognitive algorithms do not approximate Bayesian computations, but rather to claim that approximation by itself cannot guarantee tractability. (Kwisthout et al., 2011, 780)

The above authors argue that approximation is only tractable in cases where the representation of the problem provides constraint and thus tractability. A similar, weaker claim is made by Griffiths and Tenenbaum (2009),

who say that theories constrain representational structures. For a set of four entities with no assumptions about causal relationships, there are 4096 possible causal graphs linking them. If we know that two entities are germs and two are symptoms, then there are only 16 possible graphs. So a principle at the theory level (that germs cause symptoms) constrains the number of structures in $\mathcal{S}$.

But the current problem concerns situations where the theory itself, or critical constraining concepts in the theory (such as GERM) don't exist psychologically. The symbolic threshold is just such a case. In such novel contexts, Bayesian approximations thus cannot be the only explanations of representational structures, since representational structures are what make these approximations possible. A hierarchical Bayesian approach may justify why one particular taxonomic arrangement of all the animals in fig. 3.3 is more likely to explain the data than all other possible arrangements, but this involves mechanical rearrangement of the same things. It does not involve the introduction of entirely new things, nor entirely new ways of structuring those things. Otherwise, it would be hard to see what Linnaeus adds to Aristotle. Bayesianism can use the idea of germs being present in food to limit the possible trajectory of diseases in fig. 3.5, but not explain where the concept GERM comes from. Chater and Oaksford (2008) admit that algorithms approximating probability are likely to work only for domains for which the brain has a dedicated module. Prior to the symbolic threshold, there couldn't have been a module for symbolic meaning. What's more, theories are one way of solving the relevance problem of the previous subsection. If the theories aren't there (or the relevant concept, such as GERM isn't there), then the relevance problem is unavoidable.

The addition of new content is a crucial difference between abduction and induction (§3.3.1). Unlike the case with induction, there is no shortage of abductive (or analogy- or insight-based) analyses of theory changes or the creation of new concepts that play constraining roles in theories: Gentner and Markman (1997) and Myrstad (2004) on Kepler, Rosenberg (1974) on Newton and Einstein, Paavola (2006) and Weisberg (2009) on Semmelweis' germ hypothesis and Paavola (2004) on Darwin's theory of evolution. These analyses all demonstrate the role of abduction, insight or analogy in particular momentous discoveries, while Dunbar (1996) demonstrates the role of analogy and insight in more modest discoveries in biology laboratories.

Rosenberg (1974) and Aliseda (2004) argue that abduction plays a role in coming up with new explanations in science in general, and Holyoak and Thagard (1995) do the same for analogy. I eventually aim to show that these play a role in guessing the meaning of novel signs.

Given that discovery is non-rational, a much more sensible response to the problem of no-theories is to abandon the idea that it's Bayes all the way down and to allow that non-probabilistic associative mechanisms could create novel concepts, connections between concepts, or larger structures. Research on creativity abounds with such processes (Mednick, 1962; Abraham and Windmann, 2007; Hélie and Sun, 2010; Thagard and Stewart, 2011) whereas inductive accounts of creativity merely involve novel rearrangements of familiar units based on probability distributions (for instance Jern and Kemp, 2013, who give a probabilistic account of creativity in as far as its involved in putting different foods together to create new kinds of salad). In a detailed review of creativity research, Abraham and Windmann (2007) argue that novel combinations of familiar elements are not very creative at all, in the grander scheme of things.

My position here accords with a range of previous research. Fodor (2001) argues that syntactic processes cannot in principle cope with contextually unconstrained inferences such as abduction[21], which entails that rationalist approaches in general cannot. Since I argued that abduction is non-normative (while Fodor assumes an IBE — inference to the best explanation — view of abduction), I don't think this is as worrying as Fodor does. Schurz (2008) claims that abduction introduces new representations while induction is what transfers them to new instances. Deutscher (2002) argues that abduction is what produces new levels of abstraction: these correspond to theories in the discussion here. 'In those cases where agents respond to new evidence by inventing new hypotheses, the Bayesian model is silent' (Okasha, 2000, 706-7).

> In the end, we are left trying to understand what animals do and how they do it. The hard problems remain inaccessible to the tools of Bayesian analysis, which merely provide a means to

[21] For instance, Fodor argues that simplicity is context-dependent and thus not explicable by any local syntactic property such as string length. Allowing it to be explained by a non-local syntactic property engages the Frame Problem, so no syntactic property at all will work.

> select an answer once the hard problem of specifying the list of possible answers has been solved (or at least prescribed). (Anderson, 2011, 190)

Opening the black box of discovery (to adopt a phrase from Paavola, 2006) requires recognising that hypothesis generation is not inductive, and seeing whether non-rationalist processes offer any suggestions.

3.5 Empiricist approaches

In this section, I explore empiricist (i.e. associative, non-normative) approaches to hypothesis generation. Three initial reasons for optimism here are **(1)** I've built inference upon minimal rationality, which assumes an empiricist psychology (§2.2.3). **(2)** I've already reviewed evidence for associative processes in context-deciding inference (§2.6). **(3)** Abduction is not logically valid (i.e. truth-preserving) or optimal (i.e. probability enhancing) so there's no need to shoehorn it into a normative framework (§3.3.3).

I investigate two processes which can produce hypotheses in an empiricist manner: analogy and insight. The two are not always clearly distinct. Christie and Gentner (2010), for instance, describe participants having a 'relational insight' while solving an analogy problem and Gentner and Markman (1997) discuss how comparison fosters insight. Further, I will frame analogy in terms of structural similarity and insight in terms of semantic structures and Graham and Kilbreath (2007) review evidence showing that judgements of similarity can be sensitive to semantic categories and vice versa. Ansburg (2000), in an individual-differences study, shows that success at insight problems is strongly correlated with success at verbal analogies.

Both insight and analogy involve novel connections between two representations (or two representational structures), and because both involve novelty, they are not based on learnt probabilities. The discussion here has some features in common with hierarchical Bayesian models in that it posits representational structures that guide inference in novel cases, though I argued that probabilities alone cannot explain those structures, novel hypotheses or novel theories in open-ended contexts. The literature on insight and analogy explicitly addresses key issues raised throughout this chapter: how representational structures help determine similarity (Day and Gentner,

2007), creativity (Mednick, 1962), salience (Gentner and Markman, 1997) and relevance (Ross and Bradshaw, 1994).

3.5.1 Analogy

Analogy involves associations (or mappings) from sources (or bases) in prior experience to targets currently being interpreted. Here I'll outline how analogy can in principle provide hypotheses in pragmatic inference, but since it's not clear what the source analog at the symbolic threshold would have been, I will quickly move onto insight for concrete predictions.

Structure Mapping Theory (SMT, Gentner, 1983, 2010; Gentner and Markman, 1997) proposes that analogy involves mapping, or the formation of novel connections, between structured representations. The meanings underlying both 'The dog chased the cat' and 'Umbrella Corp. made a takeover bid for Acme Co.' involve structured representations, something like CHASE(DOG,CAT) and TAKE-OVER(UMBRELLA,ACME). Connections can be drawn between objects (DOG ~ UMBRELLA; CAT ~ ACME) and predicates (CHASE ~ TAKE-OVER). Higher-order mappings are possible between large-scale causal structures, such as narratives[22].

The process of structure mapping involves a number of steps. First, all possible mapping between elements are made. These may involve inconsistencies or many-to-one mappings. Second, these elemental mappings coalesce into more structured clusters, called kernels. Third, the kernels are merged into structures that are consistent according to theoretical principles (Gentner and Markman, 1997; Gentner and Medina, 1998). For instance, compare CHASE(CAT_i, MOUSE) with CHASE(DOG, CAT_j). At the first stage, relational mappings are possible between CAT_i and DOG because both do the chasing. At the same time, object mappings are possible between CAT_i and CAT_j. One principle of SMT is that mappings which reflect the argument structure of predicates are preferred, so a more consistent mapping is from CAT_i to DOG. One-to-one mappings are similarly preferred to many-to-one

[22]Gentner and Markman (1997) say that analogy should focus more on predicates and higher-order relations and less on mappings between objects. The distinction is intended to highlight the fact that analogy allows for abstract cognition: Kepler, for instance, drew an analogy between the motive force of the sun on the planets and the effects of a whirlpool on boats: there are no object matches here. Nonetheless, a range of research shows that object mappings play a role in pragmatic inference (Ross and Bradshaw, 1994; Catrambone, 2002), so I won't exclude these from discussion here.

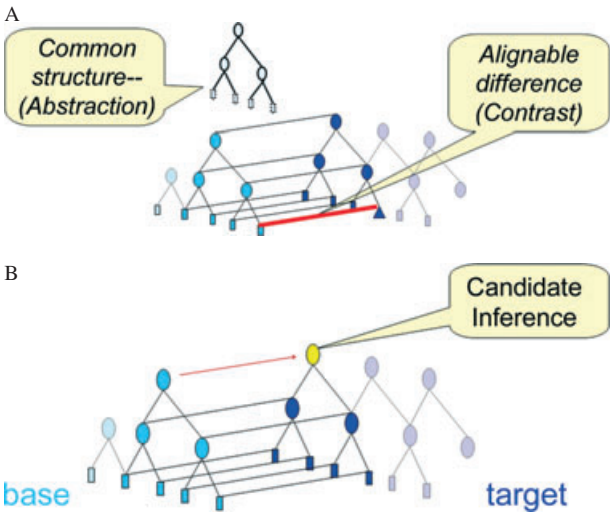


Figure 3.9: Structure mapping promotes (A) abstraction and salience; (B) further inferences (Gentner, 2010).

mappings and relational mappings consistent with higher-order or systematic mappings are preferred to relational mappings without higher-order mappings.

Once the most consistent mapping is chosen, the stage is set for further abstract cognition: inferences can be drawn from the base to the target; the target can be reinterpreted or its structure reconfigured to promote further consistency; abstractions can be drawn; and attention can be drawn to salient similarities or differences (Gentner and Markman, 1997; Gentner, 2010, see fig. 3.9). Reinterpretation and reconfiguration will be discussed below under the heading 'Insight'; the rest are explained below. As an example of the first point, the information that a dog is likely to damage a cat might prompt the hypothesis that Umbrella Corp. might have an adverse effect on Acme Co. Day and Gentner (2007) demonstrate this effect experimentally and Colhoun and Gentner (2009) show that SMT predicts human patterns of inductive inferences in causal situations.

This seems more like plausibility than probability (as I discussed in relation to the metaphor 'The investors were squirrels', §3.3.6), given that the probability of one company damaging the other is rationally independent of

the probability of a dog damaging a cat. SMT deals with novel situations, not by deriving priors from theories, but by mapping comparatively familiar or richly represented bases to comparatively unfamiliar or sparse targets, and then using those mappings to transfer information from the former to the latter. Christie and Gentner (2010) show in a child word-learning experiment that analogy can affect hypotheses generation independently of cross-situational probabilities.

Obviously, we don't constantly compare each representational structure with all others all the time: comparison between two objects or events is often prompted by something in experience, such as spatial juxtaposition (Christie and Gentner, 2010) or their being labelled with the same novel sign (Graham and Kilbreath, 2007). SMT seems compatible with whatever processes underlie the joint attentional scenes discussed in §1.5.1.2 (for instance, the Hungarian train tickets, Tomasello, 1999). If the tourist maps the Hungarian scene onto the familiar scene, then joint attention can prompt inference from base to target, where the inference is an hypotheses about the meaning of the Hungarian word. A range of studies thus shows that analogy can play an abductive role in word learning.

SMT also offers salience-deciding processes (Gentner and Markman, 1997). Compare fig. 3.10 (*a*) and (*b*). In both pictures, the boy is looking at something. This prompts a predicate mapping between LOOK(BOY,SNAKE) and LOOK(BOY,FISH). An object mapping is also possible between the dressers in each picture. According SMT's systematicity principle, the dresser is less salient in analogy because it is merely an object mapping, which is dispreferred to predicate mappings involving LOOK.

Further, analogy involves alignment of two structures, and alignable differences are more salient than non-alignable ones (Gentner and Markman, 1997). Alignable differences occur when different elements in each structure play the same role: the fish and the snake in fig. 3.10 are alignable differences because they play the same role in predicate LOOK(x, y). Non-alignable differences occur when an element of one structure has no correspondence in the other: the dog in fig. 3.10 (*a*) corresponds to nothing in (*b*). Gentner and Markman (1997) review evidence for the comparative salience of alignable differences. For instance, people are able to list more differences between hotels and motels than they are between magazines and kittens because the first pair involve many alignable differences (hotels have several stories, mo-

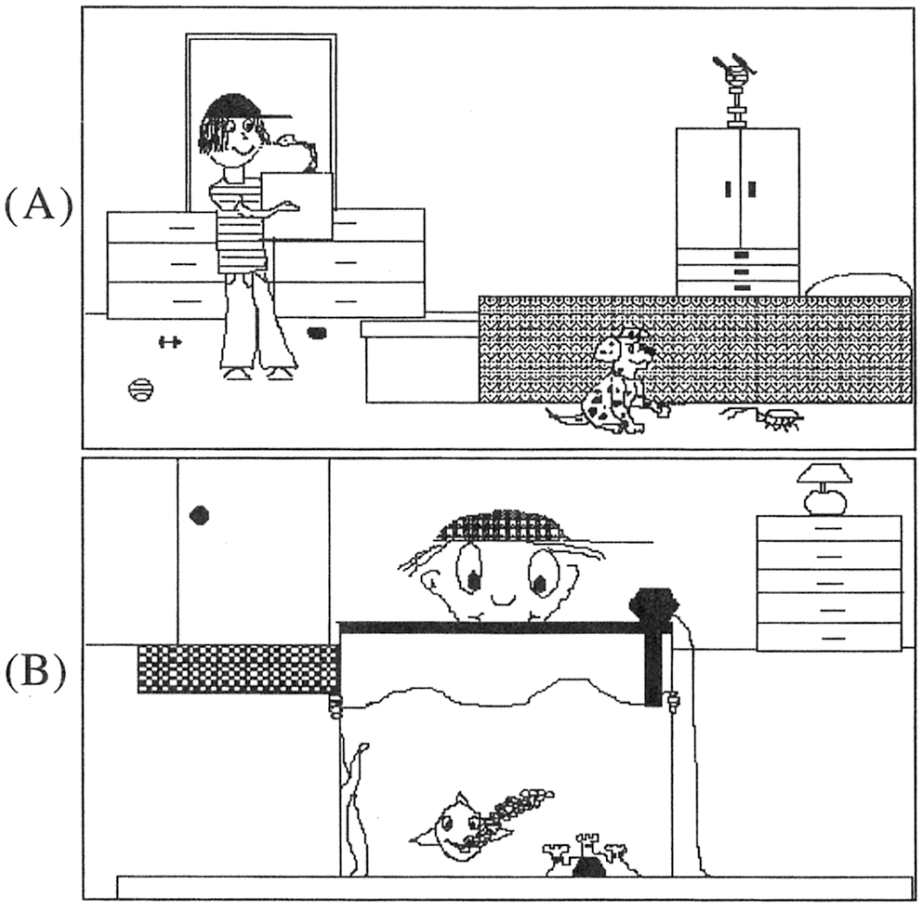


Figure 3.10: Alignable and non-alignable differences based on the predicate LOOK(x, y) (Gentner and Markman, 1997).

tels just one or two). While the second pair has more differences overall, these are not alignable and thus less salient.

In early work, Gentner (1983) claims that SMT operates over syntactic, not semantic information. I think, though, by 'syntactic' she is focusing on the fact that analogy maps predicates onto each other: dogs and Umbrella Corp. have nothing in common semantically, but play a similar role in the structured representation they find themselves in. There is a semantic link between predicates CHASE and TAKE-OVER, though, so I don't think that Gentner's sense of 'syntax' precludes semantic information from explaining how analogy plays an abductive role. Ross and Bradshaw (1994) presented participants with ambiguous texts, and found that associative object mapping affects interpretation of the ambiguity. Similarly, semantic relationships affect recall in analogy (Catrambone, 2002). In as far as recall informs accessibility, which underpins relevance-deciding inference, these results show that semantic or associative information allows similarity or analogy to play an abductive role.

I have been focusing on the work of Gentner and colleagues, but alternatives to SMT exist. A major milestone in analogy research is Holyoak and Thagard (1995). They note that analogy is creative: it plays a role in dealing with novelty because it tries to understand the novel in terms of the familiar, a process that often involves a mental leap or the formation of a novel connection between representations. They describe this as a spark jumping across a gap, an image I'll explore further under 'Insight'. In their view, analogy is not logical, but not haphazard either: they say it is subject to something like logic which they call analogic. I argued that since abduction is not logical, it is psychological. Their description does make analogic seem psychological, and they state that the 'rules' of analogy are not rigid, but relative to a person's goals, for instance. Indeed, this addition of purpose is a major difference between their approach and SMT, which is purpose-blind. Further, they state explicitly that 'analogy is a source of possible *conjectures*, not guaranteed conclusions' (1995, 30, emphasis mine).

Like SMT, their work supposes systematic mappings. They are less dismissive of object mappings than SMT is, though. While SMT downplays the role of semantic connections between representations, Holyoak and Thagard believe that '[t]he semantic connections between concepts provide important building blocks for seeing analogies' (1995, 23). In this regard, they present

a model of representational structure that accords very closely with my view of plausibility (§3.3.6). This structure is one way of deciding what counts as relevant in a mapping, much like accessibility in Sperber and Wilson (1995). They are, however, more interested in the evolutionary question of how humans evolved higher-order representations than they are in how we evolved context-deciding inferences, which is unsurprising given that Holyoak is one of the authors in Penn et al. (2008, cf. §2.5.2.3 above). They also posit a dual-process model for analogy, combining elements of associative and syntactic processing. It is thus clear that their view of analogy is compatible with it playing an abductive role, and elsewhere one of the authors of Penn et al. explicitly makes such a connection (Thagard, 2007).

3.5.2 Insight

Like analogy, insight is a matter of connections between representations. Weak or distant connections can be strengthened in insight problem solving, or novel connections can be formed: insight is creative (Mednick, 1962; Kounios and Beeman, 2009; Cushen and Wiley, 2011). An accumulation of novel or newly strengthened connections in a representation can mean that one's representation of the problem has been restructured entirely[23], as illustrated in an experiment by Durso et al. (1994). Participants had to come up with an explanation for an unusual scenario: 'A man walks into a bar and asks for a glass of water. The bartender points a shotgun at the man. The man says, "Thank you," and walks out' (1994, 95). Typical features of solving an insight problem such as this are as follows (Jung-Beeman et al., 2004; Bowden et al., 2005; Gilhooly and Murphy, 2005; Kounios et al., 2006; Kwisthout, 2012, variously):

1 Participants reach an impasse or 'blank' and are unable to progress. Then they suddenly realise the answer.

2 Once the relevant dimensions are known, the answer is obvious; but until they are known, it is non-obvious.

[23]Early or Gestaltist approaches to insight insist that it requires such restructuring, but this has been downplayed somewhat in much modern work. Since restructuring is just the result of a number of novel individual connections, we are dealing with a difference of scale, not type: both involve representational change. Hélie and Sun (2010) acknowledge the existence of a continuum here.

3 Participants usually cannot report the process by which they reach the answer.

4 They typically arrive at the solution all at once rather than through a step-by-step mechanical process, so insight problem solving is not accompanied by a feeling of increasing 'warmth' or closeness to solution.

5 They usually report feeling a flash of surprise, an 'Aha!' or a 'Eureka!' moment, much like a cartoon light bulb going off in the head.

The solution to the above problem is that the man had hiccoughs. Drinking water is one cure for hiccoughs. Receiving a fright is another, hence the barman's behaviour. Durso et al. (1994) divided participants into two groups: solvers (those who discovered the solution within 2 hours) and non-solvers (those who failed to do so). After solving the problem or failing to do so within the time limit, all participants rated the relatedness of pairs of concepts in the solution (e.g. REMEDY), or in the scenario (e.g. BARMAN) or in a bar generally (e.g. PRETZELS). Graphs of these concept associations were abstracted over each of the two groups (fig. 3.11). These graphs show that solvers and non-solvers had different representations of the problem. Solvers had formed connections that non-solvers did not, such as between REMEDY and GLASS-OF-WATER or RELIEVED and THANK-YOU.

Their claim that insight involves connections between representations coheres remarkably with Peirce's description of abduction:

> It is an act of insight, although of extremely fallible insight. It is true that the different elements of the hypothesis were in our minds before; but it is the idea of putting together what we had never before dreamed of putting together which flashes the new suggestion before our contemplation. (CP 5.181, 1903)

Another example of an insight problem[24] is a Compound Remote Associates (CRA) task. Here, participants are given three words, such as

[24]Insight is a feature of people's solutions to problems, not of problems themselves, so by 'insight problem' I mean a problem that typically is solved by insight processes. In addition, some people may solve a particular problem with insight, others might solve it more mechanically. Indeed, one person might solve a given problem with insight and another example of the same type mechanically. The CRA task here could be solved mechanically by listing all possible collocations of these words and then searching the lists for common elements. This is presumably how we'd get a computer to do it. Humans solve it mechanically some of the time, but it is typically solved with insight, according to the evidence below.

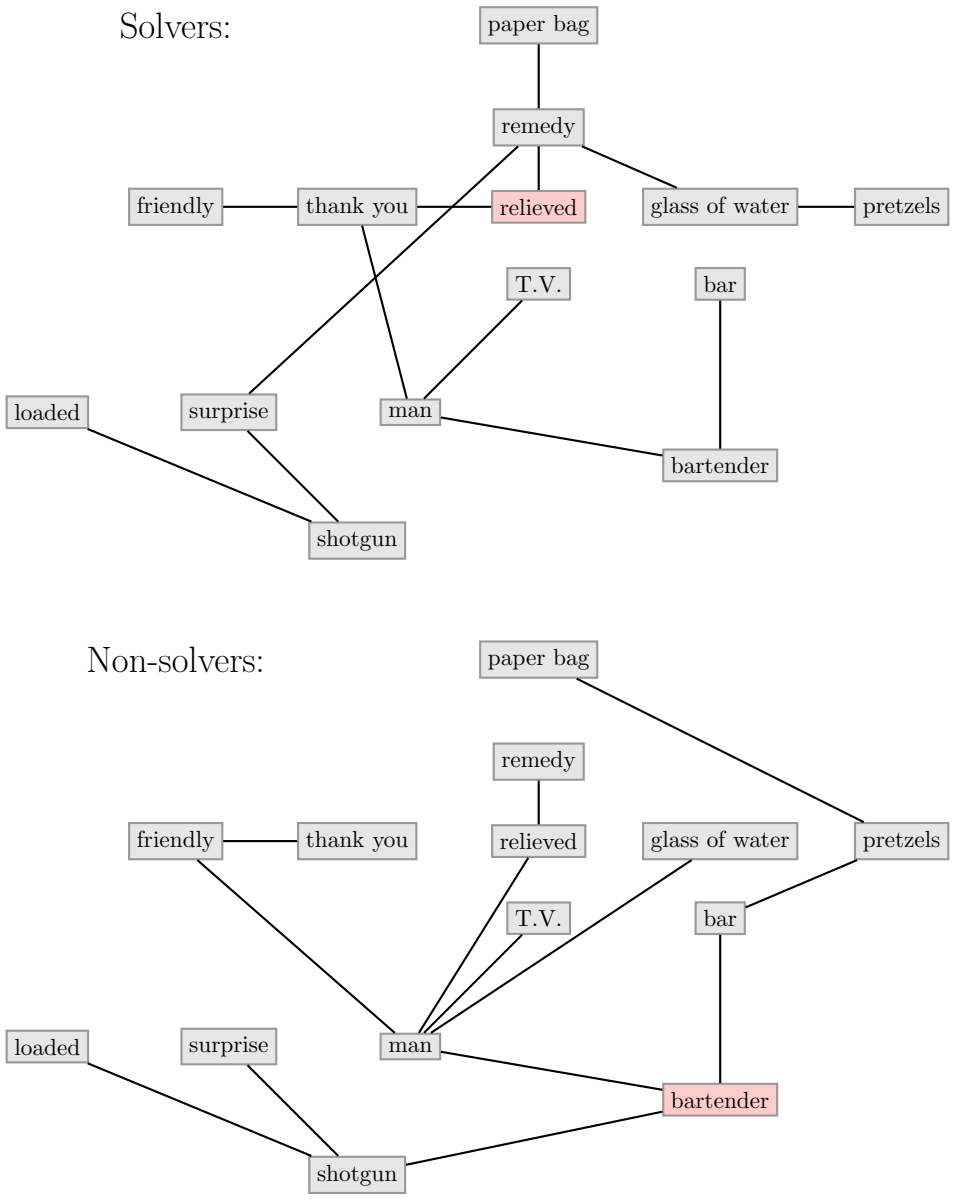


Figure 3.11: Graphs based on conceptual association showing that solvers (top) and non-solvers (bottom) had different representations of the problem. The highlighted node indicates the centre of each graph. Redrawn from Durso et al. (1994).

PINE
SAUCE
CRAB

and are asked to think of a fourth word that could be placed before or after all three. In this case, the answer is 'apple'. This doesn't require restructuring as such (though fixation on the most common collocations, such as 'pine cone', would lead to an impasse), but it does require recognition of distant associations and is usually described in terms of automatic spreading activation over semantic representations (Jung-Beeman et al., 2004; Hélie and Sun, 2010)[25]: compare this to the FOOT/GLASS/CUT example or John crying out while swimming near glass (§2.6.3).

'One critical cognitive process distinguishing insight solutions from non-insight solutions is that solving with insight requires solvers to recognize distant or novel semantic (or associative) relations; hence, insight-specific neural activity should reflect that process' (Jung-Beeman et al., 2004, 501). The authors provide two pieces of evidence for insight-specific neural activity.

In an fMRI experiment, participants were asked to solve CRA problems, and report whether they felt a flash of insight, as described above. Those reporting such a flash showed significantly more activation in the anterior Superior Temporal Gyrus in the right hemisphere (aSTG-RH) than those reporting no insight. The role of the STG-RH (and RH language processing in general) has already been discussed (§2.6): these areas are involved in distant semantic associates, and I mentioned how this was a matter of contextually unconstrained pragmatic inference. This result is not explicable by the emotional effect of having an 'Aha!' moment because the activation is present before solution. Behavioural evidence from a split-visual-field priming study supports this neurological evidence (Bowden and Jung-Beeman, 1998). In addition, the fact that this is a priming study supports the claim that insight involves spreading activation across semantic representations.

In a second experiment, Jung-Beeman et al. measured EEG responses while participants performed the same task, since these provide information about the time course of the relevant cognitive processes. The results showed a burst of gamma-band activity (implicated in transitions from non-

[25] In the case of the latter, this is explicitly situated in a dual-process framework.

awareness to awareness) in the anterior right temporal lobe for insight solutions, but not for non-insight solutions. This burst occurred approximately 0.3 seconds before participants pushed the response button, supporting the claim that insight involves sudden solution rather than a step-by-step mechanical process.

Kounios et al. (2006) show, in EEG and fMRI studies, that participants' brain states predict whether they will solve such problems insightfully or not, *before* they even see the problem. In particular, they found increased activation of the anterior cingulate cortex (ACC) before insight problem solving. The results support the description of strategy switching in the Mutilated Checkerboard problem above (§3.4.3).

> [T]he activity observed in ACC prior to insight may reflect increased readiness to monitor for competing responses, and to apply cognitive control mechanisms as needed to (a) suppress extraneous thoughts, (b) initially select prepotent[26] solution spaces or strategies, and, if these prove ineffective, (c) subsequently shift attention to a nonprepotent solution or strategy. Such shifts are characteristic of insight. (Kounios et al., 2006, 887)

Gilhooly and Murphy (2005) performed an individual-differences study, giving participants insight and non-insight problems to solve and various tests of cognitive abilities: Raven's matrices test of general intelligence, a number of verbal and spatial ability tests, as well as tests of cognitive flexibility. They found that insight problems tended to cluster with other insight problems and non-insight problems tended to cluster with non-insight problems. General intelligence was predictive of non-insight task success, but not of success in insight tasks. Measures of cognitive flexibility were important for insight task solving, but not non-insight tasks. So insight problems are cognitively different from non-insight problems. A further individual-differences study (Schooler and Melcher, 1995) showed that individuals' ability to break out of the initial context of a problem predicted insight problem solving ability, as opposed to non-insight problems.

Ansburg and Hill (2003) also conducted an individual-differences study. Participants were given tests of creative or insightful ability (remote asso-

[26] By this, they seem to mean 'predictable' or 'likely'. Cf. my discussion in §2.6.4 of how LH systems dominate when predictable relations are sufficient, but RH systems play an increasing role as broader semantic fields are needed.

ciation tasks, like those discussed above) and tests of analytic or deductive ability. They were then required to memorise printed word lists (focal cues) while a second, recorded word list was audible in the background (peripheral cues). Unbeknownst to the participants, some focal and peripheral cues were solutions to a later anagram problem. Participants capable of diffuse attention (both focal and peripheral cues) should thus prove more successful at the later anagram problems than those only capable of focal attention. Success at the creative tasks predicted ability to use peripheral cues in solving the anagrams; success at analytic or deductive tasks didn't. '[I]ndividuals who cast broad attentional nets are more likely to capture unexpected cues and to generate remote associations than are those whose cognitive resources are more narrowly focused' (Ansburg and Hill, 2003, 1142), and their results show that insightful individuals are those capable of such diffuse attention. I discussed the relationship between broad attention and hemispheric differences in contextually unconstrained inference in §2.6. Murray and Byrne (2005) tested individual differences in attention (among other things). They found that the ability to switch attention correlates with insight success; focused (selective, sustained) attention does not. Kounios and Beeman (2009) found an alpha band burst over the right occipital cortex immediately prior to the gamma band response discussed above, which they interpret to mean a suppression of focal attention.

Coren (1995) gave left- and right-handed participants tests of divergent thinking, such as combining familiar objects to make a new object with a novel function.

> It is called divergent because it often involves the consideration of several different directions, alternatives, or information sources. Divergent thinkers seem more capable of breaking sets and achieving novel solutions. For this reason divergent thinking is often listed as a major component of the psychological trait of creativity. (Coren, 1995, 313)

These terms are highly reminiscent of previous examples: broad attentional nets and breaking out of initial representations. Left-handed males scored higher than right-handed males at the divergent thinking tests, and the authors tentatively link left-handedness with a reduction in the usual leftward hemispheric asymmetry compared to right-handers. That is, left-

handers' RHs are less reduced than those of right-handers. Abdullaev and Posner (1997) show that divergent thinking is associated with increased activation in RH homologues of Wernicke's area.

Kwisthout (2012) explicitly links insight to relevance, and I have previously discussed his claims about the computational intractability of the relevance problem in abstraction (§3.4.5). He offers a slightly different (but compatible) definition of an insight problem: he considers them to be problems which are comparatively easy to solve once the relevant dimensions of the problem are known, or once the representation of the problem includes the relevant information. Finding the relevant representation is the hard part. The impasse is reached when the original representation of the problem does not contain the relevant information. Restructuring the representation is what replaces irrelevant with relevant information. Kwisthout's notion of relevance is compatible with that of Sperber and Wilson (1995). Familiarity with a given problem can constrain relevance, but misclassification of the problem leads to impasse. Novel problems (such as novel symbols) preclude familiarity, hence the difficulty of the relevance problem and the need for insight in these cases. Allott (2013) characterises hypothesis generation in Sperber and Wilson (1995) as a step-by-step mechanical process, and I think that is an accurate description of their theory, but the evidence suggests that their theory is thus not an accurate description of pragmatic inference in novel contexts.

To take an example from a different study (Luo and Niki, 2003), think of something that can move heavy logs but not a small nail. The initial representation of the problem makes certain information salient (here, weight). Weight, however, turns out to be an irrelevant dimension. The relevant dimension is density, because the answer is 'a river' in that it can usefully transport logs from one place to another while a nail would simply sink. Searching for a solution based on the initial interpretation cannot lead to the solution, hence the impasse. Nor is the relevant concept DENSITY analytically derivable from the statement of the problem; it requires a creative leap to decide what is relevant. Another example, from Murray and Byrne (2005) is this: how would you throw a ping pong ball so that travels a short distance, stops on its own, and returns? You cannot bounce it or tie anything to it. People typically have two problems with relevance here: they misconstrue the dimension of travel as horizontal when it should be vertical,

and they do not realise that gravity is relevant: you simply throw the ball directly upwards.

Before returning to the problem of word learning in novel cases, I'll quickly list a few other results that demonstrate that insight problem solving as a distinct mode of cognition is a psychological reality. Schooler et al. (1993) and Schooler and Melcher (1995) showed that verbalising one's thinking while trying to solve a problem disrupts insight problem solving, but not non-insight problem solving. Wagner et al. (2004) shows that sleep greatly facilitates insight problem solving. Jausovec and Bakracevic (1995) measured participants' heart rates in insight and non-insight problem solving, and found that they increased gradually and steadily during non-insight problem solving, but increased suddenly just before solution in insight problem solving, matching the step-by-step mechanical nature of the former and sudden-flash nature of the latter. Cushen and Wiley (2011) argued that bilinguals should perform better at insight problems, given that they have at least two words for many concepts, so each concept has a broader or more diverse range of associations, and insight involves activation over diverse associations. They found that bilinguals performed better than monolinguals at insight problems; the reverse was true for non-insight problems.

Returning now to abduction and induction, insight is a matter of spreading activation over a semantic network, hence a matter for an empiricist psychology. Because it requires diffuse attention or shifts in attention and representational changes, it is a context- and relevance-deciding process. Further, it is involved in creativity, and centres on brain regions I showed to be important for contextually unconstrained pragmatic inference. These features mean it is potentially suitable for playing an abductive role as Peirce claims. Consequently, I aim to show experimentally that abduction is in fact insightful, and that insight (and thus abduction) play a role in communicative situations sharing features with the symbolic threshold.

On the other hand, no amount of searching will alight on the solution if one has misrepresented the problem, so insight is not a matter of inductively searching an hypothesis space to choose the most likely explanation. Indeed, it is hard to see how the hiccough-shotgun problem above is explicable inductively: one may represent both frights and water as very likely remedies for hiccoughs, but that doesn't mean one comes up with the explanation for that scenario by evaluating such probabilities. The problem

is the representation itself: making a novel connection or strengthening a weak connection between WATER and REMEDY. Murray and Byrne (2005) characterise insight problems as being inherently ill defined, but induction needs a well defined hypothesis space (§3.4.2), so induction is ill equipped for insight. CRA problems, on the other hand, are *potentially* explicable inductively, but the neurological evidence discussed above shows that this is not how humans typically solve such problems.

3.5.3 Evolutionary background: evidence from animals and children

My description here coheres with the claim in Deacon (1997) that insight was crucial at the symbolic threshold, though Deacon doesn't discuss the psychological realities of insight or abduction. Indeed, he focuses entirely on the prefrontal cortex, while evidence reviewed above has focused on the RH temporal lobe. Similarly, he focuses on higher-order relations, while I focus on context.

Human insight doesn't require an evolutionary saltation, since animals are capable of insight problem solving. The difference is that they do so less spontaneously than we do, and do so domain specifically: I don't know of any evidence that doesn't relate to food retrieval. I've discussed evidence from Köhler (1927) of insight problem solving in apes stacking boxes to reach food, but this is anecdotal (§2.5.1). Foerder et al. (2011) provide similar evidence for elephants in a comparable task. Conflicting results emerge from experiments: Mendes et al. (2007) found orangutans to be highly successful in the Floating Peanut Task[27], while Hanus et al. (2011) found low solution rates in chimpanzees and gorillas. Apart from mammals, there is much evidence of insightful behaviour in corvids (e.g. Bird and Emmery, 2009). Rooks spontaneously learnt to solve a variant of the Floating Peanut Task by dropping rocks into a tube with a floating worm, such that the water level would rise and they could reach it with their beaks. They also learnt to use big rather than small rocks, which is more efficient.

Taylor and Gray (2009) argue that this doesn't involve human-like knowledge of causation, because if it did, the birds would have used only large rocks initially. However, I think this may overestimate the causal knowledge

[27] A peanut is floating out of reach in a tube. The problem is solved by spitting water into the tube to raise the water level.

of children. The only study I know that tests children's ability in the Floating Peanut Task (Hanus et al., 2011) used water poured from a pitcher, not rocks, so we have no evidence to suggest that children would start with large rocks. In the youngest age group of children tested by Hanus et al. (4 year olds), only 2 out of 24 were successful, so the bar is set rather low. In one of the experiment with chimpanzees, the success rate was 2 out of 19. Anyway, Taylor and Gray (2009) seem to be assuming that a rationalist account must explain this, while I've been arguing that insight is in fact empiricist.

Hanus et al. argue that functional fixedness (the inability to use a familiar object for a new purpose) underlay low success rates among the primates, while functional flexibility is a feature of creativity and insight in humans (§3.5.2). The experimenters compared children's success in a wet condition (water already in the tube) and dry condition (no water in the tube). Across all age groups, children in the wet condition were more likely to solve the problem. This suggests an effect of relevance: water was associated with the problem in the wet condition, but not the dry condition, where a more distant connection had to be made to realise its relevance. Primates performed badly in both conditions, so perhaps they struggle with basic problems of relevance. It would be interesting to see what would happen if the water source was made salient to them, as empty feeding tubes were in Savage-Rumbaugh and Rumbaugh (1978), which I discussed in relation to context and symbol learning (§2.5.2.2).

Taylor and Gray (2009) call insight a 'rather murky' term, but I hope I have shown that it is not, though much still needs to be done to understand it properly. Anyway, their understanding seems to be based on texts by biologists, not psychologists. Taylor and Gray also consider (and, again, reject) the possibility that the rooks are 'Popperian creatures' (Dennett, 1994), ones capable of generating hypotheses[28]. A proper understanding of inference as it relates to insight explains just how these concepts are related, and I think that's just what my inferential hierarchy and discussion of abduction have done.

The question of animal representations is a contentious one, though, and the above studies do not investigate just how animals represent these prob-

[28]Though this is not how they phrase it, and Dennett's original formulation focuses on preselected hypotheses, neglecting Popper's context of discovery, for which see §3.3.4 above

lems. It is therefore worth seeing (briefly) whether there is evidence for any abilities related to insight in animals that doesn't require knowing more about their representations. This would help avoid any accusations of evolutionary saltation. In §3.5.2, I discussed results from Ansburg and Hill (2003) showing that insight ability correlates with diffuse attention in humans, and I quoted the authors as claiming that diffuse attention allows people to 'capture unexpected cues and to generate remote associations than are those whose cognitive resources are more narrowly focused' (2003, 1142). One potential pre-adaptation for insight, then, would be a hemispheric differences between focal and diffuse attention, such that the RH is more successful at diffuse attention and the LH at focal attention.

By exposing chicks to light while in the egg, Rogers et al. (2004) were able to induce hemispheric lateralisation. Both lateralised and unlateralised chicks pecked at grains of food mixed in with pebbles while, periodically, a model resembling the silhouette of a predatory bird passed above them. Pecking at food required focal attention; watching out for predators required broad or diffuse attention. Lateralised chicks were better at distinguishing grain from pebbles than unlateralised chicks, and were also more likely to detect the predator stimulus. Indeed, lateralised chicks were able to do both simultaneously. In lateralised chicks, it was the right eye (LH) that performed best at the focal task and the left eye (RH) that performed best at the diffuse task.

Reviewing lateralisation in vertebrates generally, McGilchrist concludes that:

> Lateralization brings evolutionary advantages, particularly in carrying out dual-attention tasks. In general terms, the left hemisphere yields narrow, focused attention, mainly for the purpose of getting and feeding. The right hemisphere yields a broad, vigilant attention, the purpose of which appears to be awareness of signals from the surroundings, especially of other creatures, who are potential predators or potential mates, foes, or friends; and it is involved in bonding in social animals. (McGilchrist, 2010, 505)

3.6 Conclusions

In the previous chapter, I set out the inferential hierarchy, arguing that contextual constraint is a major determiner of inferential complexity, and that empiricist processes (such as spreading activation over coarse semantic networks in the RH) underlie some contextually unconstrained inference, especially in pragmatic interpretation.

In this chapter I've defined abduction as a type of inference that generates hypotheses. It is conjectural and can involve creative discoveries or the selection of familiar hypotheses. The latter may *seem* similar to induction, but the former is quite different, and it is the former that is relevant to understanding novel signs. Abduction is a matter of an empiricist psychology in that it is non-normative and explicable by associative processes, while induction is rationalist in that it is normative (or seeks to approximate normative computational-level functions) and often described in syntactic terms at the computational level, though these may be implemented associatively at the algorithmic level.

I showed that abduction cannot be reduced to induction because induction cannot cope with novelty or unconstrained hypothesis spaces, or the question of how a relevant representation of the problem is arrived at. Abduction is what provides hypothesis spaces for induction to evaluate, so it is logically prior to induction and a necessary complement to it.

I briefly suggested a plausibility-based view of abduction, but looked in more detail at how analogy and insight meet the requirements (being empiricist, able to handle creativity or novelty, and able to decide relevance, salience or context).

Importantly, I discussed ways of diagnosing whether insight problem solving is applied to a given problem on a particular occasion, though now I'll spell this out in a little more detail. Jung-Beeman et al. (2004) showed that subjective reporting of a feeling of insight correlates with objective measurements of insight problem solving: the brains of people who reported feeling a flash of insight when they solved a CRA problem showed increased activation in the STG-RH, which I linked in the previous chapter to contextually unconstrained pragmatic inference over coarse semantic networks. I also provided a range of behavioural, physiological and neurological evidence showing that insight problem solving is distinct from non-insight mechani-

cal or analytic processes. Insight, then, is a distinct cognitive process and subjective reporting is a reliable indicator of insight. Using subjective reporting in a word learning task can thus indicate whether or to what extent the problem is solved insightfully.

In the next chapter, I fine-tune subjective reporting as a diagnostic tool: insight has a number of features, and I show that it is the 'Aha!' experience that best distinguishes insight from analytic thinking. I then use this diagnostic to show that abduction is more insightful than induction.

In the following three chapters, I isolate features of communication at the symbolic threshold and test to what extent insight (and thus abduction) are implicated in communicative tasks displaying those features.

The first is that communication at the symbolic threshold would have involved large worlds rather than small worlds. The latter involve a small set of possible answers; the former a vast set (much like charades). Inductive accounts assume small worlds, but I predict that large worlds require increased amount of insight (and thus abduction).

The second is that communication at the symbolic threshold would have involved less contextual information: intended meanings would have been less predictable from the environment than is the case in modern child word learning. I predict that less predictable cases require increased insight.

The third is that signals at the symbolic threshold were novel, not conventional, and I predict that more novel signals require more insight. I also show that level of iconicity is not a predictor of insight levels, though much current work in symbol origins emphasises the role of iconicity.

Part II

Experiments

Chapter 4

Diagnostics of Abductive Inference

4.1 Background and Aims

I have been arguing that crossing the symbolic threshold involved context-, salience- and relevance-deciding inference about the ground of a sign. In the previous chapter, I argued that abduction is just such a kind of inference, and that it generates a focal set of hypotheses in novel situations which induction can then evaluate. I also argued that abduction requires an empiricist (non-normative, associationist) psychology. I reviewed features of insight suggesting that it meets these requirements and can thus potentially play an abductive role. In this section, I show experimentally that abduction is in fact insightful, compared to non-insight or analytic problems (i.e. induction and deduction).

Insight problem solving involves distinct cognitive mechanisms compared to non-insight problem solving (§3.5.2). I reviewed behavioural, neurological and physiological evidence, but focused on subjective reporting in Jung-Beeman et al. (2004): an fMRI study found significantly more activation in the Superior Temporal Gyrus of the right hemisphere (STG-RH) for subjects who reported feeling a flash of insight while solving a Compound Remote Associate (CRA) test than participants who didn't. This result is consistent with other evidence concerning the function of this area: it is implicated in pragmatic inferences over broad semantic networks (§2.6). By analysing the time-course of neural activation in an EEG study, Jung-Beeman et al.

also excluded the possibility that subjective reporting of a flash of insight is explainable as an emotional response, equivalent to a general notion of surprise: activation peaked before the conscious 'Aha!' moment of surprise. So subjective reporting is a reliable diagnostic for insight problem solving as a distinct cognitive mechanism.

When explaining insight to their participants to prepare them for self-reporting, Jung-Beeman et al. (2004) describe insight using a range of features (cf. §3.5.2):

> A feeling of insight is a kind of 'Aha!' characterized by suddenness and obviousness. You may not be sure how you came up with the answer, but are relatively confident that it is correct without having to mentally check it. It is as though the answer came into mind all at once — when you first thought of the word, you simply knew it was the answer. This feeling does not have to be overwhelming, but should resemble what was just described. (Jung-Beeman et al., 2004, 507)

So although subjective reporting is diagnostic of insight problem solving, this reporting bundles together a number of potentially independent features. It is not guaranteed that all features work equally well to distinguish insight from non-insight problem solving. One aim of the present experiment is to evaluate each feature individually. Hence, I decomposed this description into four criteria[1], and allowed participants to report each dimension separately. Here, each criterion contains two descriptors: the first is the insight response; the second is the non-insight response. The word in brackets after each gives the label that will identify that variable in analysis.

Criterion 1 The answer came to me all at once **vs.** I worked towards the answer step-by-step (`sudden`)

Criterion 2 I wouldn't be able to explain how I got the answer **vs.** I would be able to explain how I got the answer (`explain`)

Criterion 3 I had an 'Aha!' moment, like a lighbulb flashing on **vs.** I had no 'Aha!' moment (`aha`)

[1] I listed five criteria at the end of the previous chapter. One was a matter of computational-level analysis rather than subjective experience, though, so it has been dropped here.

Criterion 4 I'm confident my answer is good without being told **vs.** I think it's possible my answer is bad (`confidence`)

Participants in Jung-Beeman et al. (2004) gave binary responses: they reported solving a problem with insight or without it. But Hélie and Sun (2010) stress that there is a continuum here: it could be that a response just makes it over the threshold of awareness, and is thus not accompanied by a strong 'Aha!' experience. A 100-point scale for self-reporting insight was used in MacGregor and Cunningham (2008). I asked participants to self-report their insight experience by moving on-screen sliders along four independent 100-point scales (though the movement appeared continuous), one for each of the four criteria. Each end of the scales was labelled with one of the opposite-sense descriptors of the four criteria.

None of problem types (insight, abduction, analytic) forms a uniform monolithic set. Hence, I sought to test a number of subtypes for each.

Insight problems are notoriously difficult to test in experimental conditions (Bowden et al., 2005): they can take hours to solve (if solved at all, due to impasse). Nonetheless, there are three subtypes of problem explicitly called insightful in the literature that are generally solvable in comparable timeframes to the other kinds of problems. CRA problems have already been discussed (Jung-Beeman et al., 2004). MacGregor and Cunningham (2008) claim that rebus problems are similarly insightful. These are problems where typographic and semantic features are manipulated to give common phrases. For instance, `SOMething` reads as 'the start of something big' and $\frac{\texttt{exit}}{\texttt{leg}}$ reads as 'go out on a limb'. Finally, there are what might be called classic insight problems, where fixation on the dominant (but irrelevant) interpretation of an ambiguity leads to impasse. For instance, 'A man in a small town married 20 different women of the same town. All are still living and he never divorced. Polygamy is unlawful but he has broken no law. How can this be?' (Gilhooly and Murphy, 2005). The answer is that he's a clergyman: he married them in the sense of performing the marriage ceremony, not in the sense of becoming their husband.

Abductive problems are most commonly researched in expert situations like medical diagnosis or scientific discovery (Johnson and Krems, 2001), but these expert situations are beyond the ability of non-expert experimental participants. Abduction is the generation of an hypothesis to explain a surprising event, and this can be an event in daily life, so I cre-

ated scenarios for five kinds of surprising event, four from daily life and one comparatively unfamiliar, involving an alien world.

Two subtypes required participants to hypothesise a *cause* for an event; two required hypotheses about the *motivation* for someone's behaviour. There was one descriptively simple and one descriptively complex causation subtype, as well as one simple and one complex motivation subtype. An example of a simple motivation problem is this: 'You see a woman shouting at a sales assistant in a shop. Why is she shouting?' A more complex motivation problem: 'Someone comes into the room and lights a candle. Everyone who lives in that apartment is single. In the corner, someone else is watching a DVD on their laptop, looking bored. Why did someone light a candle?' A simple causation problem: 'You hear a loud noise. What caused it?' A complex causation problem: 'It's been quite a sunny July so far. You see a tree with dead leaves all around it on the ground. The tree seems free from mould and rot. There are no other trees around. Why are there leaves on the ground?' Each scenario was designed to be open-ended, with multiple possible answers.

The complex questions included additional information, relative to simple questions, prompting inferences about relevance. The additional information was derived from responses in a pilot study (not reported here) in which volunteers came up with hypotheses for equivalent simple scenarios. For instance, the simple equivalent of the above complex motivation problem was 'There is a lit candle in a room. Why did someone light a candle?' Many participants responded with the answers 'it's Valentines day,' or 'there's been a power failure.' Some of the additional information in the complex scenarios was included to inhibit some of those earlier suggestions (if everyone is single, a romantic candle-lit dinner for Valentine's day is a less likely explanation). Some additional information was included to be suggestive, rather than directly inhibitory (if there's been a power failure, people are less likely in general to be watching DVDs, except that laptops can operate without mains power). Finally, some additional information was purposefully irrelevant (someone looking bored).

The fifth abductive subtype involved descriptions of an alien world populated by fictitious Zorgs. The scenarios were based on easily recognisable scenes (for instance, people singing 'Happy Birthday' while someone blows out candles on a cake), except that the cultural signifiers that would help

us recognise this event were either removed, or described in novel terms, or described in related but atypical terms. These mutated cultural signifiers were interspersed with irrelevant information. Hence: 'A Zorg in a purple hat is surrounded by a ring of other Zorgs. They're making a pleasing sounding noise in harmony, and its face flushes a gentle orange. They give it something covered in small blinking lights, which flash on and off in what seems a random pattern. It takes out a blade and looks expectant. What's going on?'

Analytic problems included both deductive and inductive questions. These were based on templates found in reviews of kinds of reasoning (Kurtz et al., 1999; Kemp and Jern, 2014). One inductive subtype involved choosing the more likely explanation for a scenario from two alternatives: 'Which of the following is the most likely explanation for a cough, given that the patient is a smoker? a) Emphysema b) A Cold'. Another involved rating how likely a particular explanation was on a percentage scale: 'Of all possible reasons for a train being cancelled, how likely is it that snow is the right one?' The third subtype involved novel property induction (again, with probability rated on a percentage scale): 'Cow guts contain the enzyme protylase. How likely is it that all herbivores' guts contain the enzyme protylase?' The final subtype involved evaluating deductive reasoning: 'All men are mortals and Cratylus is a mortal. Is Cratylus a man? a) Definitely b) Possibly c) Definitely not.' I included only one deductive subtype (compared to 3 inductive ones) since deduction isn't ampliative, and there is thus less need to show that it is distinct from abduction.

I predict that insight problems will be rated significantly differently from analytic problems along most of the four criteria, in line with results in Jung-Beeman et al. (2004). I also predict that ratings for abductive problems will be different from analytic problems in the same direction as insight problems. This would indicate that abductive problems are insightful, compared to deduction and induction. If correct, subjective reporting can be used as diagnostic of abduction in the world-learning problems that make up the rest of this dissertation, where features of the communicative context are manipulated to simulate the symbolic threshold. These will show that abduction was crucial at the symbolic threshold.

4.2 Methodology

4.2.1 Design

A within-subjects design was used: participants attempted problems of all three types (insight, abductive, analytic), then provided insight ratings along the four criteria discussed above. The independent variable, then, is `type`, and the dependent variables, as labelled above, are `sudden`, `explain`, `aha` and `confidence`.

4.2.2 Participants

I recruited 37 participants, 19 via the University of Edinburgh's (UE) job website; 18 via Amazon's Mechanical Turk (MT), a crowd-sourcing platform[2]. Given the relatively anonymous nature of MT work and the fact that workers are motivated to maximise their earning rate by preferring tasks to be as quick as possible, biographical data was collected only for the UE participants (5 male, 14 female; average age = 24.6, SD = 6.64). All participants were paid £4.50 for their participation. UE participants signed a consent form which explained the nature of the experiment; MT participants were informed of the nature of the experiment via the MT online interface and were told that clicking the link to the experiment applet indicated consent.

4.2.3 Materials

The experiment was coded in Processing (www.processing.org), a java-based open-source platform that exports an applet that can be accessed online. UE participants used University of Edinburgh iMac computers. MT participants participated online. For each problem type, there were a number of subtypes, described above (3 insight, 5 abductive, 4 analytic). Each subtype contained 4 problems, yielding 48 problems in all. The insight problems were taken unchanged from sources (Jung-Beeman et al., 2004; MacGregor and

[2]The main reason for having these two groups of participants was that part of this methodology is comparatively new: studies cited above showed a correlation between RH activation and subjective reporting, but here I will be using subjective reporting to distinguish types of problem. I thus wanted a comparatively diverse group of participants, and crowd-sourcing seemed a quick and efficient way of providing a more diverse group than recruiting within the university alone. Mason and Suri (2011) provide evidence for the validity of MT participants' responses in behavioural experiments.

Cunningham, 2008; Gilhooly and Murphy, 2005). Abductive problems were generated as described above. Analytic problems were based on templates found in reviews of such problems (Kurtz et al., 1999; Kemp and Jern, 2014). All problems are found in Appendix 1.

4.2.4 Procedure

Participants saw an on-screen welcome that gave detailed instructions, as well as an explanation of insight problem solving like that from Jung-Beeman et al. (2004) above. UE participants were able to ask questions to check understanding; MT participants were not, given that they participated online. Participants pressed a key to begin the experiment when they were ready.

Participants then worked through all problem subtypes in a randomised order: consecutive subtypes were not necessarily of the same type. Each subtype was introduced by an explanation and an example, then the four problems of that subtype followed in randomised order. Participants were given 2 minutes to solve each problem. After 2 minutes, they were asked if they wanted more time, or wanted to proceed to the next question. Selecting more time reset the 2-minute counter. The option to skip was offered because insight problem solving can lead to an impasse. After the four problems of each subtype, the next subtype was introduced.

Depending on the problem subtype, participants either typed an answer (for instance, when hypothesising or thinking of a word), selected from a set of options (for instance, when choosing which was the most likely explanation) or moved a slider along an on-screen scale (for instance, when rating how likely an event is). Because the abductive tasks were open-ended, there was no unique correct solution, so any answer counted as a solution. Even though the other types had a unique correct solution, for the sake of uniformity across all three types, any answer was accepted as a solution. The analysis below examines only insight ratings, not the solutions themselves.

Each time a problem was solved, a screen with four scales corresponding to the four insight criteria was displayed. Each time, the criteria were presented in a different randomised order. The directionality of each scale was also randomised independently of the other scales: providing a high insight rating could involve moving the slider either to the right or the left, depending on the arrangement of the above descriptors on that particular occasion. The sliders began in the center of the scale (halfway between an insight

and a non-insight rating), and participants were not able to progress to the next problem until all sliders had been moved. They were, however, able to return the slider to its original central position if they so wished. This was to prevent participants from progressing to the next question without providing an insight rating.

4.3 Results

Results were analysed using R, an open-source statistics package (R Development Core Team, 2011), and in particular the `lme4` package (Bates et al., 2011) for linear mixed-effects models (LMEMs). This is the most appropriate analysis for this data set, given that it exhibited heteroskedasticity and significant degrees of non-normality, and the design was unbalanced given that there were fewer subtypes of insight problem than the others. None of the above are problematic for LMEMs (Baayen et al., 2008; Barr et al., 2013).

An LMEM analysis begins with the maximal model plausible given experiment design (Barr et al., 2013), though if inspection of the random-effects correlation matrix shows values close to 1 or −1, this indicates over-parameterisation (Baayen et al., 2008), in which case those effects should be removed from the model. For this reason, I removed random effect `group` (MT or EU) and the random slope for `item`. The maximal model for this data included `type` (insight, abductive, analytic) as the fixed effect, with random intercepts and slopes for `participant` and random intercepts for `item` nested in `subtype`.

The estimated parameters for all four models are summarised in table 4.1 and displayed graphically in fig. 4.1. For instance, the model estimates that, for criterion `sudden`, ratings for abduction problems are 16.241 points higher than for analytic problems, averaged across the population, while insight problems are 8.620 points higher than analytic problems. Note that `aha` shows the largest effect size, and that `confidence` ratings for insight and abduction are actually lower than for analytic problems.

There are several ways to estimate p-values in an LMEM, though Barr et al. (2013) note that there is not yet consensus within the field about which is best. I thus use all three of the below options, where applicable.

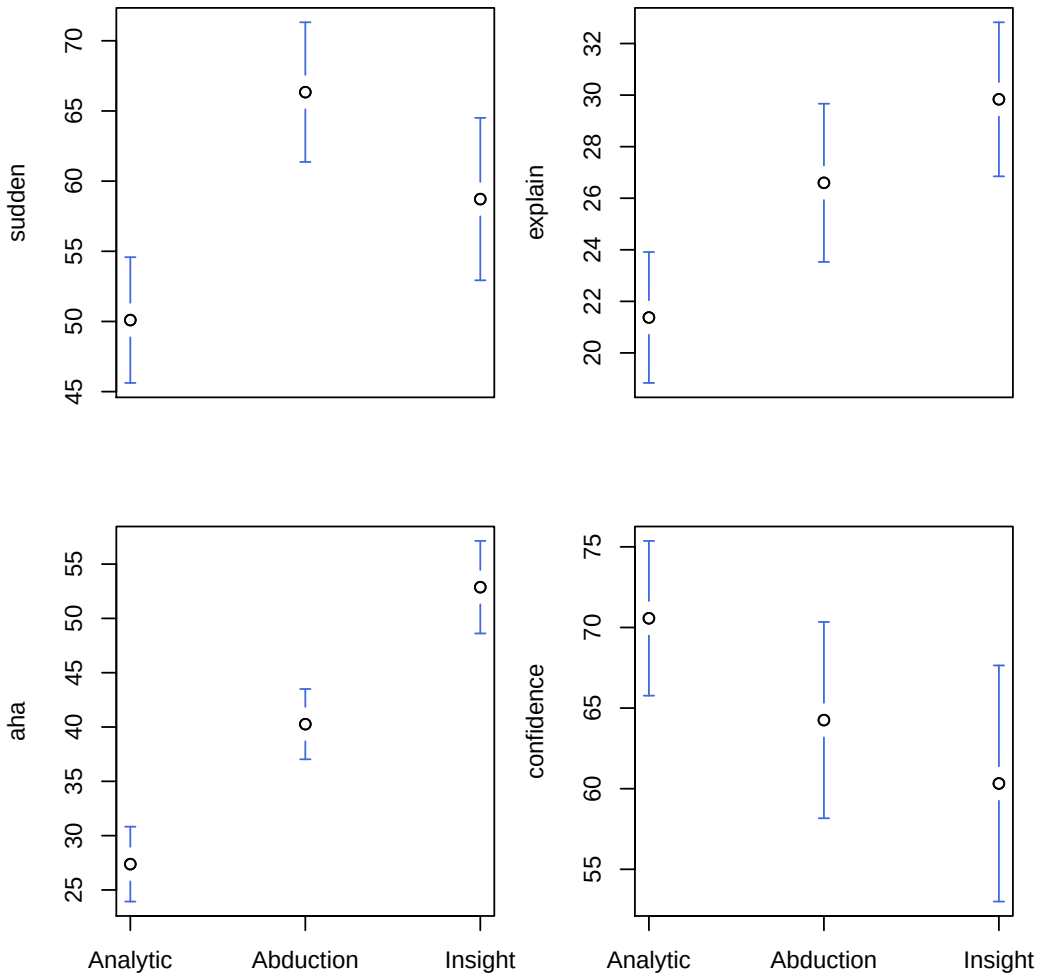


Figure 4.1: Model estimates of mean for each criterion by problem type. Bars indicate model estimates of standard error of the mean.

Criterion	Parameter estimate	SE	t
`sudden`			
Abduction	16.241	4.997	3.263
Insight	8.620	5.787	1.490
`explain`			
Abduction	5.224	3.068	1.702
Insight	8.462	2.987	2.833
`aha`			
Abduction	12.886	3.235	3.984
Insight	25.496	4.258	5.988
`confidence`			
Abduction	−6.314	6.858	−1.037
Insight	−10.246	7.417	−1.399

Table 4.1: Experiment 1 maximal model parameters for fixed effects (compared to base level `analytic`), standard errors (SE) and t-values.

Barr et al. (2013) recommend comparing the log-likelihoods[3] of two models (the maximal model and a null model, precisely the same in every respect, except with the fixed effect removed) using the `anova` function. However, while this method shows whether `type` has a significant effect on insight rating, it doesn't show whether there is a significant difference between the base level (analytic) and each of the other levels (abduction and insight). Baayen et al. (2008) suggest that an informal way of deciding whether these differences are significant is to look at the t-value (table 4.1): if its absolute value is larger than 2, the effect is probably significant. Baayen (2008, 269) derives p-values from t-values according to formula 4.1.

$$p = 2 * (1 - \texttt{pt}(\texttt{abs}(t), Y - Z)) \tag{4.1}$$

Here, `pt` is a function in R that accesses the probability distribution for t. The function takes two arguments: the absolute value of t and an upper bound for the degrees of freedom given by subtracting the number of fixed effects parameters (Z) from the number of observations (Y).

Baayen (2008) recommend using a Markov chain Monte Carlo (`mcmc`) simulation. However, this method is not implemented for models with random slopes (Barr et al., 2013), so this method is not suitable for the above maximal models. It is possible to use simplified models (i.e. excluding the

[3] One measure of how well a model fits the data.

random slope for `participant`) in order to allow `mcmc` analysis.

The derived p-values are shown in table 4.2. Though the `mcmc` p-values are generally less conservative than the values based on the maximal model, they all agree that, for `sudden`, `explain` and `aha`, `type` is a significant effect. For `confidence`, the `anova` p-value shows that `type` is not a significant effect; the `mcmc` p-values suggest that insight is significantly different, but that abduction is not (though it lies in the predicted direction). So `confidence` is not a reliable diagnostic for insight according to more conservative p_2.

P-values based on the maximal model are higher than those based on the reduced model, because the latter are anticonservative (Barr et al., 2013). There are two cases where the `mcmc`-method finds significance where the t-value method does not (excluding `confidence`): insight for `sudden` and abduction for `explain`. In each case, though, the non-significant level lies in the same direction as the significant level, relative to the analytic base level. The results for `aha` show a consistently significant effect of `type` at both levels (abduction and insight, compared to analytic problems).

	p_1		p_2		p_3	
	χ^2	p	abd. p	ins. p	abd. p	ins. p
`sudden`	9.1731	**0.0102**	**0.00112**	0.136	**0.0001**	**0.0170**
`explain`	7.38	**0.025**	0.0889	**0.00467**	**0.0136**	**0.0004**
`aha`	23.801	**<0.0001**	**<0.0001**	**<0.0001**	**0.0001**	**0.0001**
`confidence`	2.142	0.343	0.299	0.162	0.0802	**0.0098**

Table 4.2: Experiment 1 p-values derived by three methods discussed above: `anova` (p_1), t-value by formula 4.1 (p_2); `mcmc` (p_3). Significant values are in bold. Both the `anova` and t-value methods use the maximal model; the `mcmc` uses the reduced model. Both the t-value and `mcmc` methods indicate whether the difference between abductive and analytic problems (abd. p) and between insight and analytic problems (ins. p) is significant.

4.4 Discussion

This experiment sought to show that abduction is more insightful than induction, an analytic problem. After attempting different types of problem (abduction, insight and analytic), participants provided ratings used in insight research to characterise insight problems. Ratings were provided independently along four specific criteria, rather than for a general sense of 'insight'.

The LMEM analysis shows that the `confidence` criterion fails to distinguish insight from analytic problems (table 4.2), so it cannot be used to distinguish abduction from induction. This is not surprising: criterion 4 was essentially a confidence rating. For many insight problems, the moment of realising the answer is often accompanied by a realisation that the answer is obviously correct (Bowden et al., 2005). However, high levels of confidence are typical of a range of non-insight phenomena, including recall of familiar facts and many analytic problems. Indeed, this is the only criterion where ratings for insight problems (and abduction) were lower than for analytic problems.

For the `sudden`, `explain` and `aha` criteria, a significant effect of `type` was found in the `anova` analysis and in each case, abduction lay in the same direction relative to analytic problems as insight did. The question remaining is whether each difference is significant, because the `anova` analysis doesn't distinguish within the levels (abduction and insight) that are contrasted with the base level (analytic). To check whether each criteria distinguishes insight from analytic problems and whether abduction is significantly insightful, we must rely either on the t-value method based on formula 4.1 or the `mcmc` method, which is based on non-maximal models since the `mcmc` package in R cannot handle models with random slopes.

These alternatives offer contradictory advice for `sudden` and `explain`: the non-maximal models find both levels significant for both criteria; the maximal models find insight non-significant for `sudden` and abduction non-significant for `explain`. They agree, however, that `aha` strongly and reliably distinguishes insight from analytic problems and show that abduction is significantly insightful.

If we prefer the t-value method, then `sudden` fails to distinguish insight from non-insight problems, in which case criterion 1 is not diagnostic of insight as a distinct cognitive process. On the other hand, `explain` does distinguish insight from non-insight problems, in which case criterion 2 can show whether abduction is insightful. But `explain` ratings for abduction are not significantly different from ratings for analytic problems by the t-value method, though the effect does lie in the same direction as insight. If, however, we prefer the p-values from the non-maximal model, then abduction is significantly insightful according to both criteria.

So as things stand, the features of insight that are significantly different

from analytic problems are either given by the set {`sudden, explain, aha`} or the set {`explain, aha`}. The former is justified by the p-values of a non-maximal model, in which case abduction is significantly more insightful than induction across all three criteria. The latter is justified by the t-values of the maximal model, in which case abduction is significantly more insightful than induction according to criterion `aha`.

Both sets, however, contain `explain` and `aha`, so it is natural to look in a little more detail at these. The results based on `aha` are clear, so the question of whether the difference between abduction and induction is statistically significant overall depends on how much weight we give `explain` compared to `aha`.

There are theoretical and statistical reasons for lending more weight to `aha` than `explain` here. Firstly, Bowden and Jung-Beeman (2003a) claim that the 'Aha!' experience is central to insight, as do Kaplan and Simon (1990). Many cognitive processes, though, are ineffable, so `explain` is more a matter of distinguishing inference from reasoning according to my definitions of these terms. Secondly, both the parameter estimates and t-values in table 4.1 are larger for `aha` than for `explain`, so the former is a stronger diagnostic for insight.

It may simply be the case that abduction ratings by `explain` vary less strongly compared to induction than they do by `aha`. Fig. 4.2 shows `explain` ratings broken down by abduction subtype. Some problems score higher than others, with the most novel (the Zorg questions) scoring the highest. Further research could explore what effect novelty has on insight ratings by manipulating how many features of the vignette correspond to something recognisable, and how similar the descriptions of features are to conventional cultural signifiers. That lies outside the scope of this thesis, though, since it relates to manipulations of features in analogic bases (cf. §3.5.1).

The result of the experiment coheres well with claims in previous chapters. I argued that pragmatic inference about the meaning of novel signs involves context-, relevance- or salience-deciding processes (§1.5.2); that induction is incapable of dealing with contextually unconstrained novelty (§3.4); and that abduction generates novel hypotheses based on plausibility (i.e. on representational structures, §3.3.6). Insight was shown to involve activation over representational structures in the same brain areas that are involved in contextually unconstrained or novel pragmatic inference (§2.6;

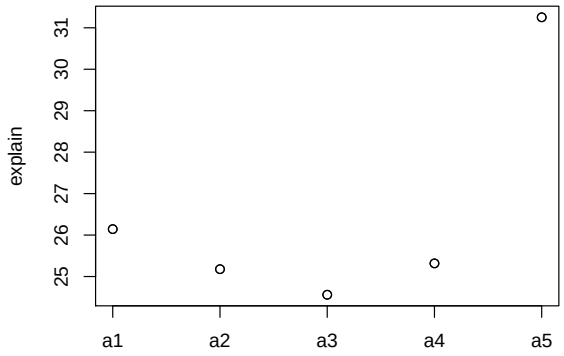


Figure 4.2: Mean `explain` ratings broken down by abduction subtype: a1 = complex causation; a2 = simple causation; a3 = simple motivation; a4 = complex motivation; a5 = Zorg world.

§3.5.2). Recall the Graded Salience Hypothesis (§2.6.4), for instance: the less conventional (or more novel) the metaphor, the harder it is to generate an relevance-deciding inference about the ground of the metaphor, and the more activity in the RH temporal lobe. The results here should be anything but surprising, therefore.

To summarise the main results, the experiment shows that abduction is, on the whole, more insightful than induction. Diagnostics for insight can thus be used to evaluate the extent to which abduction and induction each shoulder the explanatory burden in word learning in different contexts, and I will use the `aha` criterion as the diagnostic in remaining experiments. In cases where only a small set of hypotheses are possible, where hypotheses are highly predicable from the context, or signs are familiar, abduction will not need to perform a very difficult job (which is not to say it will perform no job at all) and the bulk of the work will be done by induction. In such cases, there should be a comparatively low `aha` rating. But in cases where there are many theoretically possible answers, or hypotheses are not highly predictable from context, or signs are novel, abduction will play an important role in generating a manageable hypothesis space for induction to evaluate. In such cases there will be a comparatively high `aha` rating.

Finally, I turn to consider a couple of objections.

If abduction is insightful, why are ratings for abduction not more similar to those for insight? There are a number of plausible explanations. Firstly, all insight problems involved comparatively short strings (CRAs involved just three words, and some rebus problems even fewer) while abductive problems involved several sentences, and sometimes long paragraphs. The effort involved in processing this extra linguistic stimuli would have incorporated non-insight processes, and may have thus lowered insight ratings. In fact, abductive problems involved more than just linguistic knowledge: they involved integrating linguistic knowledge with world knowledge (which involves frontal lobe processing, Menenti et al., 2009), while insight problems involved mostly linguistic knowledge. Finally (and probably most importantly), insight problems involved one correct answer, while abductive problems were more open-ended. The experience of finding the correct solution may have increased the strength of the flash of realisation that accompanies insight problem solving. What matters, however, is that these results consistently show a difference between abduction and induction.

Since Deacon (1997) claimed that insight is crucial at the symbolic threshold, why add abduction to the mix? Why not just investigate the role of insight in symbol origins? The answer has more to do with the state of the field than it does with abduction itself. Induction has been the focus of much word-learning research in language evolution. But insight and induction don't seem to have anything to do with each other. Introducing abduction into the discussion is a crucial step in that it relates these disparate processes: it is the input to induction, and it operates insightfully. Further, the central question here is essentially semiotic, and abduction is fundamental to Peircean semiotics.

Chapter 5

Word Learning and Hypothesis Spaces

5.1 Background and Aims

I have identified context-size as a major dimension along which the evolution of inference should be explored and have argued that crossing the symbolic threshold involved contextually unconstrained inference. My central claim, though, is that abduction was necessary for crossing the symbolic threshold, so I need to show that abduction is required for contextually unconstrained inferences about the meaning of novel signs. Induction, on the other hand, is a contextually constrained inference.

Complex tasks typically require both: hypotheses are generated and then evaluated. But the burden shouldered by each differs according to task type: at times it may be comparatively easy to generate a set of hypotheses while choosing the best from that focal set is difficult; at other times, coming up with any hypothesis at all is very difficult, but once the hypothesis is generated, its posterior probability is immediately recognisable as being extremely high. Different word-learning tasks, then, may depend on different levels of involvement by both forms of inference.

In this experiment, I manipulate context size (hereafter, 'world size' or WS): in small-world problems, there is a highly limited set of possible answers; in large-world problems, a vast set. Medina et al. (2011) contrasted the small-world problems typical of inductive experiments with more realistic large-world problems (see §2.5.2.2 for a discussion and illustration) and

showed that WS has a significant effect on learning. In small-world problems, it is possible to inductively evaluate all possible hypotheses, whereas in large-world problems, a focal set of hypotheses must be generated by abduction before induction can proceed. Abduction, then, would play a small role in small-world problems and a large role in large-world problems.

The previous chapter showed that subjective reporting of an insight experience (or Aha! moment) is diagnostic of increased levels of abduction, compared to induction. I thus use subjective reporting as an indicator of increased levels of abduction in the present experiment. Firstly, I predict that insight ratings in unconstrained (or vast) worlds should be significantly different from insight ratings when a focal set of hypotheses is given for evaluation. Secondly, I predict that as WS increases, insight ratings should increase, but because increases in WS can be gradual, increases in insight ratings should be gradual, too. That is, I expect an upwards trend in insight, not a series of significantly different values.

I investigate this question in a Pictionary-like game, where participants must guess what cue the drawer had in mind while producing a simple drawing. An alternative would have been to pair a novel word with video footage (as in Medina et al., 2011), but manipulating world size in that case would have been impractical. Further, there is precedent for the use of Pictionary-like games in investigating symbol origins (Garrod et al., 2007; Fay et al., 2010).

The experiments by Garrod, Fay and colleagues allowed feedback to help the guessers get the right answer. However, participants would then be processing feedback while processing novel signals, a possible confound for insight ratings. Further, I wanted the stimuli for each guesser to be identical, whereas feedback in the above experiments varied according to the particulars of a given drawer+guesser interaction. All stimuli for this experiment were thus prepared prior to the experiment, rather than created by participants during the experiment.

A way of helping guessers discover the answer without feedback is suggested by Compound Remote Associate (CRA) problems, where participants see three words and must find a fourth word that collocates with each. Each word individually has multiple collocates, but the set narrows in on a unique answer. Similarly, any drawing in a Pictionary-like game could prompt many guesses, but a series of three pictures illustrating the same cue would help

participants narrow in on one answer. Each word-guessing task here thus involved simultaneous presentation of three independently produced drawings to illustrate the same cue.

To distinguish contextually constrained and unconstrained problems, I contrasted tasks where participants had to evaluate a set of hypotheses from tasks where they had to generate their own. In the unconstrained tasks, participants saw only the trio of drawings and had to guess the cue without any further help; in the constrained tasks, they simultaneously saw a set of possible answers and merely had to decide which of those was the best answer for the stimuli. Constrained and unconstrained tasks correspond to predominantly inductive and predominantly abductive problems, respectively. Further, WS in constrained tasks could be manipulated: in a very small world, there were only 2 possible answers to be evaluated; in a medium-size world, there were 4; in a large-world, 8, while the unconstrained task allowed a vast world. I took 8 to be the upper limit of constrained WS since Smith et al. (2011) found this to be the upper limit on participants' ability to accurately track probabilities in a cross-situational word-learning task.

After guessing the answer, participants were asked to provide a subjective report of their insight experience in solving the problem. The previous experiment decomposed various features of an insight experience into four independent criteria and showed that experience of an Aha! moment was the best predictor of problem type. Hence participants were asked to rate their experience along this one dimension, though a fuller description of insight problem solving was given in the instructions.

5.2 Methodology

5.2.1 Design

Participants provided subjective reports of a feeling of insight after word-guessing tasks. Each participant tried to solve tasks without any answers given (open-context task), and tasks with possible answers given (closed-context task). There were three subtypes of closed-context task: either 2, 4 or 8 possible answers were given for evaluation. Participants saw four open-context problems and four problems for each subtype of closed-context problem (hence 16 problems in total), yielding a within-subject repeated-measures design with one independent variable (WS = 2, 4, 8, vast) and one

dependent variable (self-reported insight rating).

5.2.2 Participants

21 participants were recruited via mailing lists for undergraduate linguistics courses at the University of Edinburgh and through the university's student job website. They were paid £6 for their participation and signed a consent form which explained the nature of the experiment.

5.2.3 Materials

Sixteen cue words were selected from randomly chosen Pictionary cards. Three volunteers worked through the list of 16 cues, each time drawing a simple picture that would prompt a partner (one of three other volunteers) to guess the cue. Each drawer+guesser pair worked independently. This process produced sixteen sets of three pictures. For instance, fig. 5.3 shows three pictures drawn independently to represent 'ripe'. All stimuli are given in appendix 1.

To generate a list of possible answers for a trio of pictures, guesses for each individual picture in that trio were collected separately via Amazon's Mechanical Turk service (see previous chapter). The guesses for all three individual pictures within a given trio were then collated and ranked from most-commonly produced to least-commonly produced. The possible answers for a context-constrained problem of WS = n consisted of the real answer plus the top $n-1$ other answers on that list.

The experiment was conducted on iMac computers in a University of Edinburgh computer lab. Several participants were in the lab at a given time, but they worked separately, were widely spaced, and were unable to see each other's screens, so were effectively isolated.

5.2.4 Procedure

After a welcome screen and instructions (which included practice CRA problems to illustrate insight), participants attempted to guess the cue represented by trios of pictures. The order of conditions (WS) was randomised, and items were assigned randomly to conditions. For closed world conditions, either 2, 4 or 8 possible answers were displayed beneath the trio of pictures; for the open-world condition, no such answers were given.

Answers were typed. If participants guessed incorrectly, they were told that their answer was incorrect and were able to try again, repeatedly, until they guessed the answer. After 10 incorrect guesses, they were asked if they wanted to persevere with that problem or skip to the next problem. This is because insight tasks can result in impasse. Inflected forms of the answer were counted as correct (mile, miles); synonyms were not (purple, violet), as is normal in a game of Pictionary.

If they guessed correctly, they were asked to move a slider along a 100-point scale (though movement appeared continuous) to provide an insight rating. Each end of the scale was labelled with an appropriate descriptor: (`I had an 'Aha!' moment, like a lightbulb suddenly flashing on` vs. `I had no 'Aha!' moment`. The left-to-right order of these descriptors changed randomly for each problem. Participants were not able to progress to the next problem until the slider had been moved, though they were able to return it to its original position (central between the end points) as an answer.

5.3 Results

Across the whole participant pool, there were 14 cases where participants chose to skip the question after 10 incorrect guesses. Eleven of these were in the open world condition; 3 in closed worlds. One participant in particular seemed to have misunderstood the instructions: it seems they didn't realise (at least initially) that the words on screen were possible answers. Since insight ratings were collected only for correct guesses, the participants who skipped a question were not able to provide an insight rating for that question. The average number of attempts needed to reach the correct solution is given in table 5.1.

WS	2	4	8	vast
Attempts	1.0375	1.25	1.6375	2.3375

Table 5.1: Average number of attempts needed to reach the correct answer per world size (WS).

Two main analyses were carried out. The first compares open or vast worlds with closed or constrained worlds. The second analysis examined the effects of increasing context size in constrained worlds.

For the first comparison, a linear-mixed effects model (LMEM) was constructed and analysed using R, particularly package `lme4` (Bates et al., 2011). The dependent variable was insight rating (`insight`) and the fixed effect was `world` (open or closed). Random effects included `subject` (random intercept and slope) and `item` (random intercept only). A random slope for `item` indicated overparameterisation (a value of 1 in the random effects correlation matrix) and was thus excluded (Baayen et al., 2008).

P-values were derivable in 3 ways (described in the previous chapter): via the `anova` function comparing the maximal model with a null model (Barr et al., 2013); derived from the t-value according to the formula given in the previous chapter (Baayen, 2008); and using an `mcmc` analysis on a reduced model (removing the random slope).

Table 5.2 summarises the model parameters and derived p-values; the parameters are displayed graphically in fig. 5.1. Though p_1 is the most conservative, all show that insight ratings in open worlds are significantly higher than those in closed worlds.

	Insight	SE	t	p_1	p_2	p_3
Closed	44.359	4.471	9.922			
Open	53.773	4.534	2.076	**0.0438**	**0.0387**	**0.0057**

Table 5.2: Maximal model estimated parameters for mean insight rating per condition, standard errors (SE) and t-values, as well as p-values derived by the `anova` method (p_1), Baayen's t-value method (p_2); and the `mcmc` method (p_3). The `anova` and t-value method use the maximal model; the `mcmc` method uses the reduced model as described above.

The second main analysis investigated how insight ratings varied as context size increased. Table 5.3 shows the estimates of an LMEM model that replaces `type` (`open/closed`) with `WS` (`2/4/8/vast`), and these are displayed graphically in fig. 5.2. For instance, the model estimates that insight ratings for WS=4 were just 4.082 points higher than for WS=2. Apart from `WS=vast`, the t-values are less than 2, suggesting the difference between upper levels and base level (WS = 2) is not significant.

However, I predicted an upwards trend, rather than significant differences between each level, since insight ratings should increase gradually with context size. Page's L is a non-parametric repeated-measure trend test that compares ranked values to evaluate the extent to which observations (here, `insight`) have a particular order (here, increasing with `WS`). We know from the previous analysis that open worlds require significantly more in-

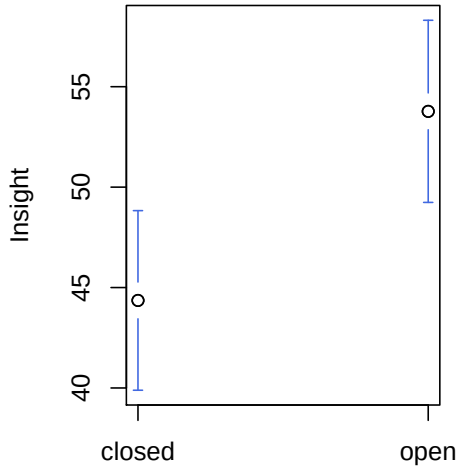


Figure 5.1: Model 1 estimates of mean for closed (`WS=2, 4, 8`) and open (`WS=vast`) worlds. Bars indicate model estimates of standard error of the mean.

sight than closed worlds, so to provide a more conservative test of whether `insight` increases with `WS`, data for vast worlds are excluded from analysis here. The test gives L = 265, $p < 0.001$, so insight ratings tend to increase with world size.

WS	mean insight rating	SE	t
2	41.633	5.569	7.476
4	45.715	5.777	0.707
8	46.016	5.148	0.851
vast	54.140	5.449	2.295

Table 5.3: Model estimates for the mean insight ratings for different levels in world size (WS), with standard errors of the mean (SE) and t-values.

5.4 Discussion

The experiment sought to show that inferring the meanings of new signs requires abduction by comparing open worlds (with no hypotheses given) with closed worlds (with a set of hypotheses give). Participants provided subjective reports of their experience of insight or 'Aha!' moment after guessing

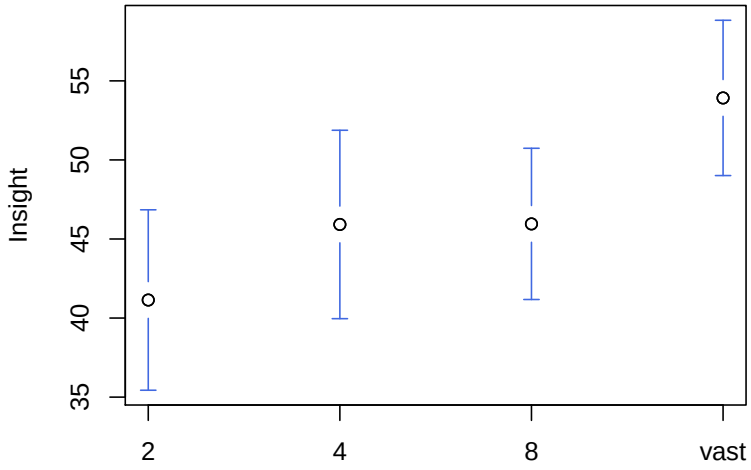


Figure 5.2: Model 2 estimates of `insight` by `WS`. Bars indicate model estimates of standard error of the mean.

the intended meanings of novel pictures in open and closed worlds. These insight ratings are diagnostic of increased levels of abduction, compared to induction.

The LMEM analysis shows that insight ratings for open worlds were significantly higher than those for closed worlds. Insight thus plays an abductive role by generating hypotheses in the former case; but has less of a role to play in the latter case since hypotheses are given. It still has some role to play, however, because the given hypotheses concern the pictures as a whole, but participants would still have generated hypotheses about what elements of the pictures mean. For instance, fig. 5.3 shows the trio of pictures representing the cue 'ripe'. Complex inferences are needed to recognise that the face in the central figure is meant to represent an *old* woman and that a person's being old is analogous to fruit being ripe; or that gravity is irrelevant for understanding the falling apple in the right-hand figure. These hypotheses were not given, and would thus have to be generated, requiring insight.

Induction alone cannot account for how this hypothesis is reached. Prior probabilities are not derivable for all English words in response to these

Figure 5.3: Pictures representing cue 'ripe'.

pictures given that the pictures are novel (§3.3.6, §3.4.2). Some probabilistic information is easily derivable from the pictures, but that information is misleading: an apple appears in all three pictures, but assigning APPLE a high prior here does not explain how the answer RIPE is arrived at. It requires insight to make the creative leap connecting representations of TIME, AGE, FRUIT, FALL or COLOUR to yield RIPE.

Further, I predicted that insight ratings would increase with context size, given that I argued that context-size is a major predictor of inference type (§2.5). Page's L showed that insight tended to increase with context. So abduction plays a larger role as context expands. Inductive accounts of word learning typically deal with contextually-constrained contexts (§3.4.1); Medina et al. (2011) argued that this is problematically unrealistic. The results here show that there there is a significant cognitive difference as contexts become less constrained.

While it is easy for Bayesian accounts to overlook the complexities of hypothesis generation in the sorts of constrained contexts typical of experiments in symbol evolution, the communicative context would have been comparatively unconstrained at the symbolic threshold. Gesture may have played a dual role at the symbolic threshold by directing attention or by providing iconic or indexical information. Inferences about the meaning of novel icons or indexes would have been just like inferences needed to understand the pictures above (that is, context-, salience- and relevance-deciding hypotheses, and thus abduction. I demonstrate this in ch. 7), whereas the attention-directing role may have helped constrain context.

Even in comparatively small contexts, though, one would still have to infer the speaker's or gesturer's communicative intentions, which would have required abduction. Even for very small contexts here (WS = 2), average

insight ratings were quite a bit higher than for purely inductive problems in the previous experiment (41.633 vs. 27.378, respectively) for reasons I set out above regarding the cue 'ripe'. In vast worlds here, insight ratings ranged as high as pure insight problems in the previous experiment (54.14 vs. 52.874, respectively).

The range of insight ratings here supports my claim in the previous chapter that there isn't necessarily a sharp contrast between abductive and inductive problems: both types are needed, and the balance between them shifts in a graded fashion. The next experiment provides more detailed support for these claims in that it involves a more graded manipulation of context size, investigating how predictable an intended word is given contextual information.

A potential extension of this experiment would involve measuring latencies while increasing WS more gradually than I have done here. If hypothesis generation were simply a matter of searching an hypothesis space (as some inductive accounts would have it), then latencies might increase linearly with context size. If, on the other hand, insight plays an increasingly large role as context size increases, then latencies would begin to plateau as abduction comes to play the dominant role.

A second extension (which could be combined with the previous one) would be to conduct an individual differences study. Participants would undertake a number of classic insight problems and receive a score based on their success, indicating how insightful they are. In constrained contexts, there wouldn't be a huge difference between insightful and less-insightful participants, since most of the work involves hypothesis evaluation. As context size increases, though, and hypothesis generation becomes important, insightful participants should gradually become advantaged over less insightful participants, so their latencies would be lower and success rates higher in large contexts.

Chapter 6

Word Learning and Predictability from Context

6.1 Background and Aims

The previous experiment investigated context size by manipulating the number of possible answers to a word-guessing problem. But even within a given context size, levels of predictability may vary: a novel sign with the relevant features made salient would be easier to understand than one with the relevant features obscured. So in addition to context size, an important feature of inferring meaning is the degree of predictability from context or environment. In this experiment, I manipulate levels of predictability in a word-guessing task. The less predictable the answer, the larger the role of abduction, so the higher the predicted insight ratings.

This requires a task that offers accurate measures of predictability. A Pictionary-like problem might provide indirect measures based on rates of success across a large sample group, but predictability would be difficult to manipulate in that case. However, a compound remote associate (CRA) task can provide direct, accurate measures of predictability: the probability of a particular word following a particular cue can be derived from a large corpus such as the Corpus of Contemporary American English (COCA, Davies, 2008). Limiting the discussion to noun+noun collocates, if one is trying to think of a noun that can follow the word 'gym', COCA shows that the most likely answer is 'bag': it follows 16.4% of the time. On the other hand, 'bag' follows 'book' only 4.01% of the time. So 'bag' is more predictable from

'gym' than from 'book'.

In §2.6.4, I reviewed evidence showing that classical language areas in the left hemisphere (LH) are active when a word is highly predictable from context, while right hemisphere (RH) insight areas are more active when a word is less predictable from context. Similarly, a conventional metaphor such as 'bright student' involves LH dominance, while a novel metaphor such as 'conscience storm' involves more RH activation (Mashal et al., 2007): 'student' is highly predictable given 'bright' while 'storm' is not very predictable given 'conscience'. In addition, Jung-Beeman et al. (2004) showed that subjective reporting of an 'Aha!' moment of insight is diagnostic of increased activation in the relevant areas. As with previous experiments, then, I use subjective insight ratings as the dependent measure. Of the following two CRA problems, then, the first should require lower levels of insight.

GYM	BOOK
PAPER	PAPER
SHOULDER	SHOULDER

Using COCA, I constructed six lists of noun+noun collocates. I limited all collocates to one syntactic class, in case syntactic class provided a confound. While Bowden and Jung-Beeman (2003b) use CRA problems where the collocate could precede or follow individual cues, and where the collocate could be separated by a space or not, I limited my stimuli to collocates that follow the cues, separated by a space, again to avoid confounds. Each list consisted of five cues that collocate with the same answer: the following list, for instance, cues 'bag'. Lists were ranked in order of how predictable the answer was given each cue: 'bag' is more likely to follow 'shoulder' than it is to follow 'grocery'. To provide three levels of predictability in CRA problems, triads of cues were grouped into three conditions (high, medium, and low predictability) as marked. I predict that insight ratings will increase as predictability decreases.

	GYM	
	PAPER	Condition 1: high predictability
Condition 2: medium predictability	SHOULDER	
	GROCERY	Condition 3: low predictability
	BOOK	

6.2 Methodology

6.2.1 Design

Participants provided subjective reports of a feeling of insight after guessing the collocate that could follow all three cues in a CRA problem. There were 6 problems in total. For each participant, each of the six problems was assigned randomly to one of three conditions (high, medium or low predictability), resulting in 2 problems per condition, yielding a within-subject repeated-measures design with one independent variable (predictability = high, medium, low) and one dependent variable (self-reported insight rating).

6.2.2 Participants

40 participants were recruited via Amazon's Mechanical Turk crowd-sourcing platform (see ch. 4). Given that predictability values were derived from an American corpus, the experiment was only made available to American participants. Given the relatively anonymous nature of MT work and the fact that workers are motivated to maximise their earning rate by preferring tasks to be as quick as possible, biographical data was not collected. They were paid £2 for their participation. After the nature of the experiment was explained, participants were then told that clicking the link to the applet constituted consent.

6.2.3 Materials

Table 6.1 shows the 5 possible cues for each answer. These were grouped into triads as described above, and one triad for each list was assigned to one of the three conditions. Rank values for each collocation were derived from COCA. For instance, 'bag(s)' is ranked 1st among all noun collocates of 'gym'. In all cases, inflected forms such as plurals were grouped together. Cloze probabilities for each collocation were also derived from COCA. COCA gives values for how many times in the corpus the answer follows each cue, as well as a total for the top 100 noun collocates of that cue. Dividing the first by the second gives an estimate value for how predictable the answer is given the cue. For instance, 'bag(s)' follows 'gym' 260 times in the corpus, and the top noun 100 collocates of 'gym' occur with it

1585 times in the data, so 'bag' follows 'gym' with an estimated probability of 0.16403. Again, inflected forms were included.

The cloze probability is merely an estimate since only the top 100 collocates are given in the COCA search results, but the estimate is serviceable since other collocates are rare and would thus have a small effect on the total. The 100th collocate of 'gym', for instance, follows it on only 2 occasions in the corpus. The 101st collocate would thus increase the total by a maximum of 2. Assuming this maximum value, the probability of 'bag' following 'gym' would then be 0.16383, a difference of just 0.0002. The smallest probability in the data set (0.0213) is still a couple of orders of magnitude larger than this difference.

Answer	sign	bag	lamp	list	button	paper
Cues	neon	gym	kerosene	wish	snooze	wrapping
	exit	paper	bedside	waiting	panic	rice
	dollar	shoulder	desk	hit	mouse	morning
	road	grocery	floor	guest	elevator	construction
	plus	book	lava	shopping	call	Sunday

Table 6.1: Ranked collocation lists for each answer word. For instance, `sign` is ranked higher among collocations of `neon` than it is among collocations of `exit`.

It turns out that the variance in cloze probability is lower in condition 3 than it is in condition 1. This is an unavoidable artefact of the distribution of collocates: fig. 6.1 shows the relative proportion of the top 20 collocates for cues `floor`, `guest` and `grocery` (adjusted so that the top collocate for each has the same height, for ease of comparison). The proportion drops very steeply at first, then levels off. High ranking collocates each account for a larger percentage of all collocates, and differences between them are comparatively large; low ranking collocates each account for a smaller percentage of all collocates, and differences between them are thus comparatively small. This simply means that values for the independent variable will tend to cluster at the lower end of the x-axis.

6.2.4 Procedure

Participants were introduced to CRA problems and given the chance to practice on 4 examples from Bowden and Jung-Beeman (2003a). Insight was explained in terms comparing the 'Aha!' experience to a light bulb suddenly flashing on. After pressing a key to begin the experiment, partici-

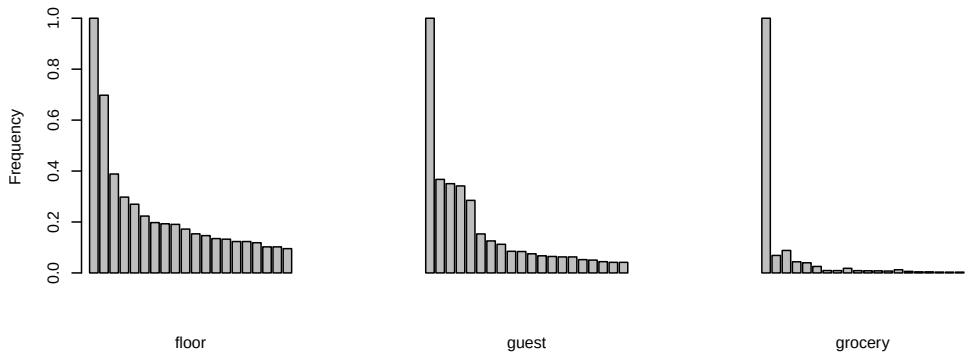


Figure 6.1: Relative distribution of the top 20 collocates for three cues.

pants were shown the 6 CRA problems in a randomised order. If they hadn't solved a problem within 90 seconds, they were asked if they wanted to skip to the next problem or persevere. This is because insight tasks often lead to impasse, and long periods of impasse would discourage participants from completing the experiment. After typing the correct answer (inflected forms were accepted as correct), participants provided an insight rating by moving an on-screen slider along a 100-point scale (though motion appeared continuous). The scale was labelled with insight descriptors ('I had an "Aha!" moment, like a lightbulb suddenly flashing on' vs. 'I had no "Aha!" moment'). The left-to-right order of these descriptors was randomised for each problem. Participants were unable to proceed until they had moved the slider from its initial position midway along the scale, though they could then return it to that position if they wished.

6.3 Results

Participants averaged 78.05% correctness across all problems (sd = 19.16269). Broken down by `condition`, average values for correctness were: `high` = 82.23%; `medium` = 75.95%; `low` = 79.49%.

Results were analysed in R using package `lme4` (Bates et al., 2011) to create linear mixed-effects models (LMEMs). The maximal model treats `condition` as the fixed effect, with random intercepts for subject and item.

Random slopes were not included since doing so produced a value of 1 in the random-effects correlation matrix, indicating overparameterisation (Baayen et al., 2008). Comparison with a null model using the `anova` function (see ch. 4) showed no significant effect of `condition` ($\chi^2 = 3.3016, p = 0.1919$).

However, `condition` was a very coarse measure of predictability, given that items within each of the three conditions could vary in predictability values. For a more detailed analysis, an LMEM was constructed with `rank` values for each cue derived from COCA. For instance, `panic` and `exit` are both the first cue in the CRA problem in the `medium` condition, but COCA shows that the answer 'button' is ranked 3rd among all the noun collocates of cue 'panic' and answer 'sign' is ranked 7th among all the noun collocates of cue 'exit'. The second LMEM thus replaced fixed effect `condition` with three fixed effects `rank1`, `rank2`, `rank3` (`rank1` referring to the 1st cue in each CRA problem, and so on).

The model found a t-value greater than 2 for `rank1` ($t = 2.535$), but not for `rank2` ($t = -1.040$) or `rank3` ($t = -1.174$). Only the first cue in the CRA is thus likely to have a significant effect on `insight`, given the rule of thumb in Baayen (2008) that t-values lower than 2 are likely to be insignificant. `Anova` comparison with null model showed a significant effect of the rank value of the first cue of the triad ($\chi^2 = 6.36, p = 0.01167$). The `mcmc` method yielded $p = 0.0163$. It is unsurprising that the first cue in each triad has a larger effect than the others, given that this is the one participants would have read first, so it may have constrained the context of the remaining cues. For instance, in condition 1, reading 'snooze' first makes 'button' likely, but makes 'attack' (the top collocate of second cue 'panic') less likely.

An even more finely grained analysis was possible, using corpus-derived estimates of cloze probabilities (described above). A final LMEM, then, treats the probability value of the first cue (`cloze`) as the fixed factor, since the rank analysis above showed that only the first cue had a significant effect on insight rating. Again, no random slopes were included since they resulted in overparameterisation. Model estimates and p-values derived by `anova` and `mcmc` methods are given in table 6.2 and displayed graphically in fig. 6.2. Both methods show that `cloze` has a significant effect on `insight`.

	β	SE	t	p_1	p_2
intercept	49.859	4.404	11.322		
`cloze`	-23.530	11.080	-2.124	**0.03665**	**0.035**

Table 6.2: Parameters from the model based on cloze probability with p-values derived by `anova` (p_1) and `mcmc` (p_2) methods.

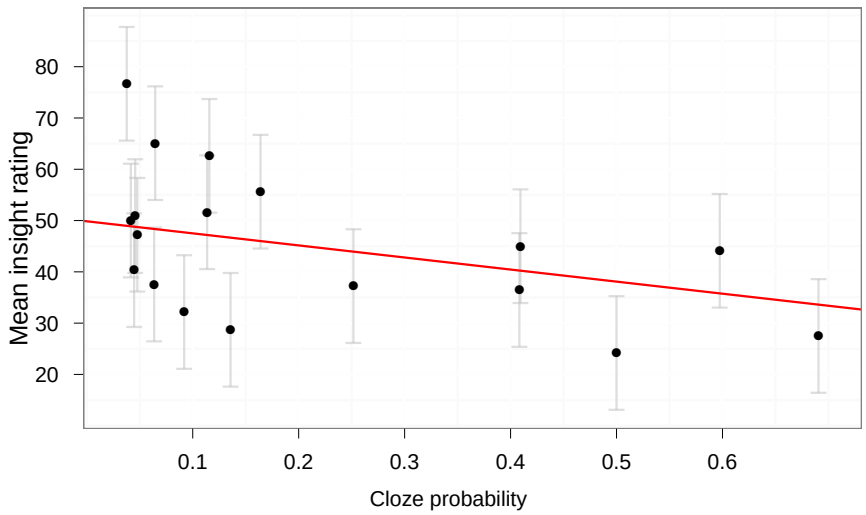


Figure 6.2: Mean insight ratings by `cloze` value of the first cue in each CRA problem, with standard errors and model estimate of the slope.

6.4 Discussion

Humans are easily able to make context-based inferences about the meaning of a word. I argued that context is one of the main dimensions along which inferential complexity depends (ch. 2) and that abduction is contextually unconstrained, unlike induction (ch. 3). The previous experiment explored context size, while this one explored levels of predictability from context. Participants solved CRA problems with varying degrees of predictability. An LMEM analysis found a significant effect of predictability on insight rating for the first cue in each problem, but not for the others. This, I argued, is unsurprising given that reading the first cue may reduce probabilities of collocates of the following cue.

The less predictable meaning is from context, the larger the role that

abduction plays in generating hypotheses about meaning. People do sometimes solve CRA problems mechanically, but these results show that the less predictable the answer, the less likely it is that a mechanical (and thus inductive) account is sufficient. If insight ratings in fig. 6.2 are compared with those for inductive and insight problems in ch. 4, it is clear that only high `cloze` values are found with anything like the insight rating of inductive problems there (which averaged 27.378). To reiterate, a given problem is not either purely abductive or purely inductive, but rather places different burdens on each, according to features of context. The present experiment demonstrates that predictability from context is one such feature.

In §2.6.4 I reviewed experiments investigating hemispheric differences between fine- and narrow-coding in semantic networks. For instance, Faust and Chiarello (1998) tested priming effects of ambiguous words on dominant or subordinate associates. The ambiguous words were situated in a sentence context that prompted either a dominant or subordinate reading. RH facilitation was found for both dominant and subordinate associates, regardless whether the context of the prime prompted a dominant or subordinate reading. LH facilitation was found only for associates of whichever meaning was prompted by sentence context.

Faust and Kahana (2002) constructed CRA-like triads made up of ambiguous words as primes for a target that was an associate of all the words. The triads were grouped into 5 conditions depending on whether the target was related to the dominant (D) or subordinate (S) meaning of each word in the triad: the triad could consist of DDD, DDS, DSS or SSS words (or unrelated words, to provide a base-line for comparison of the facilitation effect). At a low SOA (800ms), LH facilitation increased with the number of Ds in the triad, while RH facilitation was found for divergent (DDS or SSD) but not convergent (DDD or SSS) primes. At a high SOA (2500ms), statistically significant LH facilitation was found only for convergent dominant primes (DDD), while significant RH facilitation was found for all conditions.

Kircher et al. (2001) conducted an fMRI study of participants responding to sentences in three conditions: in one, participants read a completed sentence; in another, they chose between two possible endings; in the third, they generated a plausible ending. All sentences provided medium, low or very low constraint. The results show significantly more activation for the generation task in the RH temporal lobe. While the experiments by Faust

and colleagues show that the brain hemispheres respond differentially to varying levels of constraint, this shows the particular importance of the RH temporal lobe in hypothesising a meaning to fill a gap.

However, Kircher et al. (2001) refer to 'inference' in general terms, and don't mention insight at all, though Jung-Beeman et al. (2004) show that insight involves activation in the same RH temporal lobe areas investigated by Kircher et al. What my results add, then, is two-fold. Firstly, I show that the relevant kind of inference is insightful (and thus abductive). Secondly, I show a graded relation: lower constraint or predictability involves more abduction, while Kircher et al. (2001) do not compare different levels of predictability. They mention that the RH ability to deal with divergent primes might be related to creativity, but the present experiment offers more specific claims: the RH processes in question are related to creativity in that they provide abductive hypotheses through insight. Naturally, a lot remains to be done to understand just how this works. But the results so far support a central theme in this dissertation: induction must be complemented by abduction to understand open-ended word learning, though there is a tendency in the field of language evolution to suppose that induction on its own is sufficient.

In open-ended contexts, Bayesian induction is intractable (Kwisthout et al., 2011). A typical Bayesian response (e.g. Griffiths et al., 2010) is that their accounts concern the computational level, and that something at the algorithmic level probably approximates Bayesian calculations. Kwisthout et al. (2011) argue that this fails to solve the problem (and I agree), but the results here offer something positive by clarifying what happens at the algorithmic level. Contextually unconstrained processes are handled differently from contextually constrained processes, so rather than trying to work out how an algorithmic-level process approximates Bayesian induction, we must admit that there are (at least) two algorithmic-level processes that play a parallel role, and that one of them (the contextually unconstrained RH process) is not an approximation of Bayesian induction (§3.4, §3.5). The question of whether the other, contextually constrained LH process is an approximation of Bayesian induction does not concern me here.

These results also have implications for human evolution. Most of the focus in terms of inference has been on our understanding higher-order relations, and most of the focus in terms of language has been on classic

LH perisylvian regions or, in the case of Deacon (1997), the frontal lobe. However, given that crossing the symbolic threshold required contextually unconstrained inferences, and given that RH insight areas in the temporal lobe excel at these, the traditional foci must be broadened if we are to understand how humans communicate the way we do.

Chapter 7

Iconicity and Precedence in Word Learning

7.1 Background and Aims

So far, I've shown that hypothesis generation about meanings in unconstrained contexts requires insight and thus abduction. The results have thus supported my claim that abduction was essential at the symbolic threshold, since our ancestors at that point would have lacked conventional forms of directing attention or constraining context in other ways. But these claims are based on the context of the sign, not on the sign itself. Both aspects are important. In this chapter, I thus turn to examine the process of symbolisation: how a motivated sign becomes conventional over time.

A number of researchers in language evolution highlight the role of iconicity in language evolution, or language generally. Cuskley and Kirby (2013) and Kita et al. (2010) argue for the existence of an iconic or sound-symbolic protolanguage. Zlatev (2008) argues that iconic gestures played a dominant role before we developed conventional, normative signs. Perniss et al. (2010) argues that iconicity is still a general feature of language today. My focus here, though, is on how iconic signs can become symbolic through a process of symbolisation: the motivated elements of the signs are gradually reduced or lost. Garrod et al. (2007) and Fay et al. (2010) investigate the role of communicative interaction in symbolisation by having participants play repeated Pictionary-like games with the same cues. They found that, if feedback was allowed between drawer and guesser, initially elaborate iconic

representations would gradually become simpler and more abstract, until they no longer resembled their referents. That is, they became symbolic in the sense of 'non-iconic'. For an illustration (from the present experiment), see fig. 7.1 below.

This sense of the word 'symbol' is different from mine (cf. ch. 1), but my definition still applies to this scenario, even though my definition cuts across the distinction between icon and this sense of 'symbol'. I defined a symbol as a sign whose ground requires relevance-, context- or salience-deciding inference when it is *first* learned. Each stage in fig. 7.1 below is thus symbolic in my terms because the first stage requires such inference: relevance-deciding inference is needed to see the first picture as a character wearing a hat and carrying a whip and to see Harrison Ford as an actor who played a character who wore a hat and carried a whip. By the time the iconic picture has become more abstract, the guesser no longer has to make this inference about the ground. So though I consider all stages in fig. 7.1 symbolic, the latter stages are conventional symbols with a severely reduced iconic element. I argued that conventionality is not the central criterion for something being symbolic, so though most symbols are conventional, this doesn't preclude some from also being iconic.

My argument here, then, is that abduction is needed for the first stage in the above process, and that if such a process explains where symbols come from, then abduction was necessary at the symbolic threshold. Abduction is gradually required less and less by participants in the above process, until the final stages become a matter of straightforward recall: recognising a sign seen recently with precisely the same meaning.

If someone joined this toy 'speech community' at a late stage in the process, they would still have to infer the meaning of a non-iconic sign (while those participants present since the start would simply recognise it). If the inference of the late joiner is based on contextual information, and if that context is unconstrained by the environment or interlocutor, it would still require abduction. If, on the other hand, the context is highly constrained (as it often is in inductive experiments), the late joiner would only need to place a light burden on abduction. The previous experiment has already demonstrated the need for abduction in unconstrained contexts, so the focus in this chapter is on the process of symbolisation, not on late joiners.

My claims about abduction at the symbolic threshold are compatible

with, but do not require an iconic stage. My overall point is that, iconic origin or not, inferences about meaning at the symbolic threshold would have been context-, salience- or relevance-deciding (§1.5), and that a purely inductive account of the symbolic threshold is thus incomplete.

Here, I investigate the change in insight ratings over the course of a Pictionary-like experiment based on Garrod et al. (2007); Fay et al. (2010). Though the details of their methodology vary, the basic process is that pairs of participants play a communicative game. One, the drawer, reads a cue and begins drawing a picture to help their partner guess the cue. The guesser has a list of possible answers to aid them. Depending on experimental condition, interaction between the drawer and guesser may be allowed. In some conditions, interaction might involve the guesser being allowed to add to the drawing with a different colour pen; in other conditions, minimally informative interaction was limited to the guesser indicating when they had guessed the answer. After working through a list of cues, drawer and guesser switch roles. The process is repeated over a number of rounds with the same cues. Without any interaction, the pictures remain iconic across rounds; with some interaction, the pictures grew simpler and more abstract as in fig. 7.1.

The main addition I make to this methodology is that I required whichever participant was the guesser on a given round to provide an insight rating for each guess according to the 'Aha!' criterion discussed in previous experiments. Since this communication game was played on paper, not on a computer, the 100-point on-screen slider was replaced with a 7-point scale. I predict that insight ratings would be high in response to novel signs, but would drop as the participants become familiar with the signs. To control against the possibility that fatigue or boredom might cause a decrease in insight ratings, one item on the drawer's list of cues was replaced every second round. This meant that even in late rounds there were some novel pictures, and I predict that these would initially show high insight ratings which would then drop like the other, familiar signs.

In the above experiments, guessers have a list of possible answers. This constrains the context somewhat. In order to provide a less constrained context without departing from the established methodology entirely, I expanded the 20-item list of Garrod et al. (2007) to a 25-item list of possible answers. The guessers were given this list in a 5×5 grid to guess the cue.

However, a short-list of 10 items randomly selected from this long-list constituted the drawer's list for round 1. As described above, one item from the short-list was replaced from the long-list every two rounds, so by the 8th round, the guesser would only have seen signs for 13 of the 25 possible items. While the guessers knew what the possible answers were, they didn't know just which cues they would be seeing in a given round. The context here, then, was comparatively unconstrained, firstly due to the larger set of possible answers, and secondly due to uncertainty about which items from the long-list would appear in the game.

To avoid confounds arising from communication between drawer and guesser, interaction was limited to the word 'stop', which the guesser said when they thought they had the answer. No feedback was provided, so the guesser wouldn't know whether their guess was right.

According to my definition of 'symbol', and given that abduction plays an important role in generating novel hypotheses, the prediction is that insight will be higher for novel signs and lower for more familiar signs. A sign's degree of iconicity should not play a large role in determining insight ratings, except that signs tend to grow less iconic as they become familiar. But then iconicity and insight ratings are both dependent on degree of novelty or precedence, but are independent of each other: it is not the case that iconicity intermediates between novelty and insight ratings.

To test this, a second study was conducted. An independent group of participants (who had not participated in the Pictionary game) rated pictures produced in the Pictionary game for precedence (the degree to which it resembles the picture drawn in the previous round for the same cue) and iconicity (the degree to which it resembles its referent). The more novel the sign, the lower the expected ratings for precedence, while familiar signs would be very similar to the corresponding picture in the previous round and thus receive high precedence ratings. I predict that precedence should have a statistically significant effect on insight rating but that iconicity would not.

7.2 Study 1: insight in graphical communication

7.2.1 Methodology

7.2.1.1 Design

A within-subjects design was used. Pairs of participants played 8 rounds of a Pictionary-like game, alternating drawer and guesser role after each round. Each participant thus had 4 turns at each role. After making each guess, the guesser recorded an insight rating (`insight`), the dependent variable. The fixed effect, then, is `turn`; plausible random effects include `subject` (nested inside `pair`) and `item`.

7.2.1.2 Participants

22 University of Edinburgh students were recruited via the university's jobs website (5 male, 17 female; mean age = 20.08, SD = 1.86). All participants gave informed consent and were paid £6 for their participation. One pair was excluded from analysis because one participant wrote English words on her drawings, despite instructions to the contrary.

7.2.1.3 Materials

Garrod et al. (2007) and Fay et al. (2010) used a list of 20 items divided into 5 categories (places, people, entertainment, objects, abstract), either with 4 easily confusable items each, or with 3 easily confusable items and 1 distractor. I removed distractors and expanded each category to include 5 items. The original list of places, for instance, included `art gallery`, `parliament`, `museum` and `theatre`. To this, I added `university`. The original list of people included `Arnold Schwarzenegger`, `Brad Pitt`, `Hugh Grant` and `Russell Crowe`. `Hugh Grant` was the distractor, presumably because of the genres he appears in (this is not made clear in Garrod et al.). Instead, I used a list of 5 actors who are all potentially confusable in that they typically do not appear in romantic roles. Otherwise, the distractor in each list would have been salient and the others non-salient along a fairly obvious dimension, and salience would have been a likely confound. The list of 'Abstract' concepts in Garrod et al. was `loud`, `homesick`, `poverty`; in Fay et al. (2010), it was: `loud`, `homesick`, `poverty`, `sadness`. To avoid making any of these more salient than the rest, I used a list of nouns, rather than

a combination of nouns and adjectives. Since their lists contained many words of negative valence, I used only nouns of negative valence for the same reason. Since their 'Entertainment' category was just a list of genres, I expanded it to 5 items by adding `horror`. Their list of objects was rather disparate, but one thing many of the objects had in common was that they were dominated, visually, by a glass-like rectangle. I thus included `iPad` and `window` to maintain this visual uniformity.

The cues are given in table 7.1.

Places	Actors	Entertainment	Objects	Abstract
theatre	Robert De Niro	drama	television	noise
art gallery	Arnold Schwarzenegger	soap opera	computer monitor	depression
museum	Clint Eastwood	cartoon	microwave	poverty
parliament	Chuck Norris	horror	window	nausea
university	Harrison Ford	sci-fi	iPad	violence

Table 7.1: Items grouped by category.

Each participant in a pair had a response booklet. Each page in the booklet corresponded to one round, and the page for that round indicated which role (drawer or guesser) each participant was to take.

A shortlist of 10 items was randomly generated for each pair from the longlist of 25 items. These, in a randomised order, constituted the list of cues for the drawer for round 1. The list order was randomised again between each round. Every second round (rounds 3, 5 and 7) one item from the shortlist was replaced by another from the longlist, so by the end of the experiment, 7 items would have been present for all 8 rounds; 1 item for 6 rounds; 1 item for 4 rounds and 1 item for 2 rounds.

For each round, the guesser's response form had spaces for them to record their guess for each picture, in addition to a 7-point scale along which they had to provide an insight rating (by circling the relevant number on the scale) indicating the extent to which they had an 'Aha!' moment for that picture. The scale increased from left to right, such that circling 7 indicated a strong feeling of insight.

7.2.1.4 Procedure

The participants read instructions which described insight in terms of an 'Aha!' experience, comparing it to a light bulb suddenly flashing on in the head. They then had the chance to practice 4 Compound Remote Associate

(CRA, see ch. 4) problems from Bowden and Jung-Beeman (2003b) and were told that insight is a common (but not certain) experience when solving such problems. Participants were instructed not to include any English words in their drawings, or numbers, letters or other conventional signals. They were able to ask questions if unclear about any aspects of the experiment. When both participants indicated they were ready to proceed, they were given the list of possible answers and allowed a minute to familiarise themselves with the items on the list.

The drawer then read the first cue on the shortlist of 10 cues, and began drawing a picture to help the guesser infer the cue. The drawer was required to continue drawing until the guesser said 'stop', indicating that they had guessed the cue. The correctness of the guess was not confirmed. At that point, the guesser recorded their guess on their response sheet and recorded an insight rating indicating to what extent they had experienced an 'Aha!' moment. The drawer then proceeded to the next item on the list. This process repeated until all 10 items on the list for that round had been guessed. At that point, both participants turned to a new page, which indicated that they should swap roles. The procedure for each round was identical, apart from the role played by each participant. Two such rounds constitute a turn, since each participant played each role once per turn.

7.2.2 Results

Fig. 7.1 shows the series of pictures produced by one pair of participants for cue `Harrison Ford`, exemplifying the process of symbolisation found in Garrod et al. (2007).

On average, participants guessed correctly on 83.875% of problems. Since new problems were introduced throughout, but no feedback apart from the word 'stop' was allowed, they were unable to be perfectly accurate. Table 7.2 breaks these values down by `turn`. The experiment sought to measure participants' subjective experience of guessing, not guessing correctly (given that hypothesis generation is non-demonstrative), so insight ratings provided for guesses that turned out to be wrong were not excluded from analysis[1].

I first analyse the subset of the data for those items present across all 4

[1]It turns out that excluding incorrect guesses strengthens the statistical significance of the predicted effect, so this is a conservative move.

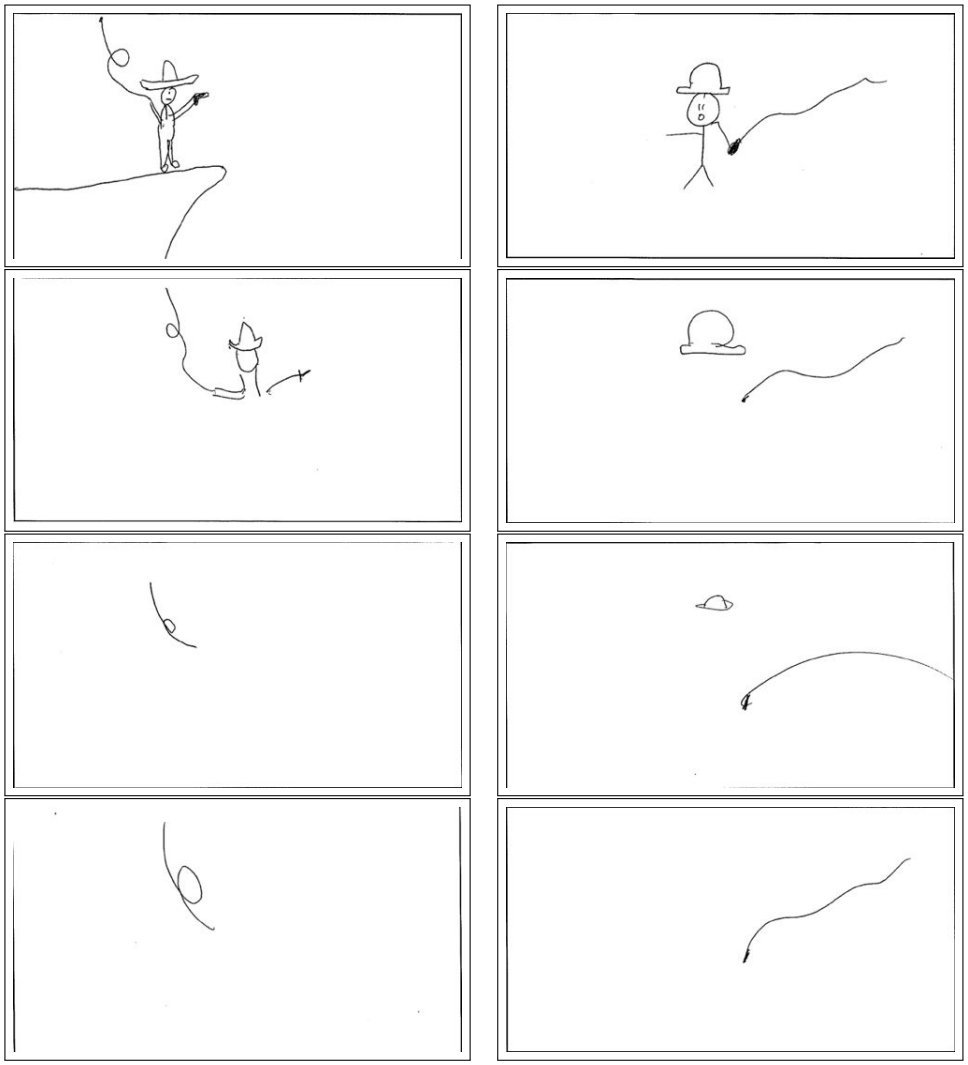


Figure 7.1: Pictures representing cue `Harrison Ford` across eight rounds for one pair of participants. Rounds increase left-to-right and top-to-bottom, in which case the left column shows pictures by one participant (the drawer in rounds 1, 3, 5, 7) and the right one shows pictures by the other participant (the drawer in rounds 2, 4, 6, 8). Each row thus constitutes one turn.

Turn	Correct (%)	sd
1	75	43.41
2	88	68.41
3	86.5	34.26
4	86	34.78

Table 7.2: Percentage guesses correct per turn, with standard deviation.

	β	SE	t	p_1	p_2
intercept	3.22689	0.24920	12.949		
`turn`	-0.27257	0.09295	-2.932	**0.006208**	**0.001**

Table 7.3: Model parameters and p-values derived by the `anova` method (p_1) and `mcmc` method (p_2, using a reduced model without random slopes).

turns to investigate the effect of `turn` on `insight`. Then I include items introduced during turns 2, 3 and 4. For the first analysis, a linear mixed-effects model (LMEM) was constructed using the `lme4` package (Bates et al., 2011) in R. The model included `turn` as the fixed effect, with random intercepts and slopes for `subject` and random intercepts for `item`. Random slopes for `item` and the inclusion of word `category` and participant `pairs` as random factors led to overparameterisation since the relevant correlation matrices included values of 1 or −1 (Baayen et al., 2008). The model parameters are given in table 7.3 and displayed graphically in fig. 7.2. These show that `insight` dropped significantly as `turn` increased.

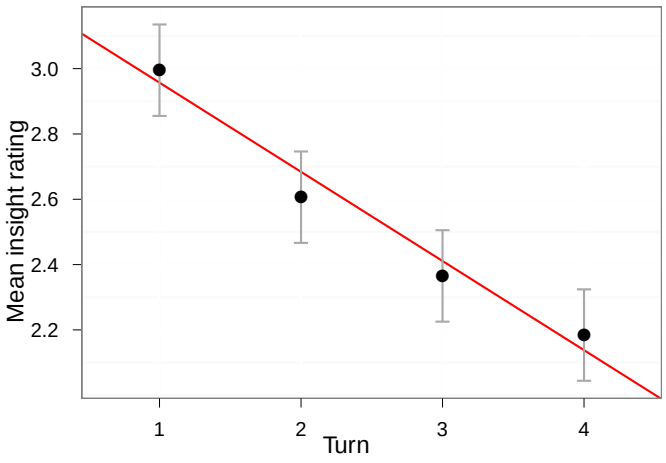


Figure 7.2: Mean insight ratings by turn with standard error. Slope from the LMEM in table 7.3.

Fig. 7.3 adds insight ratings for those items that were introduced as replacements in turns 2, 3 and 4. The graph indicates that, when new items were introduced, they initially received high insight ratings, before those

	β	SE	t	p_1	p_2
intercept	3.0202	0.2430	12.427		
`time`	-0.3136	0.0913	-3.436	**0.00182**	**0.0001**

Table 7.4: Model parameters and p-values derived by the `anova` method (p_1) and `mcmc` method (p_2, using a reduced model without random slopes).

ratings dropped as for the original items. Since turn 4 was the final turn, the data for items introduced in that turn are represented by a point, not a line. To analyse this larger data set, instead of using `turn` as the fixed effect, the time the item had been in the game was calculated (`time`). By the end of turn 3, for instance, an item that had been present since the start of the game would have been around for 3 turns, but an item introduced in turn 2 would have been around for just 2 turns, and an item introduced during turn 3 would have been around for 1 turn. An LMEM of this expanded data set was constructed replacing `turn` with `time`. The parameters and derived p-values are summarised in table 7.4. These show that `insight` decreases significantly as `time` increases and, indeed, the effect has been strengthened relative to the previous model: $|t|$ has increased from 2.932 to 3.436.

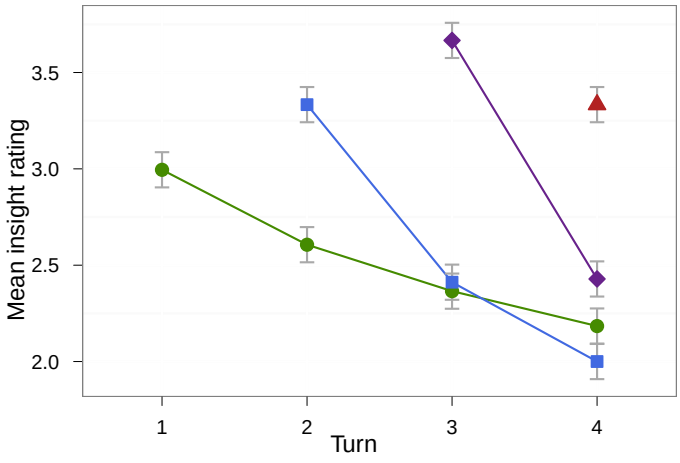


Figure 7.3: Mean insight ratings by `turn`, grouped by the number of the turn in which the items first appeared. Error bars indicate SE from the LMEM in table 7.4.

If it was the case that a decrease in `insight` over `turn` was explained by

boredom or fatigue, we would expect new additions in later turns to receive lower insight ratings than new additions in earlier turns, and similar insight ratings to familiar signs in the same later turn. If, on the other hand, the decrease in insight ratings is explained by novel signs becoming familiar, then newer additions in later turns should receive similar insight ratings to new additions in earlier turns, and higher insight ratings than familiar signs in the same later turn. The matter thus boils down to whether `turn` or `time` are better predictors for the full data set. An `anova` comparison between the two models of the full data set (one with `turn` and one with `time` as fixed effects, random effects as for models described above) shows that the one with `time` is a significantly better fit (χ^2 = 57.562, df = 0, $p < 0.001$). The graph suggests just the opposite of a fatigue effect, in fact. It's plausible that in later turns, a novel cue was salient compared to cues that had been present for a few turns already.

7.3 Study 2: iconicity and precedence

7.3.1 Methodology

7.3.1.1 Participants

Participants were recruited via Amazon's Mechanical Turk (MT) crowd-sourcing platform to provide ratings of pictures from the previous study. Given the nature of MT, participants were able to provide ratings for one picture or for several pictures as they wished. In other words, this particular crowd-sourcing task was based on number of items, not number of participants, and it is usual on MT for workers to try to perform as many tasks of the same type as they can. It turns out that the work was completed by a total of 84 participants, so most completed a large number of items (mean 52.2). Participants were assigned by MT to whichever was the next available item, so while many would have rated unrelated pictures, some would probably have rated different pictures produced for the same cue, and some would probably have rated different pictures produced by the same participant in the previous study. None of them were able, though, to respond repeatedly to the same exact task.

7.3.1.2 Design

The present study sought to test my prediction that a picture's degree of precedence (or its converse, novelty) should be a better predictor than its degree of iconicity for the insight rating provided for that picture in the previous study.

Given the limitations imposed by the nature of MT (described in the previous subsection, 'Participants'), this study cannot accurately be described as between- or a within-subjects. The focus here is on items, not participants, since I will be testing which property of a picture best predicts insight, and rather than investigating how different subjects rate precedence or iconicity.

Each picture present from round 2 through to round 8 in the previous study was rated for precedence by 10 MT participants and rated for iconicity by 10 MT participants, as described below. In the analysis, `precedence` and `iconicity` were both fixed effects, with `insight` produced during the previous study as the dependent variable.

7.3.1.3 Materials

Pictures produced during the previous study were scanned and scaled to yield jpg files (480×262 pixels).

7.3.1.4 Procedure

MT participants were presented either with a pair of pictures (for a precedence rating) or a picture and a cue word (for an iconicity rating). For a pair of pictures, participants provided a rating along a 7-point scale indicating to what extent the one picture resembled the other picture (its precedent: the drawing for that cue in the previous round). For a picture and a cue word, participants rated the picture on a 7-point scale indicating to what extent it resembled the referent of the cue. All participants were told that the pictures had been drawn in a Pictionary-like game, and were thus not supposed to be perfect representations — they were just meant to help someone guess the cue, and they should rate overall resemblance (to precedent or referent) with this in mind.

7.3.2 Results

To analyse the effects of precedence and iconicity, an LMEM similar to the previous ones was constructed, including `iconicity` and `precedence` along with `turn` as fixed effects. Since this study was limited to items present from rounds 2 through to 8, there was no distinction between `time` and `turn`. Random slopes for `subject`[2] and `item` relative to these fixed effects indicated overparameterisation, so the model included random intercepts only. The results are summarised in table 7.5. While `precedence` has a significant effect on `insight`, `iconicity` does not.

Fixed effect	β	SE	t	p_1	p_2
intercept	2.97146	0.23287	12.760		
`iconicity`	0.04572	0.03765	1.215	0.2467	0.228
`precedence`	-0.16372	0.03609	−4.537	**<0.0001**	**<0.0001**
`turn`	-0.29075	0.05538	−5.250	**<0.0001**	**<0.0001**

Table 7.5: Model estimates of fixed effects on `insight` with standard error (SE), t-values, and p-values derived by the `anova` (p_1) and `mcmc` (p_2) methods. Significant p-values are given in bold.

7.4 Discussion

Symbolisation, the process whereby an iconic signs becomes increasingly arbitrary, is one proposed route to symbolic communication. The focus in much research in this area has been on features of the sign itself (iconicity or arbitrariness) or on its use in social contexts (such as the role of feedback). However, an account of how our species evolved the cognitive skills to interpret signs at various stages along this process (whether iconic or symbolic) first needs an examination of what those skills were.

I have been arguing that abduction was crucial at the symbolic threshold because the threshold required pragmatic inference, and that abduction, unlike induction, is suitable for this contextually unconstrained process since it can operate insightfully. That is, my arguments thus far have been about cognitive differences across communicative contexts, not about cognitive differences across changes in the sign itself. In this experiment, I turned

[2] `subject` here refers to the participant in the previous study who provided the insight rating for the picture. Since the 10 ratings per picture were averaged, the MT participants in this study could not be included as random effects.

to examine the relationship between the nature of inference and the nature of the sign. The results show that abduction plays a crucial role in early stages of symbolisation, given that these had significantly higher insight ratings, and that such ratings are diagnostic of increased involvement by abduction, compared to induction. Abduction thus helped us across the symbolic threshold.

I predicted that signs at the initial stages of the process of symbolisation would require abduction, not because they are iconic, but because they are novel, and thus require inferences about relevance. The LMEM analysis shows that precedence has a significant effect on insight rating, but that iconicity does not. Signs with high levels of precedence are not novel; novel signs have low levels of precedence. This doesn't mean that iconicity plays no explanatory role in the evolution from a non-symbolic species to a symbolic species. Rather, it means that iconicity isn't what determines what kind of inference is needed for that to happen.

Fig. 7.4 shows seven participants' drawings for cue `Harrison Ford` for whichever turn that cue was introduced in. Many of them represent Harrison Ford as Indiana Jones (and Indiana Jones as a man with a whip and a hat). Some, however, represented him as a man whose last name is also that of a car. Similarly, fig. 7.5 shows representations of Clint Eastwood as a person who uses guns, and the first picture represents him as a man who is associated with places that have cactuses. Understanding each requires hypotheses about the ground of the sign. These are relevance-deciding and thus abductive.

Once that inference is made, however, relevance is established for future interactions. In the terms of Garrod et al. (2007), the symbols are 'grounded' in interaction: later pictures refer, not just to their referents, but also to earlier pictures. Once one realises that one's partner has represented Harrison Ford as the man with the whip, one no longer needs to infer the relevance of the whip in later rounds. In later stages in fig. 7.1, then, the relevance of the whip is recognised or recalled, rather than inferred. If, for whatever reason, symbolisation hadn't occurred and drawings had remained iconic throughout, insight ratings should nonetheless decrease across rounds as the icons become familiar.

As the picture becomes simplified (in fig. 7.1, being reduced to a minimal representation of the whip), interpreting the picture reduces to recognition

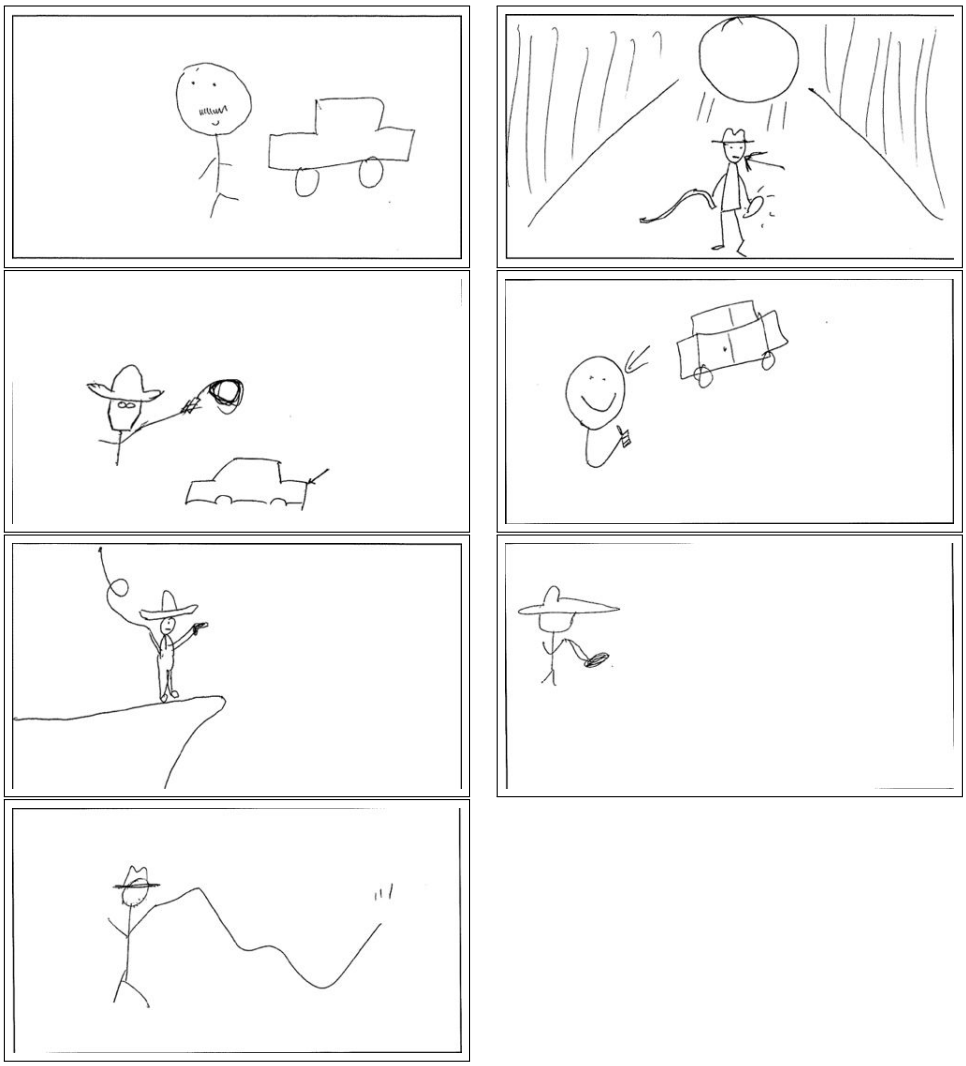


Figure 7.4: Representations of cue `Harrison Ford` drawn independently by seven participants in seven different pairs.

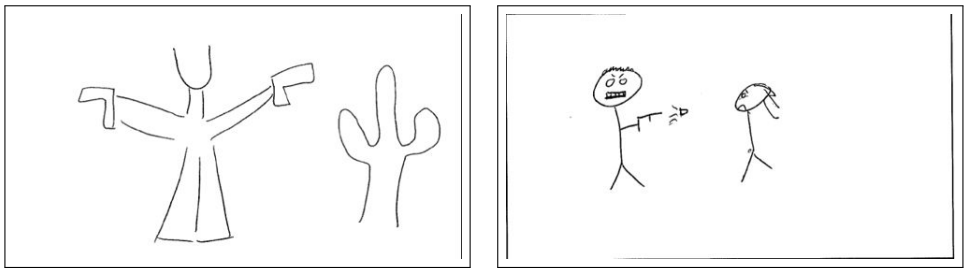


Figure 7.5: Representations of cue `Clint Eastwood` drawn independently by participants in two different pairs.

of the whip. That is, the whip either *is* a salient associate of Harrison Ford, or *becomes* a salient associate (or both). The difference lies in how we explain the whip's presence. If WHIP is a strong associate of HARRISON FORD in people's representational structures prior to the experiment, then that may explain why it appeared in so many groups' drawings in round 1 (fig. 7.4), and why simplified pictures in later rounds often focused on the whip (fig. 7.1). Call this the internal explanation. An alternative, external explanation is that dynamics of interaction are what explain why the whip became salient: perhaps if, in earlier rounds, participants (for whatever reason) said 'stop' more often as the whip was drawn than as other features were drawn, then that interaction might make the whip a salient feature of the picture.

The research by Garrod et al. (2007) and Fay et al. (2010) focuses on external processes, as does much research in language evolution (such as Kirby et al., 2008). But since abduction plays a role in symbol origins, and since it rests on plausibility, a matter of accessibility in representational structures (§3.5), this means that the process of symbolisation must be driven as much by representational structures as by social dynamics. How internal and external processes interact in sign development is a question for future research, which could, for instance, investigate correlations between word associations and judgements of salience in pictorial representations.

Another way in which representational structures play a role is in the matter of priors. In §3.4.1 I highlighted an issue with Bayesian accounts of such tasks: they require well defined hypotheses spaces, as well as prior probabilities and likelihoods for all hypotheses in those spaces, relative to

the data. Unlike in more realistic word-learning tasks, the hypothesis space here is given (though quite large). But having a constrained hypothesis space won't help one solve the problem if one cannot make the connection between the whip and Harrison Ford. The basic Bayesian response is to posit a high prior for the hypothesis that the whip indicates Harrison Ford, but if that is how we are to explain how people solve such problems in general, then we would have to allow stored priors for all possible representations of all possible concepts, which would require astronomical amounts of storage and indefinitely prolonged searches through that stored information.

Alternatively, hierarchical Bayesian models allow priors to be derived from representational structures instead of being directly represented (Griffiths et al., 2008, though I stressed in §3.4.5 that these structures are not themselves always explicable in Bayesian terms). If these representational structures mean that INDIANA JONES is highly accessible from WHIP[3] (or accessible from both WHIP and FEDORA), and if the same structures are what explain generation of the hypothesis that the picture refers to Harrison Ford (which is what an abductive account suggests), then induction simply has to evaluate this particular hypothesis, avoiding the need for storage of countless priors and lengthy searches through such priors. While Fay et al. (2010) explicitly allow inductive inference a role in symbolisation, these considerations show that inductive inference alone is insufficient: abduction would have been required too, especially in early stages of the process. Similarly, Garrod et al. (2007) highlight a few similarities and differences between their experiment and the description of the symbolic threshold in Deacon (1997). The results here suggest that insight should also be numbered among the similarities.

These results also offer an informative distinction between symbolisation and an easily confused term, ontogenetic ritualisation, which is sometimes seen in animal communication:

> Here, the learned action, and the reaction of another animal to it, both become stereotyped. A classical example is the 'nursing poke' where a baby chimpanzee starts by pushing his mother's arm aside to get at the nipple; after a while, the mother learns to raise her arm when he goes to move it, and the baby learns that

[3]Or that WHIP is highly accessible from Indiana Jones, an independent question, given that the drawer first saw the cue and then decided what picture to draw.

> he only has to touch her side to get the required effect. (Hurford, 2007, 199-200)

Interaction and imitation are features of the symbolisation process. Interaction is also found in ontogenetic ritualisation[4], and chimpanzees are capable of imitation to a limited extent (Whiten et al., 2009). It is unclear, then, why a species capable of ontogenetic ritualisation should not be capable of symbolic communication, if symbolisation is a source of symbols. Clearly, something else is needed.

Hurford (2007) suggests that this something else is partly a matter of empathy, trust and motivation, and while I agree that those are part of the picture, an additional important difference involves salience or relevance. Food is a biologically salient category, so it is unsurprising if processes of ontogenetic ritualisation apply to feeding behaviours. While animal imitation often thus relies on biologically salient categories, we are able to make complex inferences to *discover* what is salient or relevant in a complex interaction such as this experiment. We can insightfully infer relevance or salience in novel cases not constrained by biological categories, and I highlighted the role of salience in Lewis conventions and relevance in Millikan conventions (§1.4). Similarly, I've already discussed how some animals are capable of insight when it comes to food retrieval, while we are capable of insight domain generally (§3.5.2). Our ability to infer salience or relevance, then, underlies the difference between open-ended symbolisation and contextually constrained ontogenetic ritualisation.

In sum, abduction plays an important role in interpreting novel signs that require inferences about relevance or salience, and is thus required at earlier stages of the symbolisation process. If this process plays a role in explaining the origins of symbolic communication, these results mean that abduction was essential for crossing the symbolic threshold.

[4]Though Hurford (2007) stresses that a difference between this and human communication is that our interactions are reciprocal (either person in an interaction can be sender or receiver), while the chimpanzee mother's and baby's roles are asymmetric in the above example: this is an external difference.

Part III

Conclusions

Chapter 8

Conclusions

8.1 Primary argument

The primary argument of this dissertation has been that abduction was one of the cognitive faculties necessary for our ancestors to cross the symbolic threshold. I argued that symbolic communication requires pragmatic inference; that the requisite inference must have been salience-, relevance-, and context-deciding hypothesis about the ground of a sign; that such inference is comparatively complex in evolutionary terms; and that induction on its own is unable to account for word learning in novel unconstrained cases, as at the symbolic threshold. I tested Peirce's claim that abduction is insightful and found subjective reporting of an 'Aha!' moment to be diagnostic of increased levels of abduction, compared to induction. I then showed that meaning-guessing tasks require more insight (and thus abduction) when hypotheses are not given, when context size increases, when predictability from context decreases, or when signs are lacking in precedent. These were, I argued, features of the symbolic threshold that necessitated abduction.

8.2 Secondary arguments

Secondary to this main thread, I provided a detailed account of iconicity and convention as they relate to language evolution, and highlighted how an over-reliance on either description risks introducing explanatory gaps into an evolutionary account *if* one focuses unduly on perception or on the co-ordination or imitation of behaviour alone (as opposed to behaviour

and belief or intention). I acknowledged useful elements of each, though: direct iconicity provides analogical information that abduction can use in connecting representamens and objects, while indirect iconicity highlights the open-endedness of human interpretive processes compared to that of animals. Theories of convention explain how we manage to settle on anything at all, given this open-endedness.

I flagged up an incongruity in Relevance Theory: it is non-normative and places a large explanatory burden on accessibility (both features of an empiricist psychology), yet proposes that deduction is the mechanism of interpretation. Deduction is syntactic and normative and thus prototypically rationalist, which engages the Frame Problem (specifically the version of the Frame Problem that Fodor calls Hamlet's Problem). Replacing normative, syntactic deduction with non-normative, associationist abduction avoids this, and I reviewed a number of experiments that characterise pragmatic inference in empiricist terms.

The proposed inferential hierarchy, supported by behavioural and neurological evidence, gives contextual constraint a foundational role in determining inferential complexity. The hierarchy coheres with evidence of similarities and differences between human and animal cognition: many animals are capable of inference when the context is constrained by biology or experiment design. Some animals are sporadically capable of contextually unconstrained inference in domain-specific cases, particularly food retrieval. They are not generally capable of contextually unconstrained inference, however. The proposed definition of 'inference' also suggested a way to identify transitive inference cognition in animals.

I reviewed evidence supporting differences in hemispheric processing such that the LH involves fine, narrow spreading of activation while the RH involves coarse, broad spreading of activation. The LH dominates in circumstances where a word or meaning is predictable from context, or a constrained context is likely to provide coherence; the RH dominates when meaning is less predictable or a larger context is needed for coherence. Novel metaphors are more of a RH process which, given that understanding a metaphor involves inferring the ground, suggests interesting parallels with novel symbols. The discussion also showed that the RH temporal lobe (implicated in insight and creativity) is worthy of further attention in language evolution, though the frontal lobe has occupied much of the attention in this

field over the last couple of decades.

Novel, unconstrained problems were shown to represent both a practical and a principled limit on induction, since inductive attempts to deal with novelty run into problems with relevance, psychologically unrealistic hyper-theories, or innate theories. Innate theories of meaning can't explain how we evolved to cross the symbolic threshold; hyper-theories about meaning would have been unrealistic in a pre-symbolic species lacking any theory of meaning; relevance is needed to explain just which structures are extracted from semantic memory for placement in working memory; and in novel cases, there are no theories to constrain relevance. Analogy and insight, on the other hand, are well suited for dealing with novelty and unconstrained problems. Indeed, I reviewed evidence showing that analogy can determine salience and insight can determine relevance, and that both involve connections between representations or representational structures.

8.3 Directions for future work

8.3.1 Experimental

This dissertation focused on insight rather than analogy, but there is room for exploration of the role that analogy plays in hypothesis generation in iconic gesture and imitation, and of how analogy processes salience or relevance in these situations. I outlined how plausibility underlies hypothesis generation, and if plausibility is a matter of accessibility in representational structures, then it would be interesting to investigate interactions between salience in iconic signs and accessibility (for instance, in word associations). If a whip is central to most people's representation of Harrison Ford, and if representational structures determine salience and accessibility, then we might expect signallers' choices in producing a novel sign to correspond to people's word associations, and expect signs produced in accordance with accessibility to be more easily interpretable than those not. Conversely, I mentioned that 'dog' is the most common associate of 'pet', and if this was explicable by dogs being more prototypical pets, then it would be unsurprising if novel iconic gestures for 'pet' were more often dog-like than cat-like. On the other hand, dogs and cats are equally frequent, and an apple appeared in all three pictures for cue 'ripe' in ch. 5, so the study might investigate the extent to which there is a disjunction between salience and

accessibility on one hand and probability on the other.

I mentioned that the degree of similarity between George Clooney and Chuck Norris is context dependent. If a drawer or gesturer had to produce a novel sign for 'George Clooney', his decisions about salience or relevance would be quite different depending on whether the context included Toni Morrison and Alice Walker or Stephen Seagal and Jean Claude Van Damme. So manipulations of context would probably have an effect on the production of novel signs, which might in turn have an effect on interpretation.

A further study could research how salience, plausibility and probability interact across dyads as the task is iterated over time, or within dyads in cross-situational contexts. If, of two meanings, one is more salient and the other more probable cross-situationally, is it ever the case that the salient, less probable meaning would be hypothesised by participants? In an iterated learning task, would the most probable or the most salient sign (if different) be more likely to be transmitted over time?

My main measure of insight here has been self reporting, but there are a number of ways in which this could be extended. An individual differences study could measure how insightful participants are, and investigate whether insightful participants solve abductive problems more quickly or successfully than less insightful participants. I would expect that, in closed worlds, the difference between insightful and non-insightful participants would be smaller than in open worlds. A split-visual-field study could manipulate levels of novelty or predictability in a hypothesis-generating problem to see whether recognition of target words is facilitated more in the LVF/RH as novelty increases or predictability decreases. Going in the other direction, another split-visual-field study could investigate whether primes in the LVF/RH are more likely to affect hypothesis generation than those to the RVF/LH. Neuroimaging is a further possible direction: deduction, induction and insight have all been studied using various neuroimaging methods, but abduction has not.

8.3.2 Theoretical

While the above focus on the experimental side of cognitive semiotics, there is also much to explore on the theoretical side. I mentioned, for instance, how there seem to be various possibilities for points of contact between my inferential hierarchy and the mimesis hierarchy in Zlatev (2008). Research

might turn up further similarities, differences and disconnects, or overlapping or differing predictions about communication in phylogeny or ontogeny.

I explored Peirce's notion of plausibility here, but this is a very undeveloped area of cognition, and I think it merits further attention. While dual-process models (for instance, distinguishing associative from syntactic processes) are commonplace in cognitive science, it is less common to align these with two-step models (where the first step is non-normative conjecture and the second is rational evaluation of those conjectures). A further area for theoretical work, then, is the degree of overlap between syntactic and normative processes on one side and associationist and non-normative processes on the other, or the scope of non-normative processes more generally.

One particular topic in this area concerns the difference between compositionality and creative combination. If a compositional process puts familiar elements together, one expects the meaning of the composite to be predicable given the meaning of each element: if one understands 'John loves Mary', then one should understand 'Mary loves John'. On the other hand, functional flexibility rather than predictability is a hallmark of creative combination. One can know what a pole is and what a tin can is, but that doesn't mean one can predict the function of a can on a pole: one has to invent or discover that function. In one experiment reviewed in ch. 3, a participant intended this combination to be a bird nest. As the word 'intend' there suggests, this is more pragmatic than the compositional example, and it is also conjectural, plausible, or non-normative.

I think it would be quite straightforward to make the case that creative combination is what underlies novel metaphor as opposed to conventional metaphor and syntax. 'Bright student' is processed in classical LH language areas, as is syntax; 'pearl tears' is processed in creative RH language areas which are not sensitive to syntactic constraints, but which are responsible for functional flexibility. If there is a continuum between novel and conventional metaphor, then it is plausible that there is a continuum between compositionality and creative combination based on degree of predictability.

But two questions then raised are **(1)** whether any other aspects of language involve creative combination rather than compositionality and **(2)** what the evolutionary trajectory here was. I think it at least plausible that, both early on in child development and shortly after the symbolic threshold, the meanings of simple combinations of symbols needn't have been

mechanically predictable. If they weren't predictable (or not entirely predictable) this suggests we cannot assume they were entirely compositional (at least, the matter needs consideration). Conversely, if one were to argue that our species' earliest two-sign combinations were entirely compositional, this might commit one to the claim that hemispheric differences were already well developed by that point, which is quite a strong claim, one which requires evidence.

Finally, both the inferential hierarchy and the two-step dual-process model offer a framework for understanding non-experimental evidence, such as the archaeological record. 'Symbol' in archaeology often refers to artistic expression and is sometimes partly iconic. Such symbols are often used as evidence of cognition about abstract meaning, but they are also evidence of open-ended conjectural processes. My framework offers a clear account of how iconicity and open-ended creativity or conjecture are related to cognition about meaning. Conversely, I wouldn't expect the producers of Oldowan tools, which remained relatively unchanged over hundreds of thousands of years, to have been skilled in abductive inference.

Though these proposals are speculative (though anchored in points I have made in the body of the text), I hope they suggest that the introduction of abductive inference into the debate about human evolution has the potential to open up a range of avenues of research, reframe current pieces of evidence, and engage with a number of topics currently of interest.

Appendix

Sources or preparation of stimuli are described in the 'Materials' section of the relevant chapter.

Stimuli for Experiment 1 (ch. 4)

Abductive problems

Complex causation

1. You're walking down a road. Up ahead the whole road is wet. A tumbleweed blows across it, before coming to rest in a pile of garbage beside the road. It is mid afternoon. Why is the road wet?

2. You see a bald teenage boy. He seems to have a healthy complexion, and is wearing a smart suit. The suit is grey pinstripe. Why is he bald?

3. It's been quite a sunny July so far. You see a tree with dead leaves all around it on the ground. The tree seems free from mould and rot. There are no other trees around. Why are there dead leaves on the ground?

4. You see birds circling in the blazing heat, but you can't see any dead animals in the grass below. Why are they circling?

Simple causation

1. You read about a plane crash. Why did it crash?

2. You see a building collapse. Why did it collapse?

3. You hear a loud noise. What caused it?

4. Your car isn't where you left it. Why is it gone?

Complex motivation

1. Someone comes into the room and lights a candle. Everyone who lives in that apartment is single. In the corner, someone else is watching a DVD on their laptop, looking bored. Why did someone light a candle?

2. Your teenage son is acting even more weirdly than usual. The school's mandatory drugs test came back negative, and he seems happy with his girlfriend. He did well in last week's math test. Why is he acting weirder than usual?

3. A colleague said she couldn't come to your party because she's busy, but then you hear from a mutual friend that your colleague mentioned not having any plans that night. You're pretty sure she doesn't dislike you. Why did the colleague say she couldn't come?

4. A friend of yours usually drinks quite a bit, but not enough to be considered an alchoholic. One day, you offer them a drink, and they refuse. They refuse the next day, too. You've known them a long time, and they haven't mentioned any bad news recently. Why are they refusing drinks all of a sudden?

Simple motivation

1. You see a woman shouting at a sales assistant in a shop. Why is she shouting?

2. A colleague arrives for a very formal business meeting in a T-shirt and shorts.

3. Your neighbour walks out his front door, but immediately turns to go back in. Why did he turn back?

4. A friend calls you to suggest going to the zoo today. Why did they decide to go to the zoo today?

Alien world

1. You arrive at an open space in a Zorg town. Lots of Zorgs are milling about. One Zorg is coming through the crowd, which opens up to let it pass. Its clothes are in rags and it is quite tall. The other aliens pour water in its path, and bow as it goes past. Who do you think this Zorg is?

2. You're in the middle of a crowd of Zorgs, all talking quietly to each other. A Zorg with a black shirt arrives, and the others become quiet as it passes. It takes something you can't see out of its pocket and puts its hand on someone's shoulder. The crowd starts talking again, nervously. Who is the Zorg in black?

3. You see a Zorg lying twitching on the floor. Another Zorg with a red stripe on its arm kneels down beside it and holds a strange device to the first one's head. It also attaches a small wriggly thing to its leg. The Zorg on the floor seems to sag. The wriggly thing turns blue. What's going on?

4. A Zorg in a purple hat is surrounded by a ring of other Zorgs. They're making a pleasing sounding noise in harmony, and its face flushes a

gentle orange. They give it something covered in small blinking lights, which flash on and off in what seems a random pattern. It takes out a blade and looks expectant. What's going on?

Insight problems

Classic insight problems

1. How much earth is there in a hole 1m long, 1m wide and 1m deep?
2. Marsha and Marjorie were born on the same day of the same month of the same year to the same mother and the same father, yet they are not twins. How is that possible? Type your answer, then press enter.
3. You just have a candle, some matches, and a box of tacks. How can you support the candle on the wall?
4. A man in a small town married 20 different women of the same town. All are still living and he never divorced. Polygamy is unlawful, but he has broken no law. How can this be?

Rebus problems

1. poPPd
2. $\frac{exit}{leg}$
3. you just me
4. PUNISHMENT

CRA problems

1. cottage
swiss
cake
2. dream
break
light
3. right
cat
carbon
4. tooth
potato
heart

Analytic problems

Induction type 1

1. Which of the following is the most likely explanation for someone looking tanned, given that they're from Alaska?
   a) They've been on holiday b) They've been to a tanning salon
2. Which of the following is the most likely explanation for a man giving his wife flowers, given that they've been married for 40 years?
   a) He's apologising for infidelity b) It's her birthday
3. Which of the following is the most likely explanation for a cough, given that the patient is a smoker?
   a) Emphysema b) A cold
4. Which of the following is the most likely explanation for someone being arrested, given that they have a master's degree?
   a) They've committed fraud b) They've committed murder

Induction type 2

1. Of all possible reasons a friend might be late to meet you, how likely is it that missing their bus is the right explanation?
2. Of all possible reasons for a car crash, how likely is it that running a red light is the right one?
3. Of all possible reasons for a train being cancelled, how likely is it that snow is the right one?
4. Of all possible reasons for a woman collapsing in the street, how likely is it that being drunk is the right one?

Induction type 3

1. Sparrows have procampsial arches; eagles have procampsial arches; penguins have procampsial arches. How likely is it that all birds have procampsial arches?
2. Lions have decuspid molars; tigers have decuspid molars; leopards have decuspid molars. How likely is it that house cats have decuspid molars?
3. Cow guts contain the enzyme protylase. How likely is it that all herbivores' guts contain the enzyme protylase?
4. Bats' ears have protympanic membranes. How likely is it that all mammals' ears have protympanic membranes?

Deduction

1. If Zeno is Cretan and most Cretans are liars, is Zeno a liar?
   a) Definitely b) Possibly c) Definitely not.

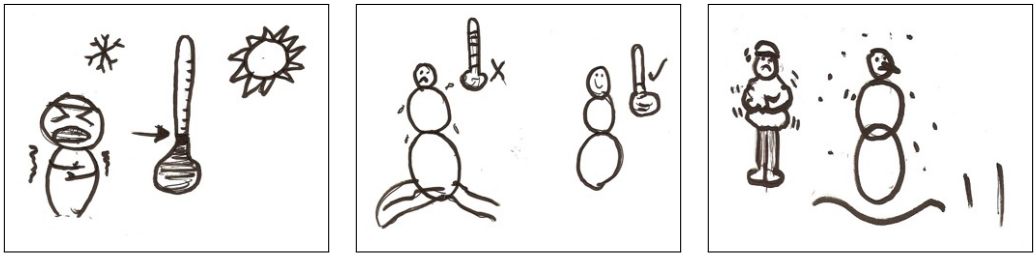
2. If Achilles isn't faster than a tortoise, and a tortoise isn't faster than an arrow, is Achilles faster than an arrow?
   a) Definitely b) Possibly c) Definitely not.

3. All philosophers are wise men and Parmenides is a philosopher. Is Parmenides a wise man?
   a) Definitely b) Possibly c) Definitely not.

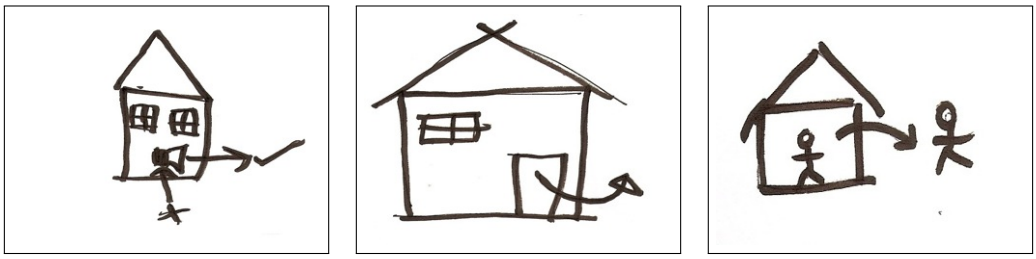
4. All men are mortals and Cratylus is a mortal. Is Cratylus a man?
   a) Definitely b) Possibly c) Definitely not.

Stimuli for experiment 2 (ch. 5)

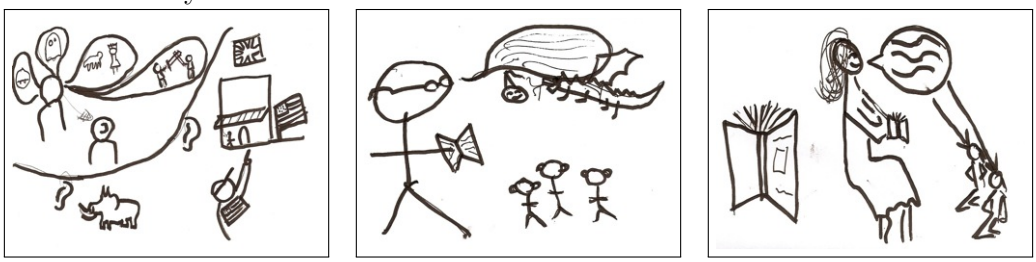
cold



outside

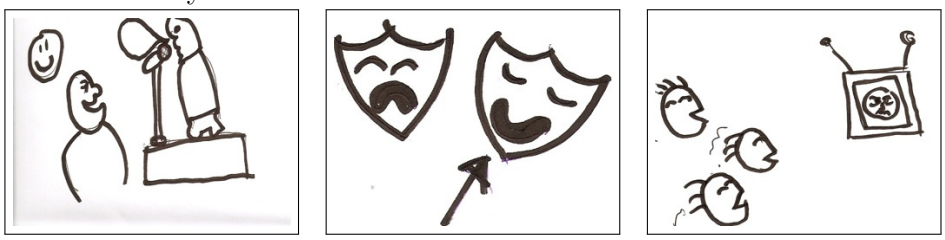


story

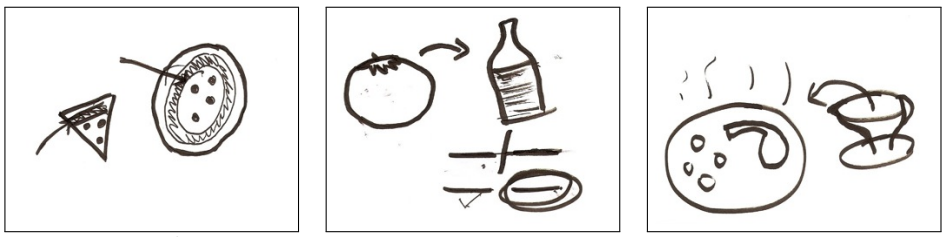


prescription

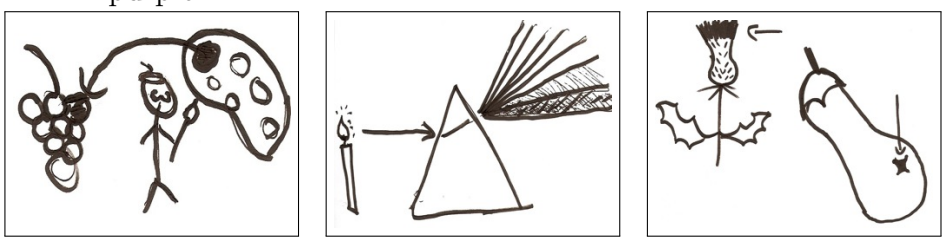
comedy



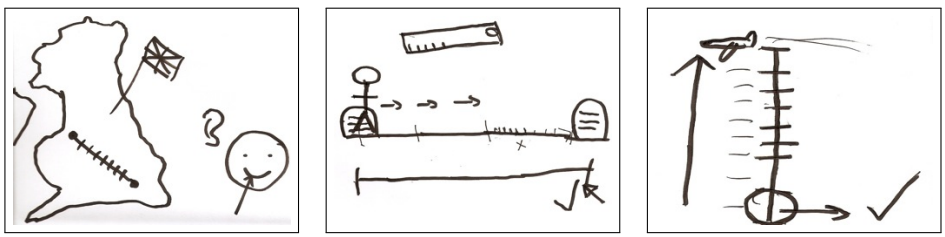
sauce



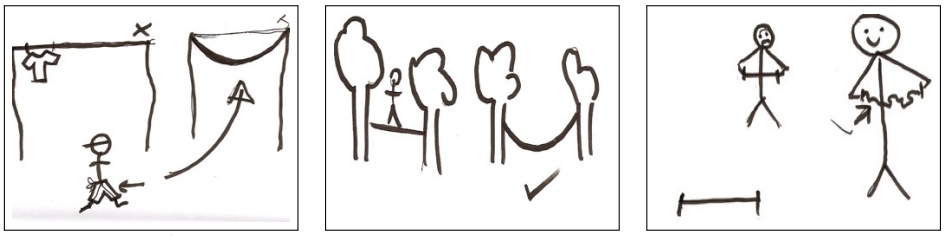
purple



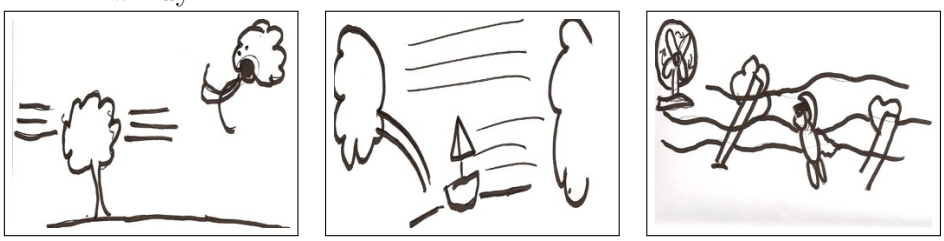
mile



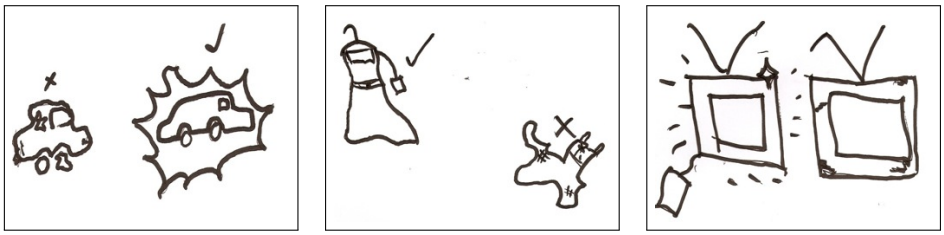
slack



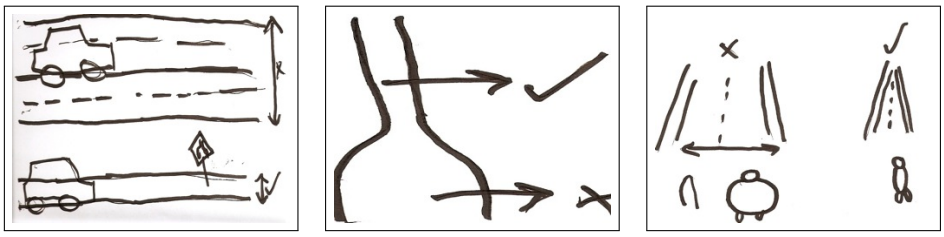
windy



new

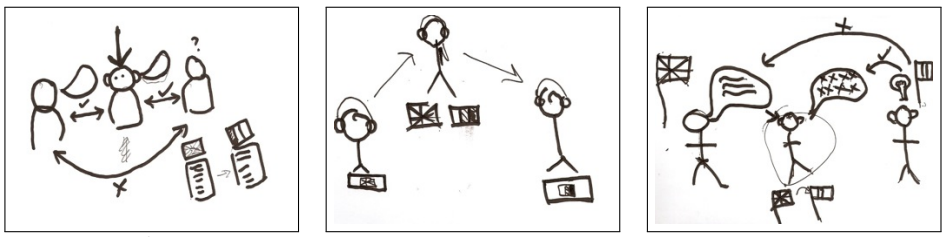


narrow

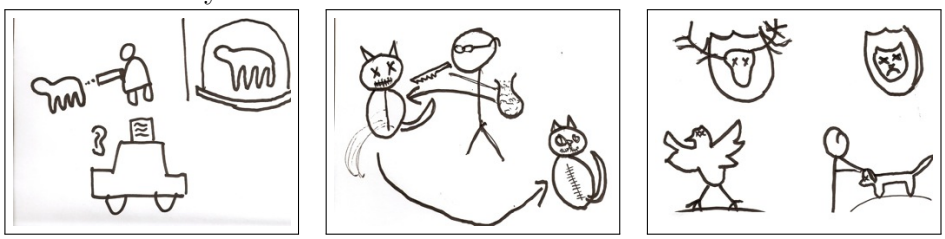


ripe

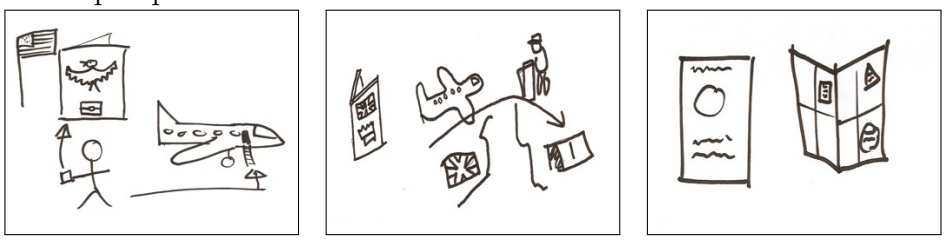
translate



taxidermy



passport



Bibliography

Abdullaev, Y. G. and Posner, M. I. (1997). The course of activating brain areas in generating verbal associations. *Psychological Science*, 8:56–58.

Abraham, A. and Windmann, S. (2007). Creative cognition: The diverse operations and the prospect of applying a cognitive neuroscience perspective. *Methods*, 42:38–48.

Aliseda, A. (2004). Logics in scientific discovery. *Foundation of Science*, 9:339–363.

Allen, C. (2006). Transitive inference in animals: Reasoning or conditioned associations? In Hurley, S. and Nudds, M., editors, *Rational Animals?*, pages 175–186. Oxford University Press, Oxford.

Allott, N. (2013). Relevance theory. In Capone, A., Piparo, F. L., and Carapezza, M., editors, *Perspectives on Linguistic Pragmatics*, chapter 3. Springer, Berlin.

Anderson, B. L. (2011). The myth of computational level theory and the vacuity of rational analysis. Comment on Jones and Love (2011). *Behavioural and Brain Sciences*, 34:169–231.

Anderson, J. R. (1990). *The Adaptive Character of Thought*. Erlbaum, Hillsdale, NJ.

Anderson, J. R. (1991). Is human cognition adaptive? *Behavioral and Brain Sciences*, 14:471–517.

Ansburg, P. I. (2000). Individual differences in problem solving via insight. *Current Psychology: Developmental, Learning, Personality, Social*, 19(2):143–146.

Ansburg, P. I. and Hill, K. (2003). Creative and analytic thinkers differ in their use of attentional resources. *Personality and Individual Differences*, 34:1141–1152.

Arbib, M. (2012). *How the brain got language: The mirror system hypothesis*. Oxford University Press, Oxford.

Armstrong, D. F. (1993). Comments on Burling (1993). *Current Anthropology*, 34(1):37–38.

Atchley, R. A., Keeney, M., and Burgess, C. (1999). Cerebral hemispheric mechanisms linking ambiguous word meaning retrieval and creativity. *Brain and Cognition*, 40:479–499.

Aumann, R. J. (1987). Correlated equilibrium as an expression of bayesian rationality. *Econometrica*, 55:1–18.

Baayen, R. H. (2008). *Analyzing linguistic data: a practical introduction to statistics using R*. Cambridge University Press, Cambridge, UK.

Baayen, R. H., Davidson, D. J., and Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, 59:390–412.

Barr, D. J., Levy, R., Scheepers, C., and Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, 68:255–278.

Bates, D., Maechler, M., and Bolker, B. (2011). *lme4: Linear mixed-effects models using S4 classes*. R package version 0.999375-42.

Beeman, M. (1993). Semantic processing in the right hemisphere may contribute to drawing inferences from discourse. *Brain and Language*, 44:80–120.

Beeman, M., Friedman, R. B., Grafman, J., Perez, E., Diamond, S., and Lindsay, M. B. (1994). Summation priming and coarse semantic coding in the right hemisphere. *Journal of Cognitive Neuroscience*, 6(1):26–45.

Binmore, K. G. (2009). *Rational decisions*. The Gorman lectures in economics. Princeton University Press, Princeton.

Bird, C. D. and Emmery, N. J. (2009). Rooks use stones to raise the water level to reach a floating worm. *Current Biology*, 19:1410–1414.

Blokpoel, M., Kwisthout, J., Wareham, T., Haselager, P., Toni, I., and van Rooij, I. (2011). The computational costs of recipient design and intention recognition in communication. In *Proceedings of the 33rd Cognitive Science Society*, Austin, TX. Cognitive Science Society.

Boesch, C. and Tomasello, M. (1998). Chimpanzee and human cultures. *Current Anthropology*, 39(5):591–614.

Bonawitz, E. B. and Griffiths, T. L. (2010). Deconfounding hypothesis generation and evaluation in Bayesian models. In *Proceedings of the Thirty-second Cognitive Science Society*, pages 2260–2265, Austin, TX. Cognitive Science Society.

Bookheimer, S. (2002). Functional MRI of language: New approaches to understanding the cortical organization of semantic processing. *Annual Review of Neuroscience*, 25:151–88.

Bottini, G., Corcoran, R., Sterzi, R., Paulesu, E., Schenone, P., Scarpa, P., Frackowiak, R. S., and Frith, C. D. (1994). The role of the right hemisphere in the interpretation of figurative aspects of language. a positron emission tomography activation study. *Brain*, 117(6):1241–53.

Bowden, E. M. and Jung-Beeman, M. (1998). Getting the right idea: Semantic activation in the right hemisphere may help solve insight problems. *Psychological Science*, 9(6):435–440.

Bowden, E. M. and Jung-Beeman, M. (2003a). Aha! insight experience correlates with solution activation in the right hemisphere. *Psychonomic Bulletin and Review*, 10(3):730–737.

Bowden, E. M. and Jung-Beeman, M. (2003b). Normative data for 144 compound remote associate problems. *Behavior Research Methods, Instruments, & Computers*, 35(4):634–639.

Bowden, E. M., Jung-Beeman, M., Fleck, J., and Kounios, J. (2005). New approaches to demystifying insight. *Trends in Cognitive Sciences*, 9(7):322–328.

Bowers, J. S. and Davis, C. J. (2012a). Bayesian just-so stories in psychology and neuroscience. *Psychological Bulletin*, 138(3):389–414.

Bowers, J. S. and Davis, C. J. (2012b). Is that what Bayesians believe? Reply to Griffiths, Chater, Norris, and Pouget (2012). *Psychological Bulletin*, 138(3):423–426.

Brighton, H. and Gigerenzer, G. (2012). *Are rational actor models "rational" outside small worlds?*, chapter 5, pages 84–109. Cambridge University Press, Cambridge.

Burks, A. W. (1946). Peirce's theory of abduction. *Philosophy of Science*, 13(4):301–306.

Burling, R. (1993). Primate calls, human language, and nonverbal communication [and comments and reply]. *Current Anthropology*, 34(1):25–53.

Burling, R. (1999). Motivation, conventionalization, and arbitrariness in the origin of language. In King, B. J., editor, *The Origins of Language: What Nonhuman Primates Can Tell Us*, chapter 9. School of American Research Press.

Bylander, T., Allemang, D., Tanner, M. C., and Josephson, J. R. (1991). The computational complexity of abduction. *Artificial Intelligence*, 49:25–60.

Call, J. (2006). Descartes' two errors: Reason and reflection in the great apes. In Hurley, S. and Nudds, M., editors, *Rational Animals?*, chapter 10, pages 219–234. Oxford University Press, Oxford.

Campos, D. G. (2011). On the distinction between Peirce's abduction and Lipton's inference to the best explanation. *Synthese*, 180(3):419–442.

Carroll, C. D. and Kemp, C. (2013). Hypothesis space checking in intuitive reasoning. In Knauff, M., Pauen, M., Sebanz, N., and Wachsmuth, I., editors, *Proceedings of the 35th Annual Conference of the Cognitive Science Society*, pages 287–292, Austin, TX. Cognitive Science Society.

Catrambone, R. (2002). The effects of surface and structural feature matches on the access of story analogs. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 28(2):318–334.

Chalmers, D. J. (1990). Why Fodor and Pylyshyn were wrong: The simplest refutation. In *Proceedings of the 12th Annual Conference of the Cognitive Science Society*, pages 340–347. Cognitive Science Society, Cambridge, MA.

Chater, N., Goodman, N., Griffiths, T. L., Kemp, C., Oaksford, M., and Tenenbaum, J. B. (2011). The imaginary fundamentalists: The unshocking truth about Bayesian cognitive science. Reply to Jones and Love (2011). *Behavioral and Brain Sciences*, 34(4):194–196.

Chater, N. and Oaksford, M. (2008). The probabilistic mind: prospects for a Bayesian cognitive science. In Chater, N. and Oaksford, M., editors, *The Probabilistic Mind: Prospects for a Bayesian Cognitive Science*, pages 3–32. Oxford University Press.

Chater, N., Tenenbaum, J. B., and Yuille, A. (2006). Probabilistic models of cognition: Conceptual foundations. *Trends in Cognitive Sciences*, 10(7):287–291.

Cheney, D. L. and Seyfarth, R. M. (1990). *How monkeys see the world: Inside the mind of another species*. University of Chicago Press, Chicago, IL.

Cherubini, P., Castelvecchio, E., and Cherubini, A. M. (2005). Generation of hypotheses in Wason's 2-4-6 task: An information theory approach. *The Quarterly Journal of Experimental Psychology*, 58A(2):309–332.

Chiappe, D. L. and Kukla, A. (1996). Context selection and the frame problem. *Behavioral and Brain Sciences*, 19(3):529–530.

Chiarello, C., Burgess, C., and Richards, L. (1990). Semantic and associative priming in the cerebral hemispheres: Some words do, some words don't ... sometimes, some places. *Brain and Language*, 38:75–104.

Christie, S. and Gentner, D. (2010). Where hypotheses come from: Learning new relations by structural alignment. *Journal of Cognition and Development*, 11(3):356–373.

Colhoun, J. and Gentner, D. (2009). Inference processes in causal analogies. In Kokinov, B., Holyoak, K. J., and Gentner, D., editors, *New frontiers in analogy research: Proceedings of the Second International Conference on Analogy*, pages 82–91, Sofia, Bulgaria. New Bulgarian University Press.

Coren, S. (1995). Differences in divergent thinking as a function of handedness and sex. *The American Journal of Psychology*, 108(3):311–325.

Crockford, C., Witting, R. M., Mundry, R., and Zuberbühler, K. (2012). Wild chimpanzees inform ignorant group members of danger. *Current Biology*, 22(2):1–5.

Csibra, G. (2003). Teleological and referential understanding of action in infancy. *Philosophical Transactions of the Royal Society B*, 358:447–458.

Cubitt, R. P. and Sugden, R. (2003). Common knowledge, salience and convention: A reconstruction of David Lewis' game theory. *Economics and Philosophy*, 19:175–210.

Cushen, P. J. and Wiley, J. (2011). Aha! Voila! Eureka! Bilingualism and insightful problem solving. *Learning and Individual Differences*, 21:458–462.

Cuskley, C. and Kirby, S. (2013). Synesthesia, cross-modality, and language evolution. In Simner, J. and Hubbard, E. M., editors, *The Oxford Handbook of Synesthesia*, pages 869–907. Oxford University Press, Oxford.

D'Amato, M. R. and Colombo, M. (1985). Auditory matching to sample in monkeys (*Cebus apella*). *Animal Learning and Behavior*, 52(3):225–236.

Danks, D. (2008). Rational analyses, instrumentalism, and implementations. In Chater, N. and Oaksford, M., editors, *The Probabilistic Mind: Prospects for a Bayesian Cognitive Science*, pages 59–78. Oxford University Press.

Davidson, D. (1975). *Inquiries into Truth and Interpretation*. Oxford University Press, Oxford.

Davidson, D. (1982). Rational animals. *Dialectica*, 36:317–328.

Davies, M. (2008). The Corpus of Contemporary American English: 450 million words, 1990-present. Available online at: `http://corpus.byu.edu/coca/`.

Davies, N. B. and Halliday, T. R. (1978). Deep croaks and fighting assessment in toads *Bufo bufo*. *Nature*, 274:683–685.

Day, S. B. and Gentner, D. (2007). Nonintentional analogical inference in text comprehension. *Memory & Cognition*, 35(1):39–49.

Deacon, T. W. (1997). *The Symbolic Species*. Penguin, London.

Dennett, D. (1978). *Brainstorms*. MIT Press, Cambridge, MA.

Dennett, D. C. (1987). *The Intentional Stance*. MIT Press, Cambridge, MA.

Dennett, D. C. (1994). The role of language in intelligence. In Khalfa, J., editor, *What is Intelligence? The Darwin College Lectures*, pages 161–178. Cambridge University Press, Cambridge.

Deutscher, G. (2002). On the misuse of the notion of 'abuction' in linguistics. *Journal of Linguistics*, 38:469–485.

Dougherty, M. R. P. and Hunter, J. E. (2003). Hypothesis generation, probability judgement, and individual differences in working memory capacity. *Acta Psychologica*, 113:263, 282.

Dretske, F. I. (2006). Minimal rationality. In Hurley, S. and Nudds, M., editors, *Rational Animals?*, pages 107–116. Oxford University Press, Oxford.

Dummet, M. (1993). *Seas of Language*. Oxford University Press, Oxford.

Dunbar, K. (1996). How scientists really reason: Scientific reasoning in real-world laboratories. In Davidson, J. E. and Sternberg, R. J., editors, *The Nature of Insight*, chapter 11, pages 365–395. MIT Press, Cambridge, MA.

Durso, F. T., Rea, C. B., and Dayton, T. (1994). Graph-theoretic confirmation of restructuring during insight. *Psychological Science*, 5(2):94–98.

Dusek, J. A. and Eichenbaum, H. (1997). The hippocampus and memory for orderly stimulus relations. *PNAS*, 94:7109–7114.

Eco, U. (1978). *A Theory of Semiotics*. John Wiley & Sons, New York.

Eco, U. (1984). *Semiotics and the Philosophy of Language*. Indiana University Press, Bloomington, IN.

Eco, U. (1986). *Semiotics and the Philosophy of Language*. Indiana University Press, Bloomington.

Eco, U. (1997). *Kant and the Platypus: Essays on Language and Cognition*. Harcourt, San Diego, CA.

El-Hani, C. N., Queiroz, J., and Stjernfelt, F. (2010). Firefly femmes fatales: A case study in the semiotics of deception. *Biosemiotics*, 3:33–55.

Evans, J. S. B. (2008). Dual-processing accounts of reasoning, judgement and social cognition. *Annual Review of Psychology*, 59:255–278.

Faust, M. and Chiarello, C. (1998). Sentence context and lexical ambiguity resolution by the two hemispheres. *Neuropsychologia*, 36(9):827–835.

Faust, M. and Kahana, A. (2002). Priming summation in the cerebral hemispheres: evidence from semantically convergent and semantically divergent primes. *Neuropsychologia*, 40:892–901.

Faust, M. and Kravetz, S. (1998). Levels of sentence constraint and lexical decision in the two hemispheres. *Brain and Language*, 62:149–162.

Favereau, O. (2008). The unconventional, but conventionalist, legacy of Lewis's "convention". *Topoi*, 27:115–126.

Fay, N., Garrod, S., Roberts, L., and Swoboda, N. (2010). The interactive evolution of human communication systems. *Cognitive Science*, 34:351–386.

Federmeier, K. D. and Kutas, M. (1999). Right words and left words: electrophysiological evidence for hemispheric differences in meaning processing. *Cognitive Brain Research*, 8:373–392.

Firth, R. (1975). *Symbols: public and private*. Cornell University Press, New York.

Fitch, W. T. (2010). *The Evolution of Language*. Cambridge University Press, Cambridge.

Fodor, J. (1987). Modules, frames, frigeons, sleeping dogs and the music of the spheres. In Pylyshyn, Z., editor, *The robot's dilemma: The frame problem in artificial intelligence*. Ablex, Norwood, NJ.

Fodor, J. (2001). *The Mind Doesn't Work That Way: The Scope and Limits of Computational Psychology*. MIT Press, Cambridge, MA.

Fodor, J. and Pylyshyn, Z. (1988). Connectionism and cognitive architecture: A critical analysis. *Cognition*, 28:3–71.

Foerder, P., Galloway, M., Barthel, T., Moore III, D. E., and Reiss, D. (2011). Insightful problem solving in an Asian elephant. *PLOS One*, 6(8):e23251.

Gabbay, D. and Woods, J. (2006). Advice on abductive logic. *Logic Journal of the IGPL*, 14(2):189–219.

Gärdenfors, P. (1995). Cued and deatched representations in animal cognition. *Behavioural Processes*, 35:263–273.

Garrod, S., Fay, N., Lee, J., Oberlander, J., and MacLeod, T. (2007). Foundations of representation: Where might graphical symbol systems come from? *Cognitive Science*, 31(6):961–987.

Gentner, D. (1983). Structure-mapping: A theoretical framework for analogy. *Cognitive Science*, 7:155–170.

Gentner, D. (2010). Bootstrapping the mind: Analogical processes and symbol systems. *Cognitive Science*, 34:752–775.

Gentner, D. and Markman, A. B. (1997). Structure mapping in analogy and similarity. *American Psychologist*, 52(1):45–56.

Gentner, D. and Medina, J. (1998). Similarity and the development of rules. *Cognition*, 65:263–297.

Gergely, G., Bekkering, H., and Király, I. (2002). Rational imitation in preverbal infants. *Nature*, 415:755.

Gergely, G. and Csibra, G. (2003). Teleological reasoning in infancy: the naïve theory of rational action. *Trends in Cognitive Sciences*, 7(7):287–292.

Gergely, G. and Csibra, G. (2006). Sylvia's recipe: The role of imitation and pedagogy in the transmission of cultural knowledge. In Enfield, N. J. and Levenson, S. C., editors, *Roots of Human Sociality: Culture, Cognition and Human Interaction*, pages 229–255. Berg Publishers, Oxford.

Gettys, C. F. and Fisher, S. D. (1979). Hypothesis plausibility and hypothesis generation. *Organizational Behavior and Human Performance*, 24:93–110.

Gigerenzer, G. and Sturm, T. (2012). How (far) can rationality be naturalized? *Synthese*, 187:243–268.

Gilhooly, K. J. and Murphy, P. (2005). Differentiating insight from non-insight problems. *Thinking & Reasoning*, 11(3):279–302.

Gillan, D. J. (1991). Reasoning in the chimpanzee: II. transitive inference. *Journal of Experimental Psychology: Animal Behavior Processes*, 7(2):150–164.

Giora, R. (1997). Understanding figurative and literal language: The graded salience hypothesis. *Cognitive Linguistics*, 7:183–206.

Glass, A., Holyoak, K., and Santa, J. (1979). *Cognition*. Addison-Wesley, Reading, MA.

Glymour, C. (1981). *Theory and Evidence*. University of Chicago Press, Chicago.

Goel, V. and Dolan, R. J. (2000). Anatomical segregation of component processes in an inductive inference task. *Journal of Cognitive Neuroscience*, 12(1):110–119.

Goel, V. and Dolan, R. J. (2004). Differential involvement of left profrontal cortex in inductive and deductive reasoning. *Cognition*, 93:B109–B121.

Goel, V., Gold, B., Kapur, S., and Houle, S. (1997). The seats of reason? An imaging study of deductive and inductive reasoning. *Neuroreport*, 8:1305–1310.

Graham, S. A. and Kilbreath, C. S. (2007). It's a sign of the kind: Gestures and words guide infants' inductive inferences. *Developmental Psychology*, 43(5):1111–1123.

Greene, A. J., Gross, W. L., Elsinger, C. L., and Rao, S. M. (2006). An fMRI analysis of the human hippocampus: inference, context and task awareness. *Journal of Cognitive Neuroscience*, 18(7):1156–1173.

Griffiths, T. L., Chater, N., Kemp, C., Perfors, A., and Tenenbaum, J. B. (2010). Probabilistic models of cognition: exploring representations and inductive biases. *Trends in Cognitive Sciences*, 14:357–364.

Griffiths, T. L., Chater, N., Norris, D., and Pouget, A. (2012). How the Bayesians got their beliefs (and what those beliefs actually are): Comment on Bowers and Davis (2012). *Psychological Bulletin*, 138(3):415–422.

Griffiths, T. L., Kemp, C., and Tenenbaum, J. B. (2008). Bayesian models of cognition. In Sun, R., editor, *The Cambridge Handbook of Computational Psychology*, pages 59–100. Cambridge University Press, Cambridge.

Griffiths, T. L. and Tenenbaum, J. B. (2009). Theory-based causal induction. *Psychological Review*, 116(4):661–716.

Grose-Fifer, J. and Deacon, D. (2004). Priming by natural category membership in the left and right cerebral hemispheres. *Neuropsychologia*, 42:1948–1960.

Grüter, C. and Farina, W. M. (2009). The honeybee waggle dance: can we follow the steps? *Trends in Ecology and Evolution*, 24(5):242–247.

Gyger, M., Marler, P., and Pickert, R. (1987). Semantics of an avian alarm call system: the male domestic fowl, Gallus domesticus. *Behaviour*, 102:15–40.

Hanus, D., Mendes, N., Tennie, C., and Call, J. (2011). Comparing the performances of apes (*Gorilla gorilla, Pan troglodytes, Pongo pygmaeus*) and human children (*Homo sapiens*) in the floating peanut task. *PLoS One*, 6(6):e19555.

Hare, B. and Tomasello, M. (2004). Chimpanzees are more skillful in competitive than in cooperative cognitive tasks. *Animal Behaviour*, 68:571–581.

Harnad, S. (1990). The symbol grounding problem. *Physica D*, 42:335–346.

Hayes, B. K. and Newell, B. R. (2009). Induction with uncertain categories: When do people consider the category alternatives? *Memory & Cognition*, 37(6):730–743.

Hélie, S. and Sun, R. (2010). Incubation, insight and creative problem solving: A unified theory and a connectionist model. *Psychological Review*, 117(3):994–1024.

Hockett, C. F. (1960). The origin of speech. *Scientific American*, 203(3):89–96.

Holland, J. H., Holyoak, K. J., Nisbett, R. E., and Thagard, P. (1986). *Induction: Processes of inference, learning, and discovery*. MIT Press, Cambridge, MA.

Holyoak, K. J. and Cheng, P. W. (2011). Causal learning and inference as a rational process: The new synthesis. *Annual Review of Psychology*, 62:135–163.

Holyoak, K. J. and Thagard, P. (1995). *Mental Leaps: Analogy in creative thought*. MIT Press, Cambridge, MA.

Horner, V. and Whiten, A. (2005). Causal knowledge and imitation/emulation switching in chimpanzees (*Pan troglodytes*) and children (*Homo sapiens*). *Animal Cognition*, 8:164–181.

Horner, V., Whiten, A., Flynn, E., and de Waal, F. B. M. (2006). Faithful replication of foraging techniques along cultural transmission chains by chimpanzees and children. *PNAS*, 103(37):13878–13383.

Hurford, J. (2007). *The Origins of Meaning*. Oxford University Press, Oxford.

Hurford, J. R. (2004). Human uniqueness, learned symbols and recursive thought. *European Review*, 12(4):551–565.

Hurford, J. R. (2010). *The Origins of Grammar*. Oxford University Press, Oxford.

Jackendoff, R. (1999). Possible stages in the evolution of the language capacity. *Trends in Cognitive Sciences*, 3(7):272–279.

Jausovec, N. and Bakracevic, K. (1995). What can heart rate tell us about the creative process? *Creativity Research Journal*, 8(1):11–24.

Jern, A. and Kemp, C. (2013). A probabilistic account of exemplar and category generation. *Cognitive Psychology*, 66:85–125.

Johnson, T. R. and Krems, J. F. (2001). Use of current explanations in multicausal abductive reasoning. *Cognitive Science*, 25:903–939.

Jones, M. and Love, B. C. (2011). Bayesian fundamentalism or enlightenment? on the explanatory status and theoretical contributions of Bayesian models of cognition. *Behavioural and Brain Sciences*, 34:169–231.

Joseph, B. D. (1987). On the use of iconic elements in etymological investigation: some case studies from Greek. *Diachronica*, 4(1/2):1–26.

Josephson, J. (2000). Smart inductive generalizations are abductions. In Flach, P. and Kakas, A., editors, *Abduction and Induction*, pages 31–44. Kluwer Academic.

Jung-Beeman, M. (2005). Bilateral brain processes for comprehending natural language. *Trends in Cognitive Sciences*, 9:512–518.

Jung-Beeman, M., Bowden, E. M., Haberman, J., Frymiare, J. L., Arambel-Liu, S., Greenblatt, R., Reber, P. J., and Kounios, J. (2004). Neural activity when people solve verbal problems with insight. *PLOS Biology*, 2(4):500–510.

Jung Von Matt (2012). Lego "imagine". http://www.jvm.com/en/work/work_subpages/lego_imagine.html. Last checked 20 February 2014.

Kacelnik, A. (2006). Meanings of rationality. In Hurley, S. and Nudds, M., editors, *Rational Animals?*, pages 87–106. Oxford University Press, Oxford.

Kantartzis, K., Imai, M., and Kita, S. (2011). Japanese sound-symbolism facilitates word learning in English-speaking children. *Cognitive Science*, 35:575–586.

Kapitan, T. (1992). Peirce and the autonomy of abductive reasoning. *Erkenntnis*, 37(1):1–26.

Kaplan, C. A. and Simon, H. A. (1990). In search of insight. *Cognitive Psychology*, 22:374–419.

Kaplan, J. A., Brownell, H. H., Jacobs, J. R., and Gardner, H. (1990). The effects of right hemisphere damage on the pragmatic interpretation of conversational remarks. *Brain and Language*, 38:315–333.

Keller, R. (1998). *A Theory of Linguistic Signs*. Oxford University Press, Oxford.

Kemp, C. and Jern, A. (2014). A taxonomy of inductive problems. *Psychonomic Bulletin and Review*, 21(1):23–46.

Kemp, C., Perfors, A., and Tenenbaum, J. B. (2004). Learning domain structures. In *Proceedings of the 26th Annual Conference of the Cognitive Science Society*, pages 672–677, Austin, TX. Cognitive Science Society.

Kemp, C., Perfors, A., and Tenenbaum, J. B. (2007). Learning overhypotheses with hierarchical Bayesian models. *Developmental Science*, 10(3):307–321.

Kiefer, M., Weisbrod, M., Kern, I., Maier, S., and Spitzer, M. (1998). Right hemisphere activation during indirect semantic priming: Evidence from event-related potentials. *Brain & language*, 64:377–408.

Kirby, S. (2002). Natural language from artificial life. *Artificial Life*, 8(185-215).

Kirby, S., Cornish, H., and Smith, K. (2008). Cumulative cultural evolution in the laboratory: An experimental approach to the origins of structure in human language. *PNAS*, 105(31):10681–10686.

Kircher, T. T. J., Brammer, M., Andreu, N. T., Williams, S. C., and McGuire, P. K. (2001). Engagement of right temporal cortex during processing of linguistic context. *Neuropsychologia*, 39:798–809.

Kiss, G. R., Armstrong, C., Milroy, R., and Piper, J. (1973). An associative thesaurus of English and its computer analysis. Available online at http://eat.rl.ac.uk.

Kita, S., Kantartzis, K., and Imai, M. (2010). Children learn sound symbolic words better: evolutionary vestige of sound symbolic protolanguage. In Smith, A. D. M., Schouwstra, M., de Boer, B., and Smith, K., editors, *The Evolution of Language: Proceedings of the 8th International Conference (EVOLANG)*, pages 206–213, New Jersey. World Scientific.

Köhler, W. (1927). *The Mentality of Apes*. Harcourt, Brace, New York.

Kounios, J. and Beeman, M. (2009). The *Aha!* moment. *Current Diretions in Psychological Science*, 18(4):210–216.

Kounios, J., Frymiare, J. L., Bowden, E. M., Fleck, J. I., Subramaniam, K., Parrish, T. B., and Jung-Beeman, M. (2006). The prepared mind: Neural activity prior to problem presentation predicts subsequent solution by sudden insight. *Psychological Science*, 17(10):882–890.

Kurtz, K. J., Gentner, D., and Gunn, V. (1999). Reasoning. In Rumelhart, D. E. and Bly, B. M., editors, *Cognitive Science: Handbook of perception and cognition*, pages 145–200. Academic Press, San Diego.

Kwisthout, J. (2012). Relevancy in problem solving: A computational framework. *The Journal of Problem Solving*, 5(1):18–33.

Kwisthout, J., Wareham, T., and van Rooij, I. (2011). Bayesian intractability is not an ailment that approximation can cure. *Cognitive Science*, 35:779–784.

Latsis, J. S. (2005). Is there redemption for conventions? *Cambridge Journal of Economics*, 29:709–727.

Leavens, D. and Hopkins, W. (2005). Multimodal concomitants of manual gesture by chimpanzees (*Pan troglodytes*). *Gesture*, 5:73–88.

Lee, H. S. and Holyoak, K. J. (2008). The role of causal models in analogical inference. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 34:1111–1122.

Lewis, D. (1969). *Convention: A philosophical study*. Harvard University Press, Cambridge, MA.

Lewis, D. (1983). *Philosophical papers*, volume 1. Oxford University Press, New York.

Lipton, P. (2004). *Inference to the Best Explanation*. Routledge, New York, 2nd edition.

Lumsden, D. (2002). Crossing the symbolic threshold. *Philosophical Psychology*, 15(2).

Luo, J. and Niki, K. (2003). Function of hippocampus in "insight" of problem solving. *Hippocampus*, 13:316–323.

Lyons, D. E., Young, A. G., and Keil, F. C. (2007). The hidden structure of overimitation. *Proceedings of the National Academy of Sciences*, 104:19751–19756.

Lyons, J. (1977). *Semantics*, volume 1. Cambridge University Press, Cambridge.

MacGregor, J. N. and Cunningham, J. B. (2008). Rebus puzzles as insight problems. *Behavior Research Methods*, 40(1):263–268.

Markman, E. M. (1989). *Categorization and Naming in Children: Problems of Induction*. MIT Press, Cambridge, MA.

Markman, E. M. (1991). The whole-object, taxonomic, and mutual exclusivity assumptions as initial constraints on word meanings. In Gelman, S. A. and Byrnes, J. P., editors, *Perspectives on Language and Thought: Interrelations in Development*, pages 72–106. Cambridge University Press, New York.

Marr, D. (1982). *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. Henry Holt and Co., New York.

Mashal, N., Faust, M., and Hendler, T. (2005). The role of the right hemisphere in processing nonsalient metaphorical meanings: Application of principal components analysis to fMRI data. *Neuropsychologia*, 43:2084–2100.

Mashal, N., Faust, M., Hendler, T., and Jung-Beeman, M. (2007). An fMRI investigation of the neural correlates underlying the processing of novel metaphoric expressions. *Brain and Language*, 100:115–126.

Mason, W. and Suri, S. (2011). Conduction behavioral research on Amazon's Mechanical Turk. *Behav Res*.

McCarthy, J. and Hayes, P. J. (1969). Some philosophical problems from the standpoint of artificial intelligence. *Machine Intelligence*, 4:463–502.

McGilchrist, I. (2010). Reciprocal organization of the cerebral hemispheres. *Dialogues in Clinical Neuroscience*, 12(4):503–515.

McGonigle, B. O. and Chalmers, M. (1977). Are monkeys logical? *Nature*, 267:694–696.

Medina, T. N., Snedeker, J., Trueswell, J. C., and Gleitman, L. R. (2011). How words can and cannot be learned by observation. *PNAS*, 108(22):9014–9019.

Mednick, S. A. (1962). The associative basis of the creative process. *Psychological Review*, 69(3):220–232.

Mendes, N., Hanus, D., and Call, J. (2007). Raising the level: orangutans use water as a tool. *Biology Letters*, 3:453–455.

Menenti, L., Petersson, K. M., Scheeringa, R., and Hagoort, P. (2009). When elephants fly: Differential sensitivity of right and left inferior frontal gyri to discourse and world knowledge. *Journal of Cognitive Neuroscience*, 21(12):2358–2368.

Mercier, H. and Sperber, D. (2009). Intuitive and reflective inferences. In Evans, J. and Frankish, K., editors, *In Two Minds: Dual-processes and beyond*, chapter 7. Oxford University Press, Oxford.

Mill, J. S. (1882/1843). *A System of Logic, Ratiocinative and Inductive*. Harber & Brothers, New York, 8th edition.

Millikan, R. G. (2005). *Language: A Biological Model*. Oxford University Press, Oxford.

Millikan, R. G. (2006). Styles of rationality. In Hurley, S. and Nudds, M., editors, *Rational Animals?*, pages 117–126. Oxford University Press, Oxford.

Moll, H. and Tomasello, M. (2007). Cooperation and human cognition: the Vygotskian intelligence hypothesis. *Philosophical Transactions of the Royal Society B*, 362:639–648.

Morewedge, C. K. and Kahneman, D. (2010). Associative processes in intuitive judgment. *Trends in Cognitive Sciences*, 14(10):435–440.

Murray, M. A. and Byrne, R. M. (2005). Attention and working memory in insight problem solving. In *Proceedings of the Cognitive Science Society*, volume 27, pages 1571–1575, Austin, TX. Cognitive Science Society.

Myrstad, J. A. (2004). The use of converse abduction in Kepler. *Foundation of Science*, 9(321-338).

Navarro, D. J. and Perfors, A. F. (2011). Hypothesis generation, sparse categories, and the positive test strategy. *Psychological Review*, 118(1):120–134.

Nelson, D. L., McEvoy, C. L., and Schreiber, T. A. (1998). The University of South Florida word association, rhyme, and word fragment norms. http://www.usf.edu/FreeAssociation. Last checked 20 February 2014.

Noble, W. and Davidson, I. (1996). *Human Evolution, Language and Mind*. Cambridge University Press, Cambridge.

Noth, W. (1995). *Handbook of Semiotics*. John Wiley & Sons.

O'Doherty, J. P. (2004). Reward representations and reward-related learning in the human brain: insights from neuroimaging. *Current Opinion in Neurobiology*, 14:769–776.

Okasha, S. (2000). Van Fraassen's critique of inference to the best explanation. *Studies in the history and philosophy of science*, 31(4):691–710.

Paavola, S. (2004). Abduction as a logic and methodology of discovery: the importance of strategies. *Foundation of Science*, 9:267–283.

Paavola, S. (2005). Peircean abduction: Instinct or inference? *Semiotica*, 153(1/4):131–154.

Paavola, S. (2006). Hansonian and Harmanian abduction as models of discovery. *International Studies in the Philosophy of Science*, 20(1):93–108.

Peirce, C. S. (1931-1935). *Collected Papers of Charles Sanders Peirce*, volume 1-6. Harvard University Press, Cambridge, MA.

Peirce, C. S. (1955). *Philosophical Writings of Peirce*. Dover Publications, New York.

Penn, D. C., Holyoak, K. J., and Povinelli, D. J. (2008). Darwin's mistake: Explaining the discontinuity between human and nonhuman minds. *Behavioral and Brain Sciences*, 31:109–178.

Pepperberg, I. M. (2000). *The Alex Studies: Cognitive and Communicative Abilities of Grey Parrots*. Harvard University Press, Cambridge, MA.

Perniss, P., Thompson, R. L., and Vigliocco, G. (2010). Iconicity as a general property of language: Evidence from spoken and signed languages. *Frontiers in Psychology*, 1:227.

Pet Food Manufacturers Association (2013). Pet population 2013. http://www.pfma.org.uk/pet-population/. Last checked 20 February 2014.

Pinker, S. (1994). *The Language Instinct*. Penguin, London.

Pinker, S. (1997). *How the Mind Works*. W. W. Norton & Co., New York.

Plutynski, A. (2011). A brief history of abduction. *The Journal of the International Society for the History of Philosophy of Science*, 1(2):227–248.

Popper, K. (1934/1968). *The Logic of Scientific Discovery*. Harper & Row, New York.

Popper, K. (1972). *Objective Knowledge: An Evolutionary Approach*. Clarendon Press, Oxford.

Postema, G. J. (2008). Salience reasoning. *Topoi*, 27:41–55.

Premack, D. and Premack, A. J. (1983). *The Mind of an Ape*. Erlbaum, Hillsdale, NJ.

Psillos, S. (2009). An explorer upon untrodden ground: Peirce on abduction. In Gabbay, D. M., Hartmann, S., and Woods, J., editors, *Handbook of the History of Logic*, volume 10: Inductive Logic, pages 117–151. Elsevier.

Pylyshyn, Z. (1984). *Computation and Cognition*. MIT Press, Cambridge, MA.

Quillian, M. R. (1968). Semantic memory. In Minksy, M., editor, *Semantic Information Processing*, pages 227–270. MIT Press, Cambridge, MA.

Quine, W. V. (1992). Structure and nature. *The Journal of Philosophy*, 89(1):5–9.

Quine, W. V. O. (1960). *Word and Object*. MIT Press, Cambridge, MA.

R Development Core Team (2011). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Ramachandran, V. S. and Spence, C. (2001). Synaesthesia: a window into perception, thought and language. *Journal of Consciousness Studies*, 8:3–34.

Ramscar, M., Yarlett, D., Dye, M., Denny, K., and Thorpe, K. (2010). The effects of feature-label-order and their implications for symbolic learning. *Cognitive Science*, 34:909–957.

Reichenbach, H. (1938). *Experience and Prediction*. University of Chicago Press, Chicago, IL.

Rendall, D., Owren, M. J., and Ryan, M. J. (2009). What do animal signals mean? *Animal Behaviour*, 78:233–240.

Ribeiro, S., Loula, A., de Araújo, I., Gudwin, R., and Queiroz, J. (2006). Symbols are not uniquely human. *Biosystems*, 90:263–272.

Richards, I. A. (1936/1950). *The Philosophy of Rhetoric*. Oxford University Press, London.

Richardson, K. (1991). Reasoning with raven — in and out of context. *British Journal of Educational Psychology*, 61:129–138.

Roberts, M. J., Welfare, H., Livermore, D. P., and Theadom, A. M. (2000). Context, visual salience, and inductive reasoning. *Thinking & Reasoning*, 6(4):349–374.

Rogers, L. J., Zucca, P., and Vallortigara, G. (2004). Advantages of having a lateralized brain. *Proceedings of the Royal Society of London B*, 271(S420-424).

Rosch, E., Mervis, C. B., Gray, W., Johnson, D., and Boyes-Braem, P. (1976). Basic objects in natural categories. *Cognitive Psychology*, 8:382–439.

Rosenberg, J. F. (1974). *Linguistic Representation*. D. Reidel Publishing Company, Dordrecht.

Ross, B. H. and Bradshaw, G. L. (1994). Encoding effects of remindings. *Memory & Cognition*, 22(5):591–605.

Saussure, F. (1959). *Course in General Linguistics*. Philosophical Library, New York, 3 edition.

Savage-Rumbaugh, E. S. and Rumbaugh, D. M. (1978). Symbolization, language and chimpanzees: a theoretical reevaluation based on initial language acquisition processes in four young *Pan troglodytes*. *Brain & language*, 6:265–300.

Savage-Rumbaugh, E. S., Rumbaugh, D. M., and Boysen, S. (1978). Symbolic communication between two chimpanzees (*Pan troglodytes*). *Science*, 201(4356):641–644.

Savage-Rumbaugh, E. S., Rumbaugh, D. M., Smith, S. T., and Lawson, J. (1980). Reference: The linguistic essential. *Science*, 210(4472):922–925.

Schelling, T. (1960). *The Strategy of Conflict*. Harvard University Press, Cambridge, MA.

Schooler, J. W. and Melcher, J. (1995). The ineffability of insight. In Smith, S. M., Ward, T. B., and Finke, R. A., editors, *The Creative Cognition Approach*, chapter 5, pages 97–133. MIT Press.

Schooler, J. W., Ohlsson, S., and Brooks, K. (1993). Thoughts beyond words: When language overshadows insight. *Journal of Experimental Psychology: General*, 122(2):166–183.

Schurz, G. (2008). Patterns of abduction. *Synthese*, 164:201–234.

Scott-Phillips, T. C., Kirby, S., and Ritchie, G. R. (2009). Signalling signalhood and the emergence of communication. *Cognition*, 113:226–233.

Sebeok, T. A. (1994/2001). *Signs: An Introduction to Semiotics*. University of Toronto Press, Toronto.

Seyfarth, R. M., Cheney, D. L., and Marler, P. (1980). Vervet monkey alarm calls: Semantic communication in a free-ranging primate. *Animal Behaviour*, 28:1070–1094.

Shafto, P., Kemp, C., Baraff, E., Coley, J. D., and Tenenbaum, J. B. (2005). Context-sensitive induction. In *Proceedings of the 27th annual conference of the cognitive science society*, pages 2003–2008, Austin, TX. Cognitive Science Society.

Shafto, P., Kemp, C., Bonawitz, E. B., Coley, J. D., and Tenenbaum, J. B. (2008). Inductive reasoning about causally transmitted properties. *Cognition*, 109:175–192.

Shanahan, M. (2009). The frame problem. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*. Winter 2009 edition.

Short, T. L. (2004). The development of Peirce's theory of signs. In Misak, C., editor, *The Cambridge Companion to Peirce*, chapter 9, pages 214–240. Cambridge University Press.

Sillari, G. (2008). Common knowledge and convention. *Topoi*, 27:29–39.

Simner, J., Cuskley, C., and Kirby, S. (2010). What sound does that taste? Cross-modal mappings across gustation and audition. *Perception*, 39:553–569.

Skyrms, B. (1996). *Evolution of the social contract*. Cambridge University Press, Cambridge.

Smith, K. (2004). The evolution of vocabulary. *Journal of Theoretical Biology*, 228:127–142.

Smith, K., Smith, A. D. M., and Blythe, R. A. (2011). Cross-situational learning: An experimental study of word-learning mechanisms. *Cognitive Science*, 35:480–498.

Sonesson, G. (2006). The meaning of meaning in biology and cognitive science: A semiotic reconstruction. *Sign Systems Studies*, 34(1):135–213.

Sperber, D. and Wilson, D. (1986/1995). *Relevance: Communication and Cognition*. Blackwell Publishing, Malden, MA, 2nd edition.

Sperber, D. and Wilson, D. (1996). Fodor's frame problem and relevance theory: reply to Chiappe & Kukla. *Behavioral and Brain Sciences*, 19(3):530–532.

Sperber, D. and Wilson, D. (2002). Pragmatics, modularity and mind-reading. *Mind & Language*, 17(1-2):3–23.

Sprenger, A. M., Dougherty, M. R., Atkins, S. M., Franco-Watkins, A. M., Thomas, R. P., Lange, N., and Abbs, B. (2011). Implications of cognitive load for hypothesis generation and probability judgement. *Frontiers in Psychology*, 2.

Taylor, A. H. and Gray, R. D. (2009). Animal cognition: Aesop's fable flies from fiction to fact. *Current Biology*, 19(17):R731–R732.

Tenenbaum, J. B., Griffiths, T. L., and Kemp, C. (2006). Theory-based Bayesian models of inductive learning and reasoning. *Trends in Cognitive Sciences*, 10(7):309–318.

Thagard, P. (2007). Abductive inference: from philosophical analysis to neural mechanisms. In Feeney, A. and Heit, E., editors, *Inductive reasoning: Cognitive, mathematical, and neuroscientific approaches*, pages 226–247. Cambridge University Press.

Thagard, P. and Stewart, T. C. (2011). The AHA! experience: Creativity through emergent binding in neural networks. *Cognitive Science*, 35:1–33.

Thomas, R. P., Dougherty, M. R., Sprenger, A. M., and Harbison, J. I. (2008). Diagnostic hypothesis generation and human judgement. *Psychological Review*, 115(1):155–185.

Thompson, R. L., Vinson, D. P., and Vigliocco, G. (2009). The link between form and meaning in American Sign Language: Lexical processing effects. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 35(2):550–557.

Thompson, R. L., Vinson, D. P., Woll, B., and Vigliocco, G. (2012). The road to language learning is iconic: Evidence from British Sign Language. *Psychological Science*, 23(12):1443–1448.

Tomasello, M. (1990). Cultural transmission in the tool use and communicatory signalling of chimpanzees? In Parker, S. T. and Gibson, K. R., editors, *"Language" and intelligence in monkeys and apes: comparative developmental perspectives*, chapter 10, pages 274–311. Cambridge University Press.

Tomasello, M. (1999). *The Cultural Origins of Human Cognition*. Harvard University Press, Cambridge, MA.

Tversky, A. and Kahneman, D. (1974). Judgement under uncertainty: Heuristics and biases. *Science*, 185:1124–1131.

Tversky, A. and Kahneman, D. (1983). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgement. *Psychological Review*, 90(4):293–315.

Vanderschraaf, P. (1995). Convention as correlated equilibrium. *Erkenntnis*, 42(1):65–87.

Vasconcelos, M. (2008). Transitive inference in non-human animals: An empirical and theoretical analysis. *Behavioural Processes*, 78:313–334.

Virtue, S., van den Broek, P., and Linderholm, T. (2006). Hemispheric processing of inferences: The effects of textual constraint and working memory capacity. *Memory & Cognition*, 34(6):1341–1354.

von Fersen, L., Wynne, C. D. L., Delius, J. D., and Staddon, J. E. R. (1991). Transitive inference formation in pigeons. *Journal of Experimental Psychology: Animal Behavior Processes*, 17(3):334–341.

von Frisch, K. (1967/1994). *The Dance Language and Orientation of Bees*. Harvard University Press, Cambridge, MA.

Wagner, U., Gais, S., Haider, H., Verleger, R., and Born, J. (2004). Sleep inspires insight. *Nature*, 427:352–355.

Washburn, D., , Thompson, R., and Oden, D. (1997). Monkeys trained with same/different symbols do not match relations. Paper presented at the meeting of the Psychonomic society, Philadelphia, PA.

Wason, P. C. (1968). Reasoning about a rule. *The Quarterly Journal of Experimental Psychology*, 20(3):273–281.

Weisberg, J. (2009). Locating IBE in the Bayesian framework. *Synthese*, 167:125–143.

Whiten, A., Goodall, J., McGrew, W. C., Nishida, T., Reynolds, V., Sugiyama, Y., Tutin, C. E. G., Wrangham, R. W., and Boesch, C. (1999). Cultures in chimpanzees. *Nature*, 399:682–685.

Whiten, A., McGuigan, N., Marshall-Pescini, S., and Hopper, L. M. (2009). Emulation, imitation, over-imitation and the scope of culture for child and chimpanzee. *Philosophical Transactions of the Royal Society B*, 364:2417–2428.

Wilson, E. O. (1975). *Sociobiology: The New Synthesis*. Harvard University Press, Cambridge, MA.

Xu, F. and Tenenbaum, J. B. (2007). Word learning as Bayesian inference. *Psychological Review*, 114(2):245–272.

Xu, J., Kemeny, S., Park, G., Frattali, C., and Braun, A. (2005). Language in context: emergent features of word, sentence, and narrative comprehension. *NeuroImage*, 25:1002–1015.

Xu, Y. and Wang, P. (2012). The frame problem, the relevance problem, and a package solution to both. *Synthese*, 187:43–72.

Zentall, T. R. and Sherburne, L. M. (1994). Transfer of value from s+ to s– in a simultaneous discrimination. *Journal of Experimental Psychology: Animal Behavior Processes*, 20:176–183.

Zlatev, J. (2008). From proto-mimesis to language: Evidence from primatology and social neuroscience. *Journal of Physiology*, 102:137–151.

Zlatev, J. (2009). The semiotic hierarchy: Life, consciousness, signs and language. *Cognitive Semiotics*, 2009(4):169–200.

Zuberbühler, K., Cheney, D. L., and Seyfarth, R. M. (1999). Conceptual semantics in a nonhuman primate. *Journal of Comparative Psychology*, 113(1):33–42.