WHAT LIES BEHIND THE DATA?
HOW SAMPLING ASSUMPTIONS SHAPE AND ARE SHAPED BY INDUCTIVE INFERENCE

KEITH RANSOM

School of Psychology
Faculty of Health Sciences
The University of Adelaide

August 2019

CONTENTS

WHEN THE BASIS FOR INDUCTION IS UNCLEAR

WHEN THE BASIS FOR INDUCTION IS UNCLEAR

WHEN THE BASIS FOR SAMPLING IS UNCLEAR

ABSTRACT

The problems of everyday cognition, from perception to social interaction and higher level reasoning, require us to predict future events and outcomes on the basis of past experience. But often (if not always) solutions to the problems we face are under-determined by our experience. So we reason inductively, drawing uncertain conclusions from incomplete information. Yet, despite our lack of first hand data, our reasoning is efficient and effective nonetheless. So how do we close the gap between the paucity of experience and the effectiveness of reason? One way that we do this is by exploiting statistical regularities that we have observed in the world, assuming (contra philosophers' counsel) that these regularities will continue to hold. In so doing, we leverage the evidentiary value of the data that we do have.

This thesis examines our assumptions about what lies beneath the data and how we leverage them to reason beyond it. In particular, it focuses on our mental models of the world – generative models that connect observations to hypotheses through their consequences. I consider the assumptions we make in solving three separate reasoning problems of increasing complexity. Firstly, in a series of related experiments I explore the effect of sampling assumptions in a categorisation task based on low-dimensional perceptual stimuli. Together, these experiments examine how reasoners weigh the value of extra data when deciding how far to generalise, and the extent to which the computations involved are influenced by their representational and sampling assumptions. In addition, I use the same experimental framework to investigate a related question: if people's sampling assumptions do alter the weighing of evidence, at what stage do these effects manifest – during learning, or only at the point of generalisation? Secondly, I examine the role of sampling assumptions in the shift from percept to concept. A key challenge for the reasoner when reasoning from high-dimensional categorical stimuli is in deciding which of the many dimensions or features represent the appropriate basis for induction. I investigate how the perceived relevance of particular features in the data is affected by people's assumptions about the representativeness of the sampling process.

In almost every sphere of human activity, we reason from data generated by others and we generate data from which others will reason. Equipped with a theory of mind, both senders and receivers of data may exploit recursive "I think, you think, I think..." reasoning to increase the evidentiary weight of data, and improve the utility of communication as a result. But when data is highly leveraged in this way, there is a downside risk. If reciprocal assumptions are not well calibrated, the reasoner may leap to the *wrong* conclusion. In the final study, I investigate the phenomena of recursive *meta-inference* in a setting where deception is warranted but lying is not an option – a setting which offers

particular advantages. Firstly, when perpetrating or avoiding a deception, some degree of meta-inferential assumption becomes a vital pre-requisite. Secondly, placing the goals of communicating parties at odds offers the potential to more easily distinguish whether people engage in genuine reflection about the assumptions of another or merely respond to constraints implicit in the sampling process.

The studies described in this thesis deal with progressively more complex challenges that we face as reasoners: how far should we generalise when the basis of induction is clear, how do we determine the relevant basis for induction in the first place, and how do we calibrate our own inductive inference with that of another. Through a combination of computational modelling and human behavioural experiments I demonstrate how our sampling assumptions influence the way we meet these challenges, and how our solution to each challenge may be inter-related.

DECLARATION

I certify that this work contains no material which has been accepted for the award of any other degree or diploma in my name in any university or other tertiary institution and, to the best of my knowledge and belief, contains no material previously published or written by another person, except where due reference has been made in the text. In addition, I certify that no part of this work will, in the future, be used in a submission in my name for any other degree or diploma in any university or other tertiary institution without the prior approval of the University of Adelaide and where applicable, any partner institution responsible for the joint award of this degree.

The author acknowledges that copyright of published works contained within this thesis resides with the copyright holder(s) of those works. I give permission for the digital version of my thesis to be made available on the web, via the University's digital research repository, the Library Search and also through web search engines, unless permission has been granted by the University to restrict access for a period of time.

I acknowledge the support I have received for my research through the provision of an Australian Government Research Training Program Scholarship.

Adelaide, August 2019

Keith Ransom

PUBLICATIONS

The following publications are wholly reproduced in this thesis:

Chapter 2: **K Ransom**, A Hendrickson, A Perfors, and DJ Navarro (2018). Representational and sampling assumptions drive individual differences in single category generalisation. In C Kalish, M Rau, J Zhu and T Rogers (Ed.) *Proceedings of the 40th Annual Conference of the Cognitive Science Society.*

Chapter 4: **K Ransom** and A Perfors (2019) Exploring the role that encoding and retrieval play in sampling effects. In A Goel, C Seifert, and C Freksa (Eds.) *Proceedings of the 41st Annual Conference of the Cognitive Science Society.*

Chapter 5: **K Ransom**, A Perfors and DJ Navarro (2016). Leaping to conclusions: Why premise relevance affects argument strength. *Cognitive Science*, *40(7)*, 1775-1796

The following publication represents a preliminary publication of work in this thesis:

Chapter 6: **K Ransom**, W Voorspoels, A Perfors, and DJ Navarro (2017). A cognitive analysis of deception without lying. In G Gunzelmann, A Howes, T Tenbrink, and E Davelaar (Ed.) *Proceedings of the 39th Annual Conference of the Cognitive Science Society*: 992-997

My own contribution to the following publication is largely reproduced in this thesis:

Chapter 3: A Hendrickson, A Perfors, DJ Navarro, and **K Ransom** (2019) Sample size, number of categories and sampling assumptions: Exploring some differences between categorization and generalization. *Cognitive Psychology* 111: 80-102

ACKNOWLEDGEMENTS

I do not acknowledge yellow shoes.
Chrysanthemums played no part in my research.
Signs of support from Mme de Forcheville were absent, but not noticeably so.

Wait a minute...if I go on weakly sampling, these acknowledgments could take a while.

Susan and Sophie, to you both I owe the greatest debt of gratitude. Without your love and support I would never have made it this far.

Amy and Danielle, thank you for letting me join your lab, for creating a stimulating research environment, and for sharing your experience and insight. Certainly none of the work in this thesis would have gotten off the ground, much less be published, without your guidance. Amy, thank you also for your advice, encouragement and patience over the past months while I have been assembling this thesis.

Wouter, thanks for being an unofficial PhD advisor, and for all your warmth, wisdom and wit. Its been a pleasure to work with you. And thank you for hosting my lab visit in Leuven, it was one of the highlights of the whole PhD adventure.

Drew, I have much enjoyed our collaboration. Thanks for your help and advice, and for the benefit of your considerable insight into the literature.

Carolyn, thanks for all your support and for adopting a stray. Likewise, thanks Anna for giving me a lab space to hide out in.

Thanks to Lauren, Steve, Wai Keen, Simon, Jess, Adam and all the other lab mates and fellow PhD students who created a great environment to come to work in.

Thanks to my family and friends for being there, and for knowing when not to ask how the write-up was coming along.

To all the lovely people at AGF+W, thanks for fuelling my caffeine addiction with such friendly spirit while I've been wrangling my thesis.

1 WHAT LIES BEHIND THE DATA

> “ *We ought then to regard the present state of the universe as the effect of its anterior state and as the cause of the one which is to follow. Given for one instant an intelligence which could comprehend all the forces by which nature is animated and the respective situation of the beings who compose it — an intelligence sufficiently vast to submit these data to analysis — it would embrace in the same formula the movements of the greatest bodies of the universe and those of the lightest atom; for it, nothing would be uncertain and the future, as the past, would be present to its eyes.* ”
>
> Pierre Simon, Marquis de Laplace, *(Laplace, 1825/1985)*

1.1 Prelude: Laplace’s daemon and the daemon within

Consider what it might take to build an intelligent daemon of the kind that Laplace had in mind. Without full knowledge of the physical laws that govern the universe, it’s going to take a lot of data. But even all the data may not be enough. As it turns out, Laplace may have overstated the prospects for such a daemon. Quantum theory, for example, suggests limits either to the determinacy or the observability of the universe (Collins, 2007). Thermodynamic irreversibility may pose problems for backward inference too (Ulanowicz, 2012). Even at the limits of knowledge, cause and effect may be underdetermined each by the other, the deepest structure of the universe may remain latent. Notwithstanding the possibility that the universe itself is a computational device (Lloyd, 2002), it looks like fool-proof computation of one state of the universe from another is off the cards. It’s not clear whether the daemon will fare any better with regard to knowing “the respective situation of the beings who compose [the universe]”. Meta-inference – inference about the conclusions drawn by other inferential agents – also faces in principle limits.Wolpert (2008), for example, shows that at any one time the universe can contain at most one inference device able to infer the conclusions of all others. Too bad if we want a spare daemon. “How unsearchable his judgments, and untraceable his ways!” may generalise more broadly than originally anticipated.

So what of the prospects for the daemon within – the inference engine that drives our everyday cognition, from perception to social interaction and higher level reasoning? Living in a world with latent structure has consequences for it too. Because the predictions that most concern us are under-determined by the data we have, we must (for the most part) treat data not as the premises in a deductive proof, but as evidence for or against alternative possibilities. Evidence implies interpretation: representation and computation. In this sense, data alone is less evidence per se, more a component to be represented in the computation of evidence. Absent the guarantees of logical entailment, evidence is sustained in part by assumptions. Hume (1748/2007), for example, expressed the idea that human reasoning ultimately relies on the assumption that "the future will be conformable to the past" – a kind of *uniformity principle*. In addition to the in-principle constraints on real world inference, our efforts are further circumscribed by our access to the physical environment and the constraints of our biology. Nonetheless, from a genetic endowment amounting to less than one gigabyte of information[1], the human infant rapidly infers a great deal about the structure of the world despite these constraints.

Supported by an evolving understanding of the structure of the world and the observations it affords, this trend of reasoning beyond the data continues (and adapts) throughout the lifespan, particularly when the data is social in origin. At first glance it might seem a paradox that we could leverage socially generated data more highly than our own first hand observations. After all, if we endorse a uniformity principle for a world governed by enduring laws, then we can reasonably restrict our predictions to conform with our past observations of how those laws operate. But when it comes to socially generated data – which are at best abstract representations of data from the physical world – the possibilities are seemingly unconstrained. The flaw in this logic is readily apparent: the uniformity principle that we endorse for socially generated data represents not a weaker but a *stronger* set of constraints than those which we ascribe to nature. As speakers, though we might say anything, we don't – wheelbarrow – usually. We constrain ourselves to abide by norms of communication (e.g., Grice, 1989). As listeners we take advantage of these conventions to fill in the gaps between what is said and what is meant. Computationally speaking, this is one of our greatest achievements. By sharing abstract representations of the world with one another – exchanging the products of inference – we leverage a computational resource distributed across space and through time. In so doing, we greatly extend the evidence base that fuels our inference.

[1] Based on an estimate of approximately 3 billion base pairs in the human genome (Venter et al., 2001) each of which require two bits to encode.

1.2 The big picture

In the broadest sense, the research I present explores the challenges of inductive reasoning – the act of making guesses about data we don't have on the basis of data we do. By making assumptions about hidden structure in the world, we leverage the value of that data as evidence. Better evidence in turn supports more accurate predictions and more efficient learning about hidden structure, allowing us to further leverage the value of our assumptions. This research is focused, in particular, on understanding a key ingredient in this virtuous cycle of inductive reasoning. Namely, how we reason about our observations in relation to the processes that govern them, the structures that constrain them and the targets of learning and inference on which our sights are set. The over-arching thesis is that such reasoning plays an important role in driving the efficiency and accuracy of generalisation. When data is social in origin, the assumptions people make by reasoning in this way not only impacts their willingness to generalise, but is bound up in how they face other key challenges of inductive reasoning including the search for relevant evidence and the framing of the inductive problem itself.

The challenge of induction and the problem with evidence

To introduce these challenges by way of an example, consider the humble egg. Assume a reasoner (call her Alice), is in possession of an egg-like object and is trying to decide whether it is one or not. In keeping with a paradigm widely accepted and adopted in cognitive science, let's consider the problem from a computational perspective. In understanding any computational problem, the important elements to consider are the information involved, the level of abstraction at which it is represented and the computations performed.

Turning first to some computational issues. Setting aside a good deal of perceptual processing (which itself will involve inductive processess during scene segmentation, feature recognition etc.), assume Alice has a suitable perceptual representation of the object in question. What now? Suppose she recalls various eggs she has seen in the past. Even assuming memory is lossless (which it isn't) it's unlikely that Alice has seen an egg that precisely matches the object in question. Regardless, she would be ill-advised to base her classification on exact matching of all available features. A key challenge of inductive inference is to determine the relevant basis on which generalisation should proceed – that is, to determine the appropriate level of abstraction at which any comparisons are made. Alice should ignore irrelevant features (surface blemishes for example), and focus on those that matter (shape and texture, for example). But even relevant features need not be matched exactly. One reason to organise experience into conceptual representations in the first place is precisely because we believe that similar things afford similar consequences

> “ *Nothing so like as eggs, yet no one, on account of this appearing similarity, expects the same taste and relish in all of them.* ”
>
> David Hume (1748, p. 50),

(Hume, 1748/2007). Presumably, Alice is interested in identifying a would be egg because she cares about some of the consequences.

Which brings us to the central computational challenge of inductive reasoning: deciding how far to generalise prior knowledge and experience to new circumstances. Strictly speaking, this is taking an important step for granted, that of framing the inductive question to be answered. Different perspectives suggest different framings. For example, viewed from a reasoning perspective, Alice's task need only involve assessing on which side of a conceptual boundary the target sits – is it an egg or not? The question of how far the boundary extends (how round an egg can be, for example) may not need to be estimated. However, this latter question may well be of interest when viewed from a learning perspective. Whether these two questions are equivalent depends upon the examples involved and the computations employed. Alice might postpone the decision of how far to stretch her concept of an egg simply by observing that the current item lies within the range of those previously encountered. This potential short-cut takes another important aspect of the problem for granted, that of framing or constraining the range of solutions considered. In taking such a short-cut, Alice makes a law-like assumption that the concept EGG automatically applies to any item falling in the region of interpolation between known examples (presumably on many relevant dimensions).

In the way I have painted it, Alice's problem seems largely a perceptual one, solved on the basis of a (roughly global) similarity comparison between the item in question and previous examples of the concept drawn from memory. But as I have already hinted, even such a seemingly simple generalisation may leverage assumptions. Suppose for example, Alice were to reason about the digestive properties of duck eggs on the basis of her experience with chicken eggs: the conceptual properties of ducks and chickens may weigh more heavily than any perceptual similarities between duck eggs and chicken eggs. Similarly, whatever she might conclude regarding the maximum size of a chicken egg it will presumably be constrained by her beliefs about the maximum size of chickens. In general, inductive reasoning may draw upon a variety of information sources including perceptual, conceptual and theoretical evidence. But reasoning from evidence in this way brings challenges of its own.

Viewed from a decision making perspective, one challenge of inductive reasoning involves selecting evidence relevant to the problem at hand. If data is not evidence per se, but instead requires interpretation to establish relevance, then the problem of selecting evidence is especially challenging. Any assumption that supports a method of evidence selection short of weighing all possible data, has enormous potential for leverage. A related challenge arises when we consider inductive reasoning from the perspective of the learner. Here the challenge is, given the observations at hand, what exactly are these observations evidence of – that is, what can be reasonably be inferred from them?

Across each of the studies I describe in this thesis, I examine the assumptions people make about the data they observe and how these assumptions affect the way they deal with these challenges that I have identified.

So what are these assumptions and where do they fit in?

Imagine for example, that Alice has not seen eggs before and receives her first mixed dozen from the "Egg of the Month" club. Given that all of the eggs fit the carton snugly, what inference might she draw about how eggs range in size? And when she receives her next mixed dozen in a similar carton, should that change her concept at all? Or consider instead that she first encounters eggs by flicking at random through the pages of an illustrated guide to eggs (e.g., Kashimori, 2017). How might this change the evidentiary value of each egg she sees? The answer depends on the *sampling assumptions* she makes about the data. What is the generative process behind these examples? Importantly, how is it that she came to see these eggs and not some others? What constraints, if any, might restrict the range of eggs that she is seeing *in a way that actually matters*?

Viewed in this way, the challenge of reasoning from data resembles the practice of statistical inference. The reasoner considers the sampling distribution from which observations are drawn and attempts to connect it with some target distribution of interest. This involves understanding the ways that the sampling distribution is, and isn't, representative of the target distribution. In the example above, if Alice decides that the size of the eggs she sees was dictated by the size of the carton then seeing additional cartons of eggs might not help her reason about their size. Conversely, if she assumes that her book contains a representative range of eggs and their sizes, then each additional egg she views may be quite informative. After seeing a dozen examples and never having observed a 50cm egg she may begin to doubt that such a thing exists. Seeing a dozen more examples should make her increasingly certain of this. Of course, representativeness is not an all or nothing property of the object in question. Even if the egg samples[2] were unrepresentative in other ways (they may all have been especially clean, for example), Alice's inference regarding size may still be valid. The reasoner's task is to consider representativeness in light of various abstractions over the data. Naturally this does not imply that it is impossible to learn from unrepresentative data. If Alice observes a 75g egg she can use this to rule out the idea that all eggs weigh less than 50g, regardless of any biased sampling.

[2] I have held back this pun for long enough!

These examples illustrate the way that different sampling assumptions licence different generalisations by changing the weight of observations as evidence for the proposition being considered. And they highlight a key challenge for the reasoner in adopting assumptions in the first place – figuring out what the data is representative of (and learning fast) and what it is not representative of (and learning slowly). This problem is complicated by the presence of latent variables which may impact the sampling distribution in a way that is not clear. At stake is the quality of inference: incorrect assumptions imply inaccuracy (the wrong things are inferred) or inefficiency (learning requires more data). Which begs the question regarding how sampling assumptions are formed in the first place and the range of factors to which they may be sensitive.

The stakes are higher when inference is based on socially generated data. Real world constraints place obvious restrictions on both the kinds of things that can be "naturally" observed and the frequency with which observations can be made. Because socially generated data is relatively unencumbered it offers obvious benefits for efficient learning. It is possible, for example, to sample from a concept of interest without regard to the naturally occurring frequencies. But the relative lack of constraints introduces downside risk too. Firstly, the sharing of evidence is achieved indirectly via a process that amounts to the resampling of data. Reasoning from re-sampled data without access to the information supporting it (including background knowledge and assumptions, for example) can be an efficient way to draw the wrong conclusions. However, where shared culture and environment promote significant overlap in knowledge and assumptions, this risk is significantly mitigated. Secondly, while the constraints on real world data that restrict its availability may also make it difficult to falsify, this is not the case for socially generated data. Because inference on the basis of socially sampled data brings increased risk and benefit, by adopting strong assumptions about the sampling process the risks and benefits are further amplified. Adopting appropriate assumptions becomes increasingly important for reasoners in this case. An open research question concerns the contextual and content-based cues that reasoners use in calibrating their assumptions when data is social in origin, and the strategies they adopt to guard against misinformation.

The focused and explicit study of people's sampling assumptions is a relatively new research agenda (actively pursued only in the past two decades). Throughout this chapter (and in section 1.4 in particular), I will canvas a number of open questions (including those I have just flagged) concerning the effects of such assumptions on the outcomes of inference. Such questions lie at the core of the investigations I pursue in subsequent chapters.

What lies behind the research?

The theoretical perspective that informs this thesis is rooted in a literature spanning many decades of research. Ahead of a necessarily selective look at some of that literature, it is worth summarising this perspective in order to give a feel for the relevance of the topics discussed. Briefly then, the view sees people as model-based reasoners and holds that:

1. People make theoretical assumptions about the structure of the world in order to supplement the predictive utility of raw data (direct sensory experience).

2. The mental representations of these assumptions (i.e. the models) when combined with data are used to interpret the data in light of the assumptions, and determine its evidentiary weight. Thus, inductive generalisation is seen as a function of the weight of evidence, rather than simply as a function of the data itself.

3. Models are likewise used to evaluate and update the theoretical assumptions in light of the data.

4. Importantly, models compete. That is, people generate and test alternative assumptions (fragments of structured representations) across many levels of abstraction.

5. Ultimately, the models that persist (have the biggest impact on inference) are the ones that strike the better balance between cognitive efficiency (generational, representational, and computational efficiency, for example) and cognitive effect (accuracy and utility, for example).

In this model-based view of inductive reasoning, the reasoner's sampling assumptions can be thought of as that part of their mental models that sits at the interface between representations of observations and representations of concepts, theories and conjecture.

Although this perspective is sufficiently abstract and hence difficult to directly falsify, the presumptions above do nonetheless suggest hypotheses more amenable to testing. For example, the idea that "models compete" suggests that we should expect to see some competition among sampling assumptions, in much the same way that we might expect to see it with regard to theoretical assumptions. The experiments I present in Chapters 5 and 6 provide some evidence that people are adapting their sampling assumptions to the data they observe which suggests they may represent more that one assumption at any given time. Similarly, Chapter 2 looks at how people's sampling assumptions may influence the way that they evaluate alternative representations of the solution space for a given inductive problem.

The shape of things to come

My goals in this chapter are threefold: to paint a picture of where sampling assumptions fit in the overall scheme of inductive inference, to present the case for why sampling assumptions effect inference in ways that actually matter, and to highlight the open questions that are the subject of my research. In service of the first goal and to flesh out the theoretical perspective that informs my research, I first provide a selective review of the literature as it relates to inductive generalisation. I focus in particular on how different theories have characterised the sources of "evidence" on which inductive inference relies. I discuss how inductive generalisation may draw upon inter-stimulus similarity, on distributional information embedded in conceptual space, and on a range of contextual information and theoretical knowledge.

After sketching the backdrop against which sampling assumptions might play a role, I next provide a brief overview of Bayesian inference and its use as a computational tool for modelling inductive inference in general but sampling assumptions in particular. Because the Bayesian framework makes explicit the contribution to inference from various sources of information as well the reasoner's sampling assumptions, it can provide a compelling prima facie justification of why sampling assumptions should matter. After illustrating the computational basis for why sampling assumptions ought to matter, I review contemporary research that demonstrates important ways in which they actually do, and highlight open issues relevant to the new research I describe in this thesis.

The studies I present examine the nature and effect of people's sampling assumptions in three distinct contexts. Each context emphasises a different challenge of induction and a different factor driving the interpretation of evidence:

1. The first context, which covers the experiments described in Chapters 2–4, involves generalisation on the basis of low-dimensional, perceptual stimuli. By providing a context that makes clear the basis on which generalisation should proceed, the experiments focus on the central inductive challenge of determining how widely to generalise on the basis of the data supplied. By varying sample size and the number of stimulus categories involved, the studies yield insight regarding how people's sampling assumptions interact with these factors.

2. The work in Chapter 5 employs a category-based induction task where the basis for induction is unclear due to use of the high-dimensional conceptual stimuli and essentially unknown ("blank") properties. Here the focus is on examining the reasoner's challenge of determining when potential evidence is and isn't relevant.

3. Lastly, Chapter 6 establishes a context where assumptions about what another reasoner is thinking are the primary means for interpreting evidence. By establishing a context where the cooperation of the communicating party cannot be taken for granted, the work examines inference in the case where the appropriate sampling assumption (regarding the way that the information provider is sampling data) is unclear. The study highlights the way that people draw on both the content of what is communicated (the data) and the context in which communication takes place in order to calibrate their assumptions about the evidentiary value of the data.

I conclude the chapter with a more detailed overview of the motivation behind the new research.

1.3 Inductive generalisation: reviewing the evidence

The following review is organised around a simple premise. The representational complexity of observations and the conceptual space in which they are embedded go hand in hand with the richness and complexity of our generalisations. And it cuts both ways. Demand for more accurate inference fuels the need for richer representations, and richer representations support more detailed inference. This relationship between representational and inferential complexity has been paralleled to a degree in the unfolding development of a theory of inductive generalisation. As we shall see, representations of stimuli as points in low dimensional metric space (such as Shepard, 1987, for example) goes a long way to describe how people (and animals) generalise on the basis of simple perceptual stimuli. But with the shift in the focus of study from percept to concept, the notion of metric distance runs into problems. Similarly, the notion of features and featural overlap (Tversky, 1977), while able to describe the non-metric properties of empirical judgments, struggles to capture phenomena that involve knowledge of relationships with stimuli not directly involved. The idea that individual generalisations may involve hierarchically structured concept representations organised on the basis of featural similarity (as in Osherson, Smith, Wilkie, Lopez, & Shafir, 1990, for example) allows further generalisation phenomena to be captured. Yet, as I will explain, the notion of a statistically coherent concept, while powerful, is insufficient to explain the role that people's theorising plays in their inductive inferences, particularly where the information from which people reason is social in origin.

Stimulus generalisation: more than meets the eye

Almost a century ago Pavlov published his landmark study of classical conditioning (Pavlov, 1927). In studying the secretion of saliva in dogs he conditioned the animals to salivate in response to a particular stimuli (the sound of a bell or whistle, for example). After conditioning, he observed that the same kind of reaction (secretion of saliva) although reduced in degree, could be evoked by similar stimuli that had not been directly conditioned. This pattern of behaviour was widely replicated by early researchers in the field both in animals and humans, raising a number of important questions and sparking considerable debate.

Does stimulus generalisation reflect a physiological constraint?

Perhaps the most fundamental question raised, concerned the very nature of stimulus generalisation and the mechanisms underlying it. One possibility was that stimulus generalisation in the form observed by Pavlov and others might be a result of stimulus confusion – that is, the failure to discriminate the conditioned stimulus from the generalised stimulus. Such a failure might, for example, be underpinned by a constraint of the perceptual system regarding the "resolution" at which the senses operate. But Pavlov's (1927) work and other studies of discrimination learning suggested that the ability to distinguish the conditioned stimuli from similar stimuli could be acquired. For example, Pavlov found that it was possible to constrain (but not eliminate) generalisation by repeating the same conditioned stimulus a number of times (over 1,000 times in some cases).[3] Notably, these experiments involved a single category of reinforcement – that is, all repeated presentations were accompanied by the same (positive) reinforcement. In contrast, when conditioning involved two categories of reinforcement (positive reinforcement and no reinforcement), it was possible to induce stimulus differentiation with a single (unreinforced) application of a series of related stimuli. Likewise, Hovland's (1937) study of galvanic skin response in humans which showed stimulus generalisation up to 100 JND[4] units distance from the conditioned stimulus, casts doubt on the notion that stimulus generalisation reflects a physiological constraint of the perceptual system.

Does the context in which stimuli are sampled affect generalisation?

As noted, Pavlov's (1927) early work demonstrated an important effect of framing: that is, whether one or two categories of reinforcement were used significantly impacted how the extent of generalisation changed with increased sample size. However, the direction of this effect was not always reliable. For example, while Hovland (1937) and Pavlov (1927, p.117) found that generalisation to other stimuli diminished as a result of repeated reinforcement of the conditioned stimulus, Margolius (1955) found the reverse effect – additional samples broadened the degree of generalisation. Margolius, in acknowledging the discrepancy with Hovland's results, dismissed the differences in experimental paradigms (he had employed operant conditioning, where Hovland had used classical conditioning). Nonetheless, it is interesting to speculate from a sampling assumptions perspective whether those differences might have influenced results. In Hovland's experiment, the association between a tone and an electric shock is made on every trial in an otherwise static stimulus environment. Therefore each trial offers a fresh example of the kind of tone that accompanies a shock. In Margolius's experiment relevant data is collected only when the subject (a white rat in this case) exhibits the target behaviour (pushing on a door in the target stimulus - a 79 cm^2 circular white disc). While the environment provides ample opportunity to learn the affordances of the same white disc, it offers no repeated opportunity to learn about other white discs.[5] In Chapters 2

[3] These experiments might be regarded as an early demonstration of the effect of sample size on generalisation.

[4] Just Noticeable Difference: the threshold below which the difference between two points along some physical dimension (such as frequency of sound, for example) cannot be perceived reliably (typically with $> 50\%$ or $> 75\%$ accuracy over repeated trials).

[5] Presuming of course that the rats in question mentally represent the spatio-temporal extent of the white disc as a singular entity.

and 3, I consider the effect of sample size on generalisation and its connection with the context in which stimuli are sampled.

The representational basis for stimulus similarity

While there was wide support for the view that both humans and animals were capable of "true generalisation" (as opposed to discrimination failure)[6] , the basis of stimulus similarity was the source of considerable debate. Many researchers held the view (emphasised by the Gestalt movement) that animals and people were responding to the situation as a whole, and generalising relational concepts such as "bigger", "brighter" and so on. Against this holistic view, was the "sensation driven" view which stressed the role of the stimuli and its absolute properties. In essence this was a debate about the representations supporting comparison and generalisation.

In framing his theory of stimulus generalisation, Hull (1943) speculated that stimuli were represented by the perceptual system through the discharge of a collection of afferent "molecules", and that receptors adjacent on the generalisation continuum might overlap in terms of the molecules discharged. Such "molecular" similarity might, he contended, explain the shape of primary generalisation gradients and the gradual acquisition of stimulus discrimination over repeated trials. The idea that experimental settings and individual stimuli could be broken down into a number of smaller *stimulus elements* would also form the basis of later statistical learning theories. For example, Estes (1950), in modelling the probability that a conditioned stimulus will evoke the target response, defined a learning rate parameter as the (mean) fraction of the relevant stimulus elements "sampled" on a given trial. As more trials go by, the probability of sampling a previously unconditioned stimulus element decays exponentially, so the probability of the conditioned response increases towards certainty. Bush and Mosteller (1951) captured stimulus similarity in a similar way, defining the similarity between two stimuli in terms of the elements in common. But by expressing their *index of similarity* between two stimuli S' and S as a proportion of all elements of S (Bush & Mosteller, 1951, Eqn. 4), they captured the inherent asymmetry involved. Their definition is closely related to two important models I will discuss later in this chapter: the contrast model of featural similarity (Tversky, 1977), and the Bayesian model of concept generalisation (Tenenbaum & Griffiths, 2001a, *cf* Eqn. 11).

Stimulus generalisation and the region of interpolation

The fact that any form of stimulus-response conditioning is possible in the first place was not without issue. Given noise in both the perceptual system and in the physical environment itself, it is highly unlikely that the same "stimulus experience" is ever precisely repeated. If this is the case, then how does a given stimulus-response connection

[6] Although for a different interpretation from this early period, see Lashley and Wade (1946).

become established if, as is typically the case, more than one reinforcement is required? Likewise, even if such a connection could be established after a single reinforcement, how could the response ever be invoked again unless the stimulus was repeated? Hull (1943) offered a solution for these two conundrums – the *stimulus learning paradox* and the *stimulus evocation paradox* – with his summation model of primary generalisation. He suggested that each stimulus presentation, even if insufficient on its own to produce a reliable stimulus-response connection, nonetheless forms some strength of association (*habit strength*) in a single trial. Further he held that there is a narrow region of generalisation around the reinforced stimuli which (when the stimulus dimension is represented in JND) follows a pattern of exponential decay. Thus, repeated reinforcements of the "same" stimuli, establish a series of these overlapping generalisation gradients emanating from nearby points in the stimulus dimension. While each single reinforcement (each individual gradient) is not enough to promote generalisation on its own, the resulting combination is, as Figure 1.1 illustrates. Hull's rule for combining gradients has strong connection to a Bayesian formulation of generalisation where each individual presentation of the stimulus is assumed to be sampled from an independent "consequential region" (as in, Navarro, 2006).[7]

Is stimulus generalisation a training or testing effect?

Another important debate in the early generalisation literature concerns the kind of effect that conditioned stimulus generalisation represents. Models such as those of Hull (1943), Estes (1950) and Bush and Mosteller (1951), are based on the idea that the nature of a generalisation gradient is determined by the "bonds" that are formed when stimuli are first represented (as afferent molecules or stimulus elements, for example). Against what we might call this "encoding" or ''learning" view of generalisation stands a "retrieval" or "testing" view. For Lashley and Wade (1946), stimulus generalisation was very much a retrieval effect because it represents, they held, a failure to discriminate. Similarly, Razran

[7] The combined generalisation response strength (habit strength) at any given point (${}_{\mathrm{S}}\overline{H}_{\mathrm{R}}$) is calculated by combining the contribution of each of the n primary gradients at that point according to the following "summation" rule:

$$_{\mathrm{SS}}\overline{H}_{\mathrm{R}} = \Sigma_1 - \frac{\Sigma_2}{M} + \frac{\Sigma_3}{M^2} - \cdots + (-1)^{n-1}\frac{\Sigma_n}{M^{n-1}}. \tag{1.1}$$

where Σ_i represents the sum of all products of contributing points taken i at a time. M represents the physiological limit (maximum) of habit strength, and acts to ensure that each successive term in the above equation is lower in magnitude than the previous. This seemingly complicated formula has an interesting probabilistic interpretation, as follows. Suppose that for any given test stimulus there is some chance that the target response will happen based on the resemblance of the test stimulus to one of the reinforced stimuli, and that this probability is reflected in the primary generalisation curves (dividing the habit strength by the physiological limit to achieve the unit interval). Further assume that each reinforced stimuli acts as an independent "cause" of responding. Determining the overall probability that the target response will be generated amounts to calculating the union of each of the independent events, the formula for which follows the same sum of products rule captured in Equation 1.1.

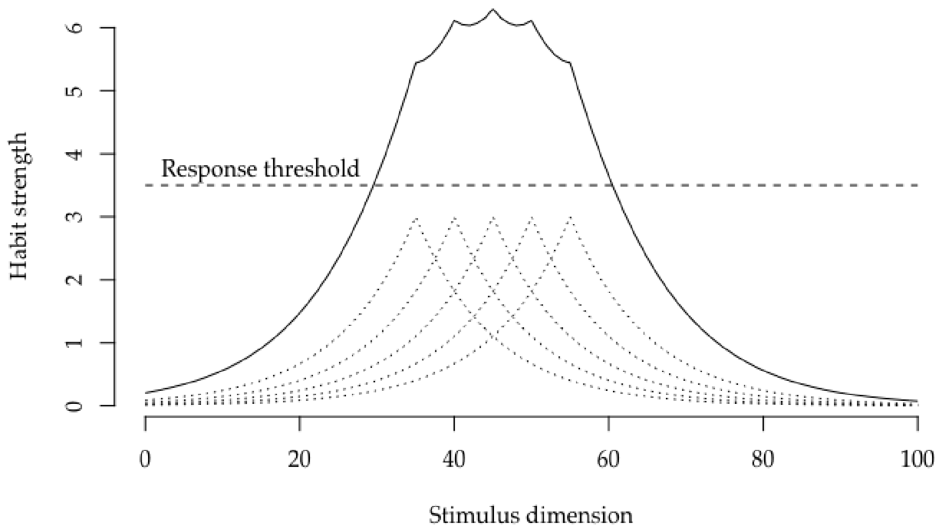


Figure 1.1: A plot illustrating Hull's summation rule (derived from, Hull, 1943, Fig. 46), predicting stimulus-response generalisation (solid line) on the basis of gradients arising from individual reinforcements (dotted lines). While generalisation on the basis of any single presentation is below the response threshold, the effect of repeated presentation (perceived as adjacent points on the stimulus dimension, due to perceptual noise) is sufficient for generalisation anywhere in the region of interpolation.

(1949) in advocating his *subsequent testing hypothesis*, held the view that "all effects of generalisation are generated during tests of generalisation". In an experiment using unfamiliar words, he contrasted the generalisations of people given conflicting meanings before and after conditioning (but before testing) with people given meaning only after conditioning. He concluded that the meanings given prior to conditioning had no effect on generalisation. At the centre of this debate is the issue of whether the evidentiary weight of data (perceptual stimuli in this case) is captured during encoding, or merely evaluated during reasoning when the data is retrieved. I consider this issue further in Chapter 4, where I attempt to distinguish whether sampling assumptions affect generalisation at the point of learning or only when reasoning.

Generalisation gradients in psychological space

Another challenge facing early researchers of stimulus generalisation concerned how to represent the value of stimuli in order to make sense of generalisation data. Hull (1943) suggested that JND represented the appropriate unit for what he called the *afferent generalisation continuum*. Others noted that the use of JND in this regard was not without its problems (e.g., Humphreys, 1939). Bush and Mosteller (1951) left the connection between the physical dimensions of stimuli and their index of similarity unspecified, speculating that any universally reliable measure of stimulus similarity might remain elusive. Shepard (1957), however proposed a solution to just this problem with his model of stimulus (and response) confusion. He proposed to turn the problem on its head. Instead of trying to use an arbitrary measure of psychological distance (such as JND, for example) in an effort to understand confusion probabilities, he started with generalisation data (confusion probabilities) and looked for invariant monotonic function whose inverse would transform numbers into distance in some psychological space which obeyed the metric axioms of distance. He found that an exponential decay function had the necessary mathematical properties and captured empirical data. Shepard's finding was also consistent with Hull's (1943) earlier use of the exponential decay function to capture generalisation data.

* * *

Early studies of stimulus generalisation, particularly those employing low dimensional perceptual stimuli, tended to focus on issues of discrimination, identification and classification. When considering such "whole of stimulus" questions, the literature had placed much emphasis on understanding the basis of global similarity judgments between pairs of items. But to generalise our experience of the world by comparing sensory snapshots would cost too much (in terms of capacity and computational complexity) and yield too little (in the way of reliable inference). So what kinds of representational abstractions do people employ to support efficient and reliable generalisation? I turn now to outline some important developments in the literature concerning generalisation with respect to conceptual representations that are both richer and more abstract than perceptual encodings of sensory data.

From percept to concept: representation-rich generalisation

In his seminal work *The Organization of Behavior*, Hebb (1949) laid out his theory of *schema* and the *cell assemblies* of which they are composed. He was seeking to bridge two important gaps that he perceived at the time. One was a theoretical and empirical gap between physiological and psychological phenomena; the other, not entirely unrelated, concerned the holistic Gestalt view of perception on the one hand and the association-focused view of learning theorists on the other. Although he didn't express it in quite this way, Hebb's schema reflected the idea that the evidentiary value of data was influenced over time by the configurations in which the data was typically observed. He contended that we learn to perceive abstract configurations of stimuli (a triangle, for example) as a function of practice, and that the schema that we form in so doing makes subsequent identification easier. Importantly, Hebb saw this process of abstraction as something that operated at multiple levels from the perceptual to the conceptual.

Attneave (1957) sought to test Hebb's notion that schema formation benefited later identification. In two experiments using a paired-associates task, he found that people pre-trained on a prototype stimulus made fewer identification errors than those who received no such pre-training. Subsequently, Hinsey (1963) clarified that the pre-training was indeed more effective when the stimuli used reflected the central tendency of the test stimuli (as was the case in Attneave's study) rather than a more peripheral example. In a landmark study in the development of prototype theory, Posner and Keele (1968) built on these earlier results by demonstrating that people could extract a central prototype without explicit pre-training. In many instances the previously unseen prototype was classified with greater accuracy than were the control items with which people had been trained. Relevant to the discussion in Chapter 4, the authors found they could draw no conclusion whether the formation of abstraction happened during the learning process or spontaneously during testing as part of the classification process. Rosch (1973) was less circumspect however, suggesting that categories in domains such as colour and form may develop around perceptually salient focal points (which need not be central). She provided an elegant empirical demonstration (Rosch, 1975) of the asymmetry inherent in such category formation – namely, that non-focal stimuli were considered closer to focal stimuli than the other way around. Rosch and Mervis (1975) introduced the notion of *family resemblance* as a way of accounting for graded category membership. They argued that semantic categories, rather than represent a list of "must have" properties, should be thought of in terms of a network of overlapping features. Carving up a conceptual space according to basic-level categories was seen as a process of maximising family resemblance (featural overlap) within categories while minimising it between categories.[8] Their empirical results demonstrated strong correlation between an item's family resemblance score and subjective ratings of proto-typicality.

Featural similarity

A related notion of featural overlap formed the basis of Tversky's (1977) *contrast model* of similarity. The model defines the similarity between two items x and y as:

$$S(x,y) = \theta f(\mathcal{Y} \cap \mathcal{X}) - \alpha f(\mathcal{Y} - \mathcal{X}) - \beta f(\mathcal{X} - \mathcal{Y}) \tag{1.2}$$

where $\mathcal{X}$ and $\mathcal{Y}$ are features sets representing the items x and y respectively. The parameters θ, α and β reflect the relative importance of the common and distinct features, and the function f defines a measure of salience for a given feature set which may involve intensity, frequency, familiarity and so on. The simplicity and generality of Tversky's model belies its significance. The model has the parameters that it does because Tversky painstakingly demonstrated that both feature salience and the relative importance of common and distinct features varies with the stimulus context and with the nature of the experimental task. However, viewed from a different angle, the free parameters in the model reflect just what it is that remains elusive when it comes to understanding

[8] In line with Carnap's (1967) earlier idea that a set of items is a kind if its members are more similar to each other than any given thing outside the set.

similarity and its role in inductive generalisation. For example, what constitutes a feature in the first place and which features are relevant in a given setting? I consider this latter problem in Chapter 5, in light of the relevance theory of induction (Medin, Coley, Storms, & Hayes, 2003), and demonstrate that the context in which the stimuli were sampled has an important role to play in determining when seemingly relevant featural overlap is indeed relevant.

Tversky's model of featural similarity, as well as Rosch and colleagues' work on prototypes and family resemblance, represent an important milestone in the understanding of inductive generalisation. Accompanying a shifting trend in the focus of study from percept to concept, the models highlighted the need to consider a more flexible mixture of cue validity and psychological distance as the principle determinant of stimulus similarity and the driver of generalisation behaviour. Notions of asymmetric featural overlap between items as well as the distribution of features within and between related categories would inform subsequent models of property induction.

Generalisation of abstract properties

Structuring our internal psychological space by organising our observations of the world into conceptually coherent concepts offers the potential to reduce computational complexity and ease the burden on long term memory. And insofar as the structure of our category representations reflect statistical regularities in the world, we can use such representations to infer knowledge about the world by projecting properties from one category to another. Which raises a number of questions about how such *property induction* might work. What, for example, constitutes good evidence in support of a particular induction? How does this depend on the nature, number and range of known examples? And how is existing knowledge of the categories involved and of the property itself, brought to bear?

In one of the first empirical studies of its kind, Rips (1975) asked people to project a relatively abstract property (an unknown disease) from one animal species to another. He was interested in the the notion of graded category membership (of the kind proposed by Rosch, for example) and whether this might influence property projection. He found an effect of similarity on property induction that would later be widely replicated — namely that willingness to project a property from the given item to the target item was influenced by the similarity of the two. Thus, people were more willing to generalise ROBINS → SPARROWS than ROBINS → HAWKS, because robins and sparrows were judged the more similar. However, he also observed a fundamental asymmetry in such projections. People were more willing to generalise from typical members of a category to less typical members (e.g. ROBINS → GEESE) than the reverse (GEESE → ROBINS). But this effect of typicality could be made to disappear if people were provided more specific information about the distribution of the property in question. Rips concluded, that in the absence of more specific knowledge, people were leveraging the distributional information in the superordinate category (BIRDS, for example) and not merely the features of the items involved. Properties belonging to more typical members were more likely to be shared by others than properties belonging to atypical members. Rips's stance on the role of premise items in inductive judgments reflected a view echoed in later theories – namely, that an item's evidentiary value reflects its direct overlap with the target in terms of observed (largely perceptual) attributes *as well as* theoretical properties derived from the distributional information implicit in the categories involved and the conceptual space in which they are embedded.

Osherson et al. (1990) adopted a similar stance in developing their *similarity-coverage* model designed to capture argument strength ratings in category-based induction tasks, like the following:

> Elephants have BCC in their blood.
> Monkeys have BCC in their blood.
> ―――――――――――――――――――
> Antelopes have BCC in their blood.

where one or more premise statements are given (above the line) in support of a conclusion (below the line). To avoid the effect of property-specific knowledge on judgments, the premise statements relate to so-called *blank predicates* about which people should have limited prior beliefs. According to the model, an argument derives its strength from two sources of evidence. The first source, the *similarity* component, derives from a direct featural comparison between premise and conclusion items – specifically, the shortest path in psychological space between the conclusion category and any one of the premise categories. It reflects the evidence of the premise items for the specific conclusion at hand, and is independent of any fixed category hierarchy. The second source, the *coverage* component, is more indirect, deriving from membership in a common superordinate category (the smallest one containing the premise and conclusion items). It reflects the weight of evidence that the premise items represent for *any* item sharing the common relationship – in effect, the shortest path between conclusion and premise items, averaged over all potential conclusions.

Osherson et al. (1990) applied their model to successfully account for a wide range of previously observed phenomena including similarity, typicality and diversity effects. Yet despite the model's crucial reliance on the coverage component in accounting for many of these phenomena, Sloman (1993) achieved comparable accounts with his purely feature-based model of property induction. In this model, argument strength relates to how well the features of the conclusion category are covered (overlapped) by the features of the premise categories. Thus, when holding the conclusion category constant, argument strength increases as the featural overlap between premise and conclusion categories increases. And when holding featural overlap constant, argument strength decreases as the featural complexity of the conclusion category increases. Even without a mechanism for leveraging knowledge implicit in a common superordinate category, Sloman gave a plausible account for all effects accounted for by Osherson et al., excluding that of premise non-monotonicity where adding positive evidence can weaken a conclusion. According to Sloman's model, adding premise items can never decrease featural overlap (by definition), so argument strength can never weaken as consequence.[9]

The diversity of evidence

Indeed, it seems intuitively reasonable that adding further positive evidence should act to increase argument strength, or at least have no effect. Certainly philosophers of science (e.g., Hempel, 1966; Horwich, 2016) have emphasised the role of diverse evidence in strengthening the support for a hypothesis. Osherson et al., however, make an important distinction between evidence coverage and evidence diversity, one which attempts to explain when premise diversity does and does not promote argument strength. The following arguments illustrate the distinction:

$$\begin{array}{l} \text{Flies have property X.} \\ \hline \text{Bees have property X.} \end{array} \tag{1.3a}$$

$$\begin{array}{l} \text{Flies have property X.} \\ \text{Moths have property X.} \\ \hline \text{Bees have property X.} \end{array} \tag{1.3b}$$

$$\begin{array}{l} \text{Flies have property X.} \\ \text{Cats have property X.} \\ \hline \text{Bees have property X.} \end{array} \tag{1.3c}$$

Argument 1.3(b), with the addition of the second premise: MOTHS, increases both the diversity and coverage of the argument. The addition of CATS in argument 1.3(c), in contrast, increases premise diversity but *decreases* the argument's coverage by changing the common reference class from insects to animals. Thus, under the similarity-coverage model, the second argument is predicted to be stronger than the first while the third is predicted to be weaker.

In Chapter 5 I explore a different kind of diversity effect, or rather a *non-diversity* effect, where the addition of premises that are insufficiently diverse may act to decrease argument strength. This kind of premise non-monotonicity cannot be accounted for by the similarity-coverage model, since by definition additional premises can never reduce the similarity term, nor the coverage term where the covering category remains the same. I argue that the phenomena depends on the reasoner holding a particular kind of theory about the process by which arguments are constructed. In a similar vein, Hayes, Navarro,

[9] Sloman (1993) suggested that his model could be extended by adding a regularisation parameter (designed to reduce overfitting in connectionist models) in the form of weight decay. In this way, the model could plausibly account for premise non-monotonicity.

Stephens, Ransom, and Dilevski (2019) show that people's general tendency to draw stronger conclusions from more diverse evidence is somewhat contingent upon the model of premise selection that they have in mind. The inability of the similarity-coverage model to capture such effects reflects the fact that, by design, the model is fundamentally similarity-based. While it does incorporate one source of the reasoner's background knowledge – namely, a category hierarchy (specifically a biological taxonomy in the case of the particular arguments studied) – it has nothing to say about other ways in which the reasoner's background knowledge and theories may be brought to bear.[10] And because its notion of similarity is essentially a global one (i.e. based on all available features of a category), it cannot make predictions about changes in feature salience that depend upon the property being projected. So if recruiting property specific knowledge and determining feature relevance is an important part of what the reasoner does, then it is an important challenge for cognitive theory to explain it. In the next section, I outline various efforts aimed at incorporating such explanations into the theory of inductive generalisation.

Theoryful induction: finding reasons to generalise

Similarity and its role in supporting inductive generalisation was much scrutinised in the latter half of the 20th century. Philosophers found it difficult to define (e.g., N. Goodman, 1972; Quine, 1969), and experimental results showed that judgments of similarity were highly context sensitive (Medin, Goldstone, & Gentner, 1993; Tversky, 1977). Such scrutiny led to a considerable broadening of the concept of similarity. As a result, a more detailed view of inductive generalisation begun to emerge. One where reasoning involves judgments of similarity that are theoryful (based on theories, hypotheses and explanations) as well as theoryless (i.e. pre-theoretic, based on global comparisons and distributional statistics). In what follows, I highlight some important conceptual turning points towards a more "theoryful" theory of inductive reasoning.

From similarity of kinds to kinds of similarity

As I have mentioned already, while the models proposed by Osherson et al. (1990) and by Sloman (1993) can both accomodate considerable feature richness in terms of the premise items involved in an argument, they are both limited in another important respect. Neither model has anything to say about the involvement of the predicate itself (i.e. the property being projected). This is unsurprising since both models were designed to capture reasoning with blank predicates. Nonetheless, as N. Goodman (1955) suggested,

[10] Such design limitations notwithstanding, Osherson et al. (1990) marks an important theoretical and empirical contribution to the development of a comprehensive theory of induction. In a parsimonious fashion, it offers a mechanism by which indirect but related evidence – distributional knowledge of relevant superordinate categories – may be brought to bear on a variety of problems.

not all properties are equally projectable. Consider the following arguments, for example:

$$\frac{\text{This piece of copper conducts electricity.}}{\text{All pieces of copper conduct electricity.}} \tag{1.4a}$$

$$\frac{\text{This man in the room is a third son.}}{\text{All men in the room are third sons.}} \tag{1.4b}$$

Assuming you did not already know that copper conducts electricity, the first argument still seems stronger than the second.

Unsurprisingly, N. Goodman's intuition is borne out in the laboratory. People are more willing to project homogenous biological properties such as skin colour, for example, than more heterogeneous ones such as obesity (Nisbett, Krantz, Jepson, & Kunda, 1983). Furthermore, the variability in property projectability may itself vary as a function of the category in question. Keil, Smith, Simons, and Levin (1998), for example, showed that children as young as five considered "number of inside parts" to be more definitive than "dusty" for a given category (animal or machine), whereas "surface markings" was important for animals but not machines, and "size" was somewhat important for machines but unimportant for animals.

Heit and Rubinstein (1994) examined the notion of projectability by testing whether people take more than a single kind of similarity into account when projecting a property. Their results demonstrated that people make stronger inferences when the property to be projected (anatomical or behavioural) matched the kind of similarity exhibited between the animal categories used (Heit & Rubinstein, 1994, Expt. 1 & 2). They further observed that while both anatomical and behavioural similarity influenced projection of behavioural properties, only anatomical similarity influenced anatomical inferences. Heit and Rubinstein suggested that inductive reasoning, is a dynamic process where people identify those features of the categories involved that are relevant to the property being inferred. The suggestion, which is echoed in later theories (such as Medin et al., 2003, for example), makes a good deal of sense. From a philosophical standpoint, N. Goodman (1972) had argued that a category's features may be unbounded. Empirically, Barsalou (1989) had observed that when people were asked to list a category's features, the resulting feature sets tended to be context dependent. And there was precedent for the suggestion in Tversky's (1977) contrast model of similarity, where the problem of determining property-specific feature relevance is analogous to the context-sensitive problem of determining non-zero feature weights. An obvious question arises however. If the property being inferred is used to determine feature salience in similarity judgments, what kinds of knowledge or theories might people be recruiting? As I hope to demonstrate in the discussion that follows, consideration of issues such as these has helped to stimulate the development of a richer theory of inductive inference.

From comparison to reasoning: the emergence of theory

N. Goodman (1955) in attempting to address his own riddle of induction, was arguing in effect against a theoryless notion of similarity as the sole driving force behind induction. He proposed that the way we determine whether a property represents a suitable basis for induction is itself a kind of second order induction. That is, by trying out a property as the basis for a generalisation and finding some success, the property may come to be *entrenched*. And if success breeds success, the rich, inductively speaking, get richer. One property is thus a more suitable basis for induction than another to the extent that it has become the better entrenched. These second order inductions (or *overhypotheses* in N. Goodman's terms) may be viewed as a kind of proto-theory adopted by the reasoner to circumvent the otherwise computationally intractable task of comparing similarity on the basis of a potentially unbounded number of features.

Quine (1969), like N. Goodman (1972) was wrestling with the slipperyness of similarity, and in particular its suitability as a measure of conceptual coherence. Quine makes an important distinction between concepts whose coherence derives from the proximity of its members in similarity space (what he terms our *innate quality space*), and those organised on the basis of a more sophisticated *theoretical* understanding of what lies beneath the similarities. Quine suggests that our development from early childhood follows a gradual trajectory whereby our concepts are restructured increasingly on the basis of theoretical coherence. Murphy and Medin (1985) build on Quine's idea regarding the theoretical coherence of concepts, suggesting that neither (theoryless) similarity alone, nor the statistical correlation of attributes is sufficient to capture the coherence of people's internal representations of real world categories. They suggest that concepts seem coherent to the degree that they match people's background knowledge and assumptions, making a kind of "theory first" argument about conceptual organisation. This view suggests that correlational structure is represented by causal theories or explanations, and that similarity "may be a by-product of conceptual coherence rather than its cause". Importantly, Murphy and Medin emphasise the bi-directional influence of mental concepts and theories; theories are composed from concepts, and in turn constrain the features represented in concepts.

The idea that Murphy and Medin (1985) along with Quine (1969) were driving at reflects an important shift in thinking about the way in which concepts are structured, and the *raison d'être* for such structure in the first place. The theory-centric view of concepts places a different emphasis on what the reasoner is trying to achieve. The focus shifts from answering "how similar are these things?" to "why are these things similar?". Computationally speaking, the two questions (in principle at least) might involve different trade-offs regarding representational complexity and generalisability. Questions of "how similar?" might be better addressed by concept representations structured around a theoryless notion of similarity. While for questions with a "why?" focus, a theoretically coherent organisation of the kind that Murphy and Medin (1985) described, might be more appropriate. Certainly the ability to answer "what lies behind the data?" offers potential benefits. A reasoner who incorporates such thinking into their inductive inferences gains the ability to leverage past experience in a wider variety of situations than might otherwise be possible – effectively learning more from less. Whether the right thing is learned will depend of course on the particulars of the reasoner's theories and assumptions. But, as I attempt to highlight through the studies in this thesis, people's implicit answers to "why this?" style questions impact how far they generalise (Chapters 2–4), how they determine what's relevant (Chapter 5) and how they perpetrate and avoid misdirection (Chapter 6). While the nature of concept representation is a source of ongoing debate, the ideas around explanatory coherence and theory-guided generalisation influenced a number of subsequent theories of induction.

Explanatory coherence

Sloman (1994), for example, investigated the evidentiary value of explanatory coherence using pairs of arguments like the following:

$$\frac{\text{Most investment bankers are at risk for heart attacks.}}{\text{Most firefighters are at risk for heart attacks.}} \tag{1.5a}$$

$$\frac{\text{Most investment bankers retire early.}}{\text{Most firefighters retire early.}} \tag{1.5b}$$

where there were clear but potentially conflicting explanations for the property to be projected. He found that people's ability to generate an explanation for the premise and conclusion had a significant impact on their perception of argument strength. Where the explanation generated to explain the premise was also a plausible account of the conclusion (as in the first example above), the premise increased people's willingness to endorse the conclusion. But where the explanations were seen to be unrelated (as in the second example), endorsement *decreased*. In the latter case, this demonstration of what amounts to premise non-monotonicity is quite striking. That an unrelated observation should offer no evidence in support of a conclusion seems reasonable. That it should somehow count as negative evidence requires some explanation. One possibility, is that people are pre-disposed to look for a single causal explanation. For example, learning that investment bankers retire early (presumably to enjoy the fruits of so much hard labour) might thus reasonably discourage the belief that early retirement is an option for firefighters (whose labours bear less fruit). Sloman raises another possibility – that people engage in a form of *abductive* inference (Harman, 1965; Pierce, 1955), and that an available explanation applying to the premise but not the conclusion (as in Argument 1.5b) may nonetheless inhibit the generation or retrieval of an alternative. A further intriguing possibility is that people make assumptions about the pragmatic context in which arguments are generated which raise heightened expectations of premise relevance

(Medin et al., 2003). The availability of an alternative explanation may be sufficient for people to conclude that the premise is intended as negative evidence. The idea that the rules of evidence change when data is socially generated is a potent one that may help to explain many reasoning phenomena which otherwise seem unjustified or illogical on the basis of "naturally occurring" evidence. This is not to say that socially generated data licences an anything goes approach to inference.[11] Indeed a goal of the research described in this thesis (and Chapter 6 in particular) is to explore the systematic ways that inference on the basis of socially generated data is related to and builds upon the more "primitive" kinds of generalisations we make.

Sloman's (1994) *competing explanations* are essentially alternative theories that the reasoner holds regarding feature relevance. Namely, which feature or features should support the projection of the property in question. In computational terms, we might think of such explanations as solving the reasoner's problem of how to select or represent the evidence. The problem still remains of weighing the evidence in favour of the target conclusion – namely, does the target of induction have the relevant feature in question. Depending on the nature of the property in question, the answer may not be immediately obvious and may also involve further theory-guided generalisation.

E. E. Smith, Shafir, and Osherson (1993), introduced the gap theory of induction to account for ways that people use background knowledge to reason about *threshold* features – something which was not captured in their earlier similarity-coverage model (Osherson et al., 1990). The essence of the theory is that argument strength reflects both similarity between premise and conclusion categories as well as the plausibility of premise and conclusion. Further, the theory suggests how these two aspects trade-off. When there is significant uncertainty regarding the relevant basis for projection, as when blank properties are used, then reasoners should rely heavily on similarity. Conversely, in examples like the following:

$$\frac{\text{Dobermans can bite through wire.}}{\text{German shepherds can bite through wire.}} \tag{1.6a}$$

$$\frac{\text{Poodles can bite through wire.}}{\text{German shepherds can bite through wire.}} \tag{1.6b}$$

where background knowledge or theory suggests a clear explanation for the property in question (sufficient size and strength in this case), then the structure of the corresponding feature space becomes relevant. Thus, in the above example, people tend to rate the second argument as stronger than the first despite the fact that German shepherds bear a closer resemblance to Dobermans than to Poodles. According to gap theory, it is the prior implausibility or "surprise value" of the premise relative to the conclusion that adds weight to the conclusion. It does so by prompting the reasoner to update her criterion about the property in question (in this case, the strength required to bite through wire).

[11] Although an hour spent viewing social and news media in 2019 might easily convince you otherwise.

Sloman's findings and those of E. E. Smith et al. (1993) highlight the role that people's background knowledge and theories play in property induction, lending support to the idea that people strive for explanatory coherence when interpreting evidence. Together with similarity-based accounts (such as Osherson et al., 1990; Sloman, 1993, for example), the work paints a rich picture of the variety of evidence that people recruit in order to close the gap between premise and conclusion. But what is missing from the picture is an account of how these different kinds of evidence may be combined and weighed in the balance. In the next section, I look at how the mathematics of Bayesian inference can be used to describe inductive generalisation where the reasoner draws on multiples sources of evidence.

1.4 Sampling matters

My principal aim in this section is to convey a sense of why sampling assumptions *should* matter, and to outline some empirical evidence which suggests that they probably do. In support of the first point, I begin by discussing the theoretical foundations of Bayesian inference and the model of Bayesian generalisation that it supports. It is within this framework that clear quantitative predictions emerge regarding the effect of sampling assumptions on the interpretation of data. In essence, it is an investigation of the further consequences of these core predictions that drives the new research I present in subsequent chapters.

The Bayesian framework for weighing data as evidence

Thus far I have considered three different sources of evidence that the reasoner might employ to fuel an inductive inference. Firstly, there is evidence drawn from inter-stimulus similarity based on direct comparison. The weight of evidence in this case is thought to relate to some measure of psychological distance (e.g., Shepard, 1987) or featural overlap (e.g., Tversky, 1977) between the data and the target of induction. Secondly, there is the kind of evidence reflected in concept representations involving distributional information over collections of abstract features (e.g., Rosch & Mervis, 1975). Thirdly, there is the evidence drawn from explicit prior knowledge (e.g., Heit, 1998), explanations (e.g., Sloman, 1994), or intuitive theories which may be abstract (e.g. N. Goodman's, 1955, *overhypotheses*), domain-specific (e.g., Heit & Rubinstein, 1994) or property-specific (e.g., E. E. Smith et al., 1993). The experiments I describe in this thesis examine how people interpret each of these forms of evidence in light of their sampling assumptions.

Many of the accounts of inductive generalisation that I have so far discussed, focus on single source of evidence to explain some aspect of inductive inference. The Bayesian approach in contrast, allows for the specification of models of inductive generalisation where the reasoner draws upon and integrates any such evidence. Bayesian models have been successively applied to inductive inference in a variety of task settings including stimulus generalisation (Shepard, 1987), category learning (Anderson, 1991), property induction (Heit, 1998; Kemp & Tenenbaum, 2009; Sanjana & Tenenbaum, 2003), word and language learning (Xu & Tenenbaum, 2007b), pedagogical teaching and learning (Shafto, Goodman, & Griffiths, 2014), and more.

Bayesian inference represents a method of belief updating consistent with the axioms of probability (see Jaynes, 2003, for example). Because of this axiomatic grounding, it confers certain desirable properties with respect to evidence calibration. For example, it allows the reasoner to avoid a so-called *Dutch book* – a series of gambles that reveals the reasoner's probabilistic beliefs to be mutually inconsistent (Horwich, 2016, p. 18). As such, Bayesian inference has been used as the basis for normative or "rational" models of cognition (e.g., Anderson, 1991; M. C. Frank & Goodman, 2012; Fried & Holyoak, 1984; Oaksford & Chater, 1998; Shafto et al., 2014; Shepard, 1987). But as Tauber, Navarro, Perfors, and Steyvers (2017) point out, setting optimality aside, the Bayesian framework has equally important application as a descriptive tool for capturing people's prior biases and sampling assumptions and testing particular cognitive theories. It is in this descriptive sense that the Bayesian models put forward in Chapters 5 and 6 are intended.

Bayesian models of inductive inference reflect a model-centric view of reasoning. Like the theory-centric view of induction and concept formation discussed earlier (*viz.* Murphy & Medin, 1985; Quine, 1969), the model-centric view changes the conception of the reasoner's task from one of comparison to one of explanation. Instead of directly comparing the source and target of induction to determine whether they are "similar enough" to warrant the induction at hand, the reasoner compares alternative explanations for why the induction is (or isn't) valid. Under the model-centric view, the reasoner forms an abstract representation (a model) of some relevant aspect of the world encompassing a number of hypothetical possibilites (*hypotheses*). The range of hypotheses considered is typically constrained by the reasoner's theories, overhypotheses or structured background knowledge (Kemp, Perfors, & Tenenbaum, 2007; Kemp & Tenenbaum, 2009). Particular hypotheses are validated or otherwise on the basis of new observations.

To see how the Bayesian inference framework may be used in practice, consider the following inductive argument:

$$\text{QUAIL EGGS} \xrightarrow{\textit{has property } C} \text{GOOSE EGGS}$$

The reasoner's model in this instance revolves around her ideas about which eggs have some novel blank property C. In Bayesian terms, she considers a set of hypotheses (more generally, an *hypothesis space*) $\mathcal{H}$, where an individual hypothesis $h \in \mathcal{H}$ represents one partitioning of the category members (EGGS, in this case) according to whether or not they exhibit the property in question. If there are N types of egg in her representation of the category, there are potentially 2^N hypotheses that she needs to consider, as Figure 1.2

illustrates. In the example shown, the reasoner is aware of three different kinds of egg, and so her hypothesis space consists of $2^3 = 8$ hypotheses (see panel a).

Crucially, the Bayesian reasoner updates the strength of her belief in each hypothesis h (represented by a probability $P(h) \in [0,1]$) on the basis of the data she encounters. She does this according to Bayes' theorem:

$$P(h|x) = \frac{P(x|h)P(h)}{\sum_{h' \in \mathcal{H}} P(x|h')P(h')}, \tag{1.7}$$

which relates her *posterior* belief $P(h|x)$ after seeing the data x to her *prior* belief $P(h)$ and her *likelihood function* $P(x|h)$.

In Bayesian computational models, the prior serves to specify the kinds of background information that the reasoner draws on in addition to the evidence of direct observations. If the reasoner has a particular theory or overhypotheses concerning the property to be generalised or the relevant domain, then that theory may be captured in the prior. Returning to the above example, consider the property *boils under 8 minutes* in place of the blank property C. This is the kind of property for which background knowledge might reasonably be brought to bear. Say, for example, that the reasoner believes that bigger eggs take longer to boil and can thus order all $N = 3$ types of EGG on the basis of size. In this case, there are only $N + 1 = 4$ hypotheses she need consider (see panel b, Figure 1.2). What amounts to deductively invalid possibilities according to the rule like implications of her background theory are represented as zero (or negligible) prior belief. In this way, the prior acts to constrain the (prior) representation of the reasoner's hypotheses space $\mathcal{H}$, and significantly reduce the number of possibilities that she considers. Bayesian models have used the prior to capture the evidence of similarity in hierarchically structured categories (Sanjana & Tenenbaum, 2003), as well as taxonomic, spatial and causal relationships amongst related categories (Kemp & Tenenbaum, 2009).

The likelihood function also serves a central role in Bayesian inference by describing how the evidence in the data bears on the reasoner's beliefs. Critically, it supports disconfirmation of hypotheses by capturing the reasoner's sense of which hypotheses are compatible with the data, and which are not. In our example, the observation that QUAIL EGGS boil in under eight minutes ($x = +$QUAIL EGGS) would invalidate the hypothesis h_0 that none of the known eggs do, as reflected in the likelihood $P(x = +\text{QUAIL EGGS}|h_0) = 0$ – see Figure 1.2 (panel c).

The second role of the likelihood function is to capture the relative strength of evidence that the data represents under each hypothesis with which it is compatible. For example, according to the likelihood function shown in Figure 1.2 (panel d), the observation $x = +$QUAIL EGGS represents stronger evidence for the hypotheses that only quail eggs boil in under eight minutes (h_1) than it does for the hypotheses that all known eggs do (h_7), by a ratio of 3 : 1. It is in this second sense that the likelihood function plays a central role in the computational analyses described throughout this thesis, supporting a quantitative assessment of the impact of various assumptions on the outcome of inference.

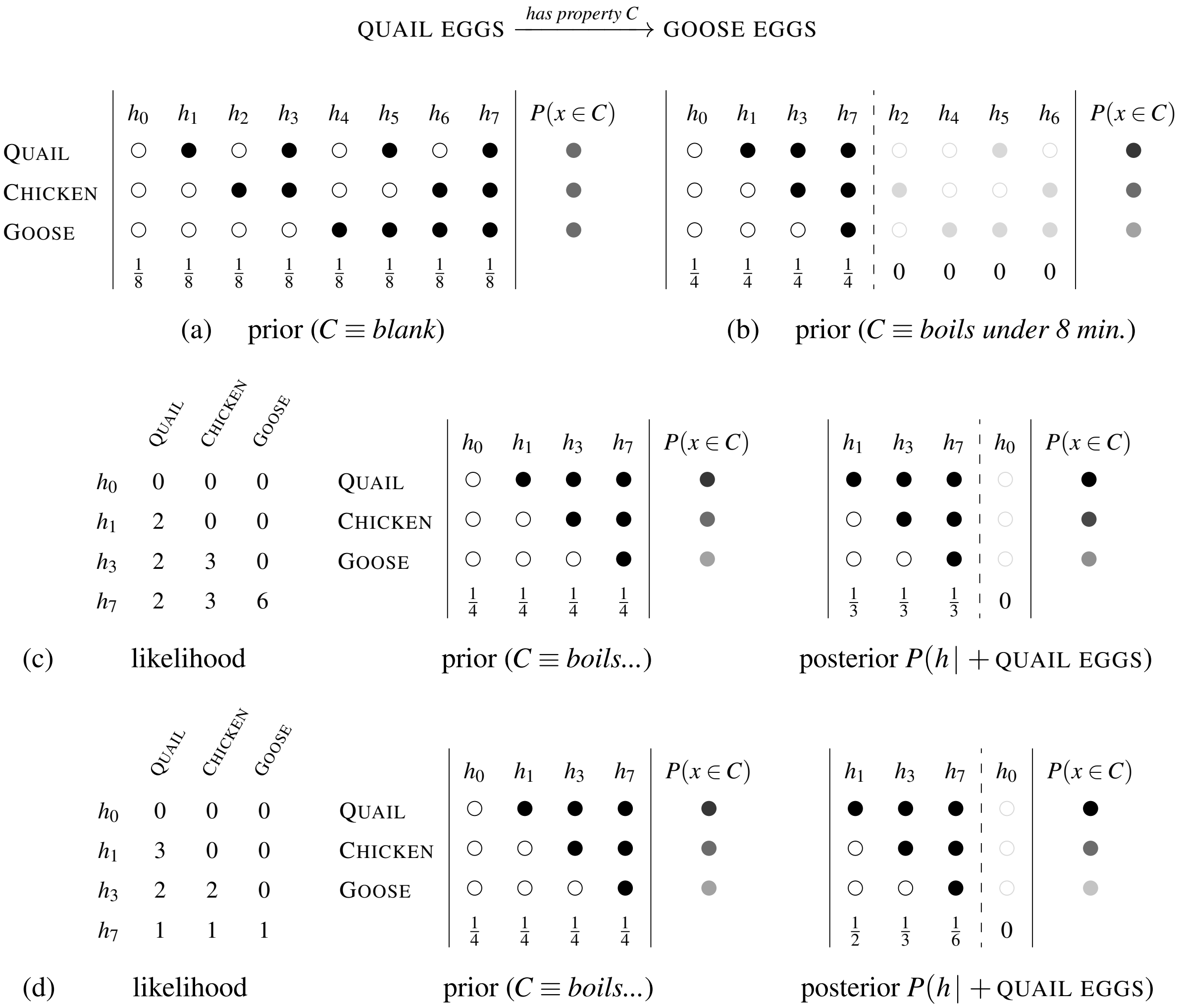

QUAIL EGGS $\xrightarrow{\text{has property } C}$ GOOSE EGGS
h_0 h_1 h_2 h_3 h_4 h_5 h_6 h_7 $P(x \in C)$
QUAIL
CHICKEN
GOOSE
$\frac{1}{8}$ $\frac{1}{8}$ $\frac{1}{8}$ $\frac{1}{8}$ $\frac{1}{8}$ $\frac{1}{8}$ $\frac{1}{8}$ $\frac{1}{8}$
(a) prior ($C \equiv$ *blank*)
h_0 h_1 h_3 h_7 h_2 h_4 h_5 h_6 $P(x \in C)$
$\frac{1}{4}$ $\frac{1}{4}$ $\frac{1}{4}$ $\frac{1}{4}$ 0 0 0 0
(b) prior ($C \equiv$ *boils under 8 min.*)
QUAIL CHICKEN GOOSE
h_0 0 0 0
h_1 2 0 0
h_3 2 3 0
h_7 2 3 6
h_0 h_1 h_3 h_7 $P(x \in C)$
$\frac{1}{4}$ $\frac{1}{4}$ $\frac{1}{4}$ $\frac{1}{4}$
h_1 h_3 h_7 h_0 $P(x \in C)$
$\frac{1}{3}$ $\frac{1}{3}$ $\frac{1}{3}$ 0
(c) likelihood
prior ($C \equiv$ *boils...*)
posterior $P(h \mid +$QUAIL EGGS$)$
QUAIL CHICKEN GOOSE
h_0 0 0 0
h_1 3 0 0
h_3 2 2 0
h_7 1 1 1
h_0 h_1 h_3 h_7 $P(x \in C)$
$\frac{1}{4}$ $\frac{1}{4}$ $\frac{1}{4}$ $\frac{1}{4}$
h_1 h_3 h_7 h_0 $P(x \in C)$
$\frac{1}{2}$ $\frac{1}{3}$ $\frac{1}{6}$ 0
(d) likelihood
prior ($C \equiv$ *boils...*)
posterior $P(h \mid +$QUAIL EGGS$)$

Figure 1.2: The panels illustrate how reasoning about a given inductive argument (top row) may be captured as Bayesian inference. [**panel a**] If the reasoner knows three different kinds of egg, then prior to seeing the data (the argument premise in this case), there are eight hypotheses that she might consider regarding the set of eggs with the given property (black dots). In the absence of any relevant background information, her prior belief may be uniformly distributed across the hypotheses: $P(h) = \frac{1}{8}$ (bottom row of panel). The prior probability that each item has the property in question $P(x \in C)$, is uniformly even as a result (grey dots). [**panel b**] The rule like implications of a property-specific theory may simplify the hypotheses space by assigning zero probability to certain hypotheses *a priori*. [**panel c**] The Bayesian likelihood (left) captures those observations which are compatible with each hypothesis (non-zero values indicate compatibility), allowing the reasoner to update her beliefs from her prior (middle) to her posterior (right), after seeing the data (x = +QUAIL EGGS). [**panel d**] The likelihood also captures the evidentiary weight that an observation represents for each hypothesis with which it is compatible. When these values differ (left), posterior belief may be unevenly redistributed across the compatible hypotheses (compare posteriors in panels c and d).

The example I have used to illustrate the workings of Bayesian inference has involved discrete observations and a finite hypothesis space with set-valued hypotheses. Nonetheless, when reasoning about evidence represented in continuous spaces, the same principles apply. The reasoner entertains hypotheses (albeit infinitely many) and updates her prior belief on the basis of observations (of continuously varying quantities) in accordance with a likelihood function (expressed over a continuous domain). Because the hypothesis space reflects a continuum rather than a discrete set, the summation in Bayes' theorem (Equation 1.7) becomes integration in the limit.

Having sketched the basics of Bayesian inference, I now want to describe two important milestones in the development of a Bayesian theory of generalisation, each of which is built around a different likelihood function. Between them, these two theories, and the two sampling assumptions they embody are central to the new research I describe throughout this thesis.

Weak sampling and a strong prior bias

Shepard (1987), used a Bayesian analysis in an effort to uncover a universal law that governs generalisation across a broad range of inductive problems in humans and might even operate across species. The theory he developed had a strong connection to his earlier work (Shepard, 1957) seeking to understand stimulus and response confusion, but reflected a broader outlook. Shepard (1987) conceived of generalisation, not as a failure of discrimination, but as a cognitive act, noting that we generalise between situations "because we judge that they are likely to belong to a set of situations having the same consequence" (Shepard, 1987, p. 1322).

In an appeal to evolutionary theory, Shepard argued that stimuli important to an individual are not one of a kind, but members of some natural class which he linked to a particular mental representation: a convex *consequential region* in psychological space. Viewed in this way, the problem of generalising from one stimulus to another becomes a form of ad hoc concept generalisation: i.e. assuming the original stimulus is an example of some latent concept, the reasoner attempts to decide whether the target stimulus would also be an example of that concept. In computational terms, once a learner has observed that one stimuli is consequential (predicts a desirable outcome, for example) she can calculate the probability that the concept should generalise to a novel stimuli by considering consequential regions of various size, shape and position as alternative hypotheses, and averaging over them in a Bayesian fashion.

Central to Shepard's analysis, is the idea that the learner makes an assumption about the likely size of the consequential region. In the uni-dimensional case, he showed that if the learner assumes only that the region has some finite size on average, but otherwise adopts the principal of maximum ignorance[12], then the resulting generalisation gradients

[12] The maximum entropy prior in this case is an Erlang distribution with finite scale. Without this stipulation of finite scale, the generalisation gradient would degenerate to linear form.

predicted by Bayes' rule follow the pattern of exponential decay exhibited in a wide range of empirical studies (see Shepard, 1987, p. 1318, for a comprehensive list). Further, he showed that the prediction of an (approximately) exponential pattern of decay is robust with respect to considerable variation in the prior distribution. By aligning the idea of stimulus similarity to proximity in psychological space, and linking concept representation to consequential regions within that space, Shepard demonstrated how a Bayesian framework could incorporate the evidence of similarity and capture the pervasive intuition that similar things have similar consequences.

Shepard's analysis focused on the problem of generalising on the basis of a single stimulus. When there is only a single data point to learn from, the generative model underlying the data is less important than it might otherwise be. In Shepard's model, the learner observes that a single stimulus has some consequence, but assumes that the two were sampled *independently* of one another. The fact that the stimulus and the consequence coincide is seen as nothing more than that – a *coincidence*. In Bayesian terms this *weak sampling* assumption is captured by a likelihood function of the form:

$$P(x|h) = \begin{cases} P(x) & \text{if } x \in h \quad \text{(i.e. } x \text{ is compatible with } h\text{)} \\ 0 & \text{otherwise} \end{cases} \tag{1.8}$$

where x corresponds to the stimulus observed, and h represents an hypothesis of interest (about the consequential region, in Shepard's case). Beyond discriminating those hypotheses that are consistent with the data from those which are not, the likelihood function in this case is silent. However likely or unlikely the observation of x may be, it represents the same weight of evidence for any consistent hypothesis. We can see this mathematically, through a rearrangement of Equations 1.7 and 1.8:

$$\frac{P(h_1|x)}{P(h_2|x)} = \frac{P(x|h_1)}{P(x|h_2)} \times \frac{P(h_1)}{P(h_2)} = \frac{P(x)}{P(x)} \times \frac{P(h_1)}{P(h_2)} = \frac{P(h_1)}{P(h_2)} \tag{1.9}$$

which shows that for any two hypotheses h_1 and h_2 compatible with the data x, the relative plausibility (or *odds ratio*) of the two reamins unaltered after seeing the evidence. If instead $\mathbf{x}$ represents a number of observations, then Equation 1.9 still holds.

When stimuli are represented in a continuous metric space (as in Shepard, 1987), new observations outside the range of previously observed examples will still broaden the region of generalisation, but the shape of the curve that dictates how far people reason beyond that range will be largely unchanged.[13] Importantly, additional observations within the previously observed range – those which are in effect, certain examples of the concept in question – these will have no effect on generalisation. Where the notion of a "generalisation curve" no longer directly applies, when observations are represented in set theoretic terms for example (as in Heit, 1998), the same underlying principles hold:

[13] Any change that there is comes about either as a result of boundary conditions which induce a form of linear rescaling along the axes of generalisation, or is contingent upon a non-uniform prior.

observations represent uniform evidence for all consistent hypothesis, and highly similar or repeated observations will have little or no effect. The upshot of all of this is that under a weak sampling assumption, the work in shaping generalisation *beyond* the data is done largely by the prior.[14] Despite this potential limitation, weak sampling has been successively used in Bayesian models to capture a variety of inductive phenomena (e.g., Heit, 1998; Kemp & Tenenbaum, 2009; Shepard, 1987).

Strong sampling and the size principle

Tenenbaum and Griffiths (2001a) built on Shepard's analyses in framing their Bayesian model of concept generalisation (see also Tenenbaum, 1999). The model defines the probability of generalising a concept C to a novel item y as an average calculated over each of the hypothetical forms that the concept might take. That is:

$$P(y \in C \,|\, \mathbf{x}) = \sum_{h:y\in h} P(h \,|\, \mathbf{x}) \tag{1.10}$$

where $P(h \,|\, \mathbf{x})$ is the probability of a given hypothesis after learning from the observations $\mathbf{x}$, per Equation 1.7.

Tenenbaum and Griffiths (2001a) capture the idea that people may assume that exemplars are sampled from the concept of interest, thereby making a *strong sampling* assumption. In its simplest form, strong sampling implies that the probability of seeing a particular example consistent with some hypothesis is proportional to the size of that hypothesis. By extension, if $\mathbf{x}$ represents a collection of n independently sampled observations, then the size of the hypothesis is raised to the nth power. Mathematically, strong sampling may be captured by the following likelihood function

$$P(\mathbf{x} \,|\, h) = \begin{cases} \frac{1}{|h|^n} & \text{if } x_1, x_2, \ldots, x_n \in h \\ 0 & \text{otherwise} \end{cases} \tag{1.11}$$

where $|h|$ denotes the *size* of hypothesis h.[15] As a comparison of Equations 1.8 and 1.11 makes clear, the likelihood function serves the same role in falsifying hypotheses under both strong and weak sampling. What strong sampling adds however, is a way of weighing

[14] Were it indeed universal, weak sampling together with Shepard's suggestion of "evolutionary internalisation" would suggest a certain irony if taken to extremes – namely, that while the shape of generalisation gradients may be sensitive to and driven by data gathered over evolutionary timescales, it would nevertheless be insensitive to change in response to further observations made by the individual.

[15] Precisely what the notion of hypothesis size maps onto depends upon the nature of the hypothesis space in question. In the case where alternative hypotheses represent a discrete set of examples (animals that can fly, for example), hypothesis size is equivalent to the number of examples consistent with the hypothesis. Where hypotheses represent continuously varying quantities that are "everywhere dense" (such as the perceptual dimensions of size and color), the size of an hypothesis represents its *measure* which corresponds to the notions of length, area or volume in Euclidean space (depending on the dimensionality of the quantity concerned).

those remaining hypotheses that are consistent with the data. As before, we can use Equations 1.7 and 1.11 to see how the relative plausibility of two consistent hypotheses changes after viewing the data:

$$\frac{P(h_1\,|\,\mathbf{x})}{P(h_2\,|\,\mathbf{x})} = \left(\frac{|h_2|}{|h_1|}\right)^n \times \frac{P(h_1)}{P(h_2)} \tag{1.12}$$

This shows that under a strong sampling assumption Bayesian inference embodies a *size principle*: the evidence of **x** favours the *smaller* of the two hypotheses. Just how this effects belief overall will also depend on the prior. Under a prior that already embodies a size principle (like the Erlang distribution that Shepard (1987) suggests), the smaller of the two hypotheses will be favoured from the start and will become increasingly favoured with each additional (consistent) observation.

The size principle matters

Tenenbaum and Griffiths (2001a) argue that the size principle, which emerges under a strong sampling assumption, may be a fundamental force driving similarity comparisons. In section 1.3, I reviewed a number of works which emphasise the importance of similarity in driving generalisation (e.g., Osherson et al., 1990; Quine, 1969; Rips, 1975; Sloman, 1993; E. E. Smith et al., 1993). The alternative suggestion that judging similarity involves making generalisations (Tenenbaum & Griffiths, 2001a), is thus an intriguing one.

To illustrate, consider that a reasoner is presented with some object x (by an experimenter, for example) and that a second object y is introduced calling for a comparison. Further, suppose that the reasoner makes a strong sampling assumption about the original object – namely, that x is sampled from some consequential set C say.[16] From a generalisation perspective, the reasoner's task in generalising from the first object x to the novel object y amounts to weighing various hypotheses about a consequential set that contains both items, against alternative proposals containing x only. The link between similarity judgments (feature weighted comparison) and generalisation (hypothesis averaging) emerges if we consider that the reasoner's hypothesis space used to represent consequential sets corresponds to a set of abstract *features* (whatever they may be). If the reasoner's hypothesis space is represented in this way, then individual hypotheses correspond to individual features, and an object belongs to some hypothesised set h if it has the corresponding feature. In this way, hypothesis weights and feature weights are equivalent, and small heavily weighted hypotheses correspond to small (specific) heavily weighted features.

[16] The implicit assumption here is that the second "test" object is weakly sampled, but other assumptions are certainly not unreasonable. It seems more likely for example, that having selected one cocker spaniel for me to consider you might select another as the basis for comparison, rather than say one of Mozart's arias. Nonetheless, pragmatic assumptions made by participants regarding the sampling of experimental test items are typically ignored in the literature (which is somewhat justified when exhaustive pairwise similarity comparisons are involved).

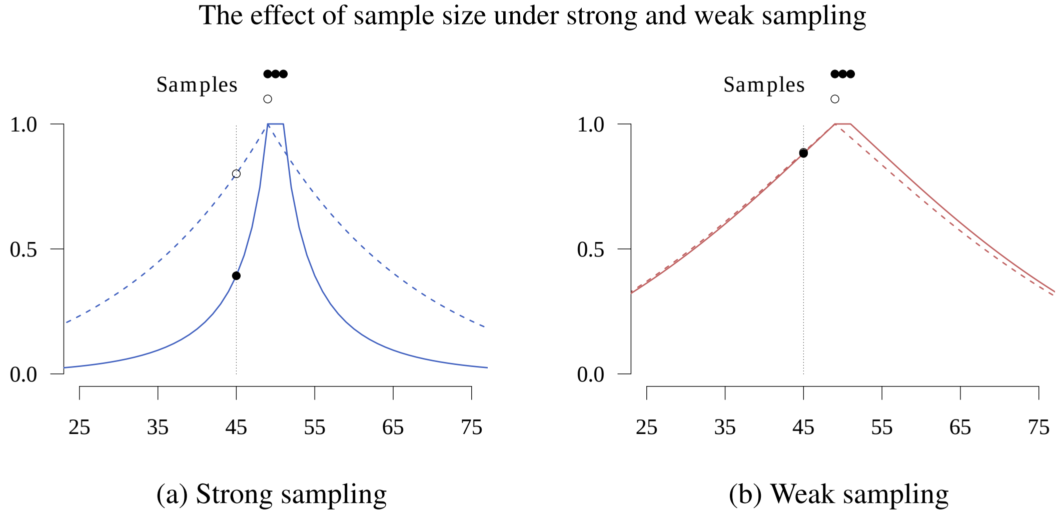
Reasoning along these lines, Tenenbaum and Griffiths (2001a) demonstrate that the Bayesian generalisation model (Equation 1.7) subsumes a version of Tversky's (1977) contrast model (Equation 1.2). And at a deeper level, their analysis shows that the size principle offers a principled reason for why more specific features should be weighted more heavily in featural comparisons.

Navarro and Perfors (2010) present an interesting counterpoint to Tenenbaum and Griffiths's (2001a) analysis, arguing that a size principle can emerge as a consequence of representational efficiency. That is, if the learner seeks to carve up the similarity structure in the environment according to a set of coherent features, then smaller features will be preferred. The argument goes something like this. Assume there is a pool of candidate features, varying in size (number ot items indexed) and coherence (average pair-wise similarity among the items that share it). Measured across larger features, coherence is more likely to approach the limit –"average" coherence. Smaller features in contrast, should have greater variability in coherence on both the high side and low side. Optimising for high coherence would thus tend to select for smaller features, while weeding out features with low coherence would mean that the smaller features that were retained would tend to be the more coherent. Either way, the argument demonstrates that a size principle governing featural similarity may emerge from optimal encoding principles, rather than a strong sampling assumption.

What does strong sampling predict that weak sampling doesn't?

I have been attempting to show why, according to the principles of Bayesian inference at least, sampling assumptions play an important role in determining the outcome of inductive reasoning. I have discussed the importance of strong sampling and the size principle in light of Tenenbaum and Griffiths's (2001a) analysis of featural similarity. Yet, I have also illustrated what can be achieved under a weak sampling assumption alone, if a size principle is built into the prior (as in Shepard, 1987) or is derived from principles of efficient representation (per Navarro & Perfors, 2010). This begs the question: where should strong sampling make a difference in a way that can't be accounted for under weak sampling and a suitably strong prior?

To illustrate the predictions of the Bayesian model in this regard, let us revisit Alice. Suppose Alice wants to avoid using eggs weighing less than 45g. Receiving eggs from her supplier, she weighs the first egg at 49g. Subsequently, she weighs three in all, finding weights of 49g, 51g, and 50g, respectively. Regarding the possibility of an underweight egg ($\leq$ 45g), how might Alice revise her opinion after making the additional measurements? Under a weak sampling assumption, she should be none the wiser. Yet this seems counterintuitive. In this example based on a single continuously varying property, the larger sample seems more informative – it now seems less likely that she might find an underweight egg. As Figure 1.3 illustrates, this kind of tightening of generalisation on the basis of increased sample size is precisely what the Bayesian model predicts under



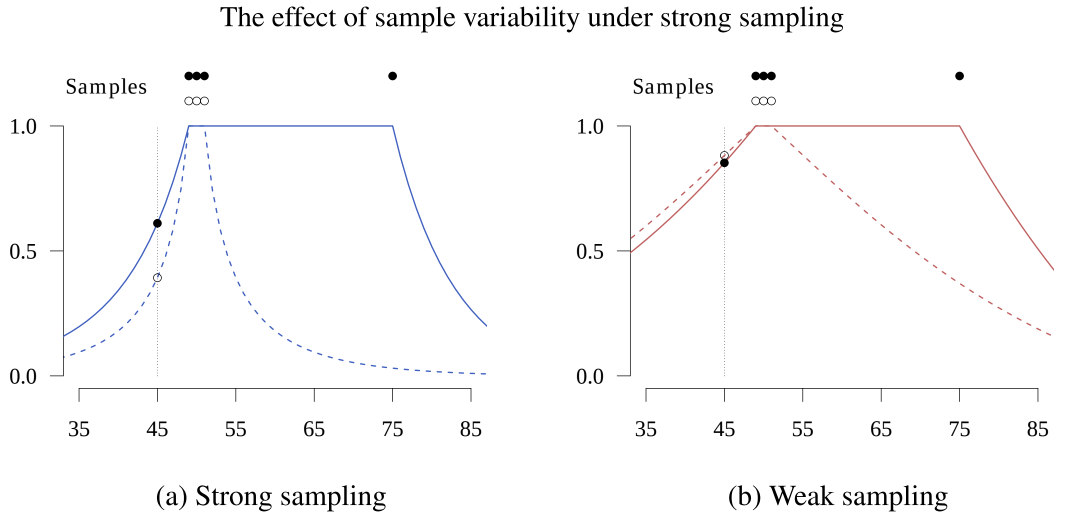
(a) Strong sampling

(b) Weak sampling

Figure 1.3: Simulated concept generalisation as a function of sampling assumption and sample size. The graphs plot the probability of generalising the concept in question (egg weights in this example) after seeing a single example (dashed lines) and three examples (solid lines). (a) Under a strong sampling assumption, the additional examples lead to a considerable tightening of generalisation around the range of egg weights observed. As a consequence the learner may considerably revise her opinion about the question of interest – the possibility of eggs weighing 45g or less – from relatively likely (open circle) to less likely (solid circle). (b) In contrast, under a weak sampling assumption, the same additional examples have little effect.

a strong sampling assumption, and has been demonstrated across a variety of inductive reasoning tasks (e.g., M. Frank & Tenenbaum, 2011; Hayes, Banner, Forrester, & Navarro, 2019; Hendrickson, Perfors, Navarro, & Ransom, 2019; Hsu & Griffiths, 2016; Lewis & Frank, 2016; Navarro, Dry, & Lee, 2012; Tenenbaum, 1999, 2000; Vong, Hendrickson, Perfors, & Navarro, 2013; Xu & Tenenbaum, 2007a, 2007b).

In a similar vein, imagine that Alice now weighs a fourth egg, this time finding a large one (75g). Now her sample consists of: 49g, 51g, 50g, and 75g eggs. Should she conclude anything differently, having discovered the large egg? Once again, a weak sampling assumption implies that she should not change her mind with respect to underweight eggs, yet once again this feels counterintuitive. Given there is now high variability in her sample, the possibility of a 45g egg seems different in the two cases. Indeed, as Figure 1.4 illustrates, the discovery of a large egg may make her more likely to expect a small one. Related effects of sample variability (diversity) have been noted in the literature (e.g., Fried & Holyoak, 1984; Hayes, Navarro, et al., 2019; Osherson et al., 1990). A core assumption of the *category density model* (Fried & Holyoak, 1984), which can also account for such effects, is that learners acquire a schematic representation of the distribution of exemplars along feature dimensions of interest. Other models (e.g., Osherson et al., 1990), while less explicit about how distributional information is acquired, nonetheless invoke the idea in accounting for effects of sample diversity. What



(a) Strong sampling

(b) Weak sampling

Figure 1.4: Simulated concept generalisation as a function of sampling assumption and sample variability. The graphs plot the probability of generalising the concept in question (egg weights in this example) after seeing three non-diverse examples (dashed lines) and after an additional diverse example is observed (solid lines). (a) The observation of a 75g egg in this example, by increasing the variability of the sample, has the effect of broadening generalisation. As a consequence the learner may considerably revise her opinion about the question of interest – the possibility of eggs weighing 45g or less – from somewhat unlikely (open circle) to more likely (solid circle). But only under a strong sampling assumption. (b) Under a weak sampling assumption, the change in variability of the sample is uninformative regarding the question of interest.

these models share is an implicit assumption which diverges from the weak sampling assumption in a fundamental way. When reasoning about some matter of interest (e.g. underweight eggs) on the basis of distributional information, the assumption is that the sample distribution learned (e.g. sample egg weights) is representative of the concept of interest (e.g. range of egg weights) – in essence, a form of strong sampling assumption.

The two cases I have just illustrated represent core predictions of the Bayesian generalisation model highlighting the difference between strong and weak sampling. In sum, while the model predicts that the *location* of the region of generalisation will adapt to encompass the data equally under both strong and weak sampling, the prediction regarding the *gradient* of generalisation differs. Under weak sampling, gradients are dominated by the prior. Under strong sampling generalisation can more flexibly adapt to data. In the absence of negative evidence, additional positive examples of a concept can lead to a tightening of generalisation gradients causing the region of generalisation to contract. Conversely, a generalisation gradient can expand in a given direction without further positive examples in that direction, by increasing sample diversity overall.

Empirical tests of the strong versus weak distinction

Following the formulation of strong and weak sampling within a Bayesian generalisation model (Shepard, 1987; Tenenbaum & Griffiths, 2001a), there have been numerous demonstrations in the literature of the differing effects of these assumptions on inductive reasoning (e.g., Fernbach, 2006; Hayes, Banner, et al., 2019; Hayes, Banner, & Navarro, 2017; Hayes, Navarro, et al., 2019; Lawson & Kalish, 2009; Navarro et al., 2012; Tenenbaum, 2000; Vong et al., 2013; Voorspoels, Navarro, Perfors, Ransom, & Storms, 2015). A common experimental approach has involved attempts to manipulate people's sampling assumptions between experimental conditions. By presenting people the same sample of training data but varying the explanation for how it was produced, it is possible to observe any effects of sampling assumptions in the results of subsequent generalisation tests. Such manipulations have typically involved a cover story that describes the way that the training samples were collected. Each of the experiments described in the following chapters employ this technique. Other related studies, notably those demonstrating sampling sensitivity in infants (Gweon, Tenenbaum, & Schulz, 2010) and pre-school children (Rhodes, Gelman, & Brickman, 2010; Xu & Tenenbaum, 2007a), have used more explicit manipulations involving an experimenter actively sampling data for participants and giving ostensive cues regarding the method of selection.

Although the point is an obvious one, it is worth emphasising that there is a difference between the sampling assumption that an experimental manipulation emphasises and the one that the reasoner adopts. Setting aside the usual scope for individual variation, there is room for widespread divergence as well. Even seemingly explicit suggestions, such as "these examples were sampled by a helpful teacher from [the category]", may not have the effect intended. For this reason, some care must be taken in interpreting the results of such studies (and indeed, the new research I describe in this thesis). Nonetheless, the pattern of results exhibited across these studies is broadly consistent with the predictions of the Bayesian framework under an assumption of strong or weak sampling, as appropriate for the experimental condition.

Conservative sampling assumptions

In a seminal study, Navarro et al. (2012) found considerable individual differences regarding the extent to which people adopt a strong sampling assumption even when a cover story manipulation was highly suggestive that one was appropriate. Their findings raise an interesting question regarding the source of such differences. There are of course numerous differences in cognitive capacities that might be posited in explanation. For example, the findings I discuss in Chapter 4 suggest that patterns of generalisation may be less responsive to additional exemplars (less in accordance with the predictions of strong sampling) when people must recall exemplars from memory – a cognitive capacity where reasonable individual variation is to be expected. Navarro et al. (2012) suggest that a form of inferential conservatism may be involved. This suggestion is intruiging because it immediately begs the question concerning the reason for such a stance. Two distinct but related possibilities come to mind, both of which are relevant to the investigations described in this thesis.

The first possibility, which might account for the prevalence of conservatism (but less so the variation among individuals) is simply that a strong assumption may frequently be unjustified. In a tautological sense, a positive example of a given concept can always be viewed as if it were a draw from the concept of interest. When viewed in that way, the question facing the reasoner is whether the draw was a fair one, or whether instead the generative process behind the sample has introduced any systematic bias. There are numerous ways in which a sampling process might introduce bias. For example, exemplars of a given concept might vary by location (consider antipodean swans, for example) or over time (e.g. telephones) in ways that effectively censor the full distribution. Even when there is no obvious censoring that might impose outright restrictions on the availability of particular exemplars, learning from a non-uniform exemplar distribution may give a distorted view of data if the learner assumes that exemplars are uniformly distributed (per strong sampling). Somewhat related is the distinction between *types* and *tokens*. If, as is often the case, the learner is interested in learning the range of exemplars that form a given concept (such as breeds of dog, for example), then learning about a new type (e.g. a *"cavoodle"*) is useful, while observing additional tokens of the same type (seeing the same dog numerous times, for instance) may be uninformative (e.g., Perfors, Ransom, & Navarro, 2014; Xie, Hayes, & Navarro, 2018).

As Navarro et al. (2012) suggest, a weak sampling assumption is equivalent to a low-variance, high-bias estimator, while a strong sampling assumption is low-bias but high-variance.[17] Whether in the long term weak sampling leads to systematic underfitting or strong sampling leads to systematic overfitting of data, will depend on the frequency with which the assumptions of strong sampling hold. Short of actually knowing how data is sampled, there can be no normative stance regarding the appropriate behaviour. As Navarro et al. (2012) suggest, learners might reasonably adopt a sampling assumption that varies in strength anywhere along a continuum between weak and strong sampling. To capture this in computational terms, they introduce the *mixed sampling* likelihood:

$$P(x\,|\,h) = \begin{cases} \theta\frac{1}{|h|} + (1-\theta)\frac{1}{|\mathcal{X}|} & \text{if } x \in h \\ 0 & \text{otherwise} \end{cases} \tag{1.13}$$

where $|\mathcal{X}|$ denotes the number of items that might possibly be observed, and θ reflects the probability that the premise item x was strongly sampled. Setting $\theta = 0$ yields weak sampling as a special case, while $\theta = 1$ is equivalent to strong sampling. A plausible interpretation of such a mixed assumption is that it reflects a conservative disposition towards the risks and rewards of a stronger sampling assumption, even when one might reasonably hold. Conservatism of this form might thus be more a longer term strategic

[17] See (Hastie, Tibshirani, & Friedman, 2009) for a statistical discussion of the bias-variance trade-off

outlook, and less a response to the specific context in which data is observed. The work of Gigerenzer and colleagues (e.g., Gigerenzer & Brighton, 2009) supports the idea that more robust reasoning may emerge from a "less is more" approach, where aspects of the data are ignored in favour of greater efficiency and accuracy in the long-term.

In Chapter 6, I describe an experiment where people play a strategic communication game, and must reason from evidence provided by a teammate or an opponent. I examine whether people adopt a conservative stance regarding the way they reason from messages or whether instead they adopt assumptions based on the context in which communication takes place as well as the message content itself.

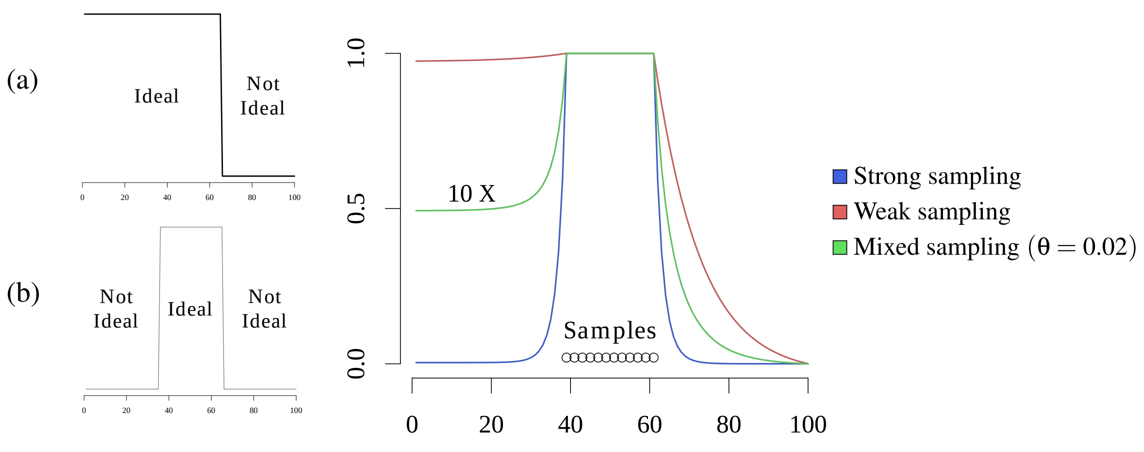
The pros and cons of stronger sampling assumptions

Despite the potential benefits of exercising a cautious inferential stance, and the inherent uncertainty surrounding the generative process, people nonetheless attempt to leverage limited data by making strong assumptions. Indeed in some settings – *pedagogical* settings, for example, where data is provided by a trusted and knowledgable teacher – people interpret the data in excess of what the $\frac{1}{|h|^n}$ size principle would suggest (Shafto et al., 2014). So taking for granted a situation where strong sampling (or something like it) is a reasonable description of the generative process, how does a stronger sampling assumption pay off in that case?

So far, I have considered the implications of a strong sampling assumption from a computational perspective. The key advantage of a stronger sampling assumption over a more conservative approach is that generalisation gradients can better adapt to data in the absence of explicit negative evidence, overcoming prior beliefs in the process. A further benefit to the learner emerges from this – the ability to trade-off representational complexity and accuracy. For example, consider a scenario (depicted in Figure 1.5), where a reasoner is attempting to infer the "ideal level" of some continuously varying (but finite) property. Importantly, the reasoner in this scenario has a strong prior belief that there is a natural cut-off point that partitions the space into ideal values at the lower end and non-ideal values at the higher end (as in Figure 1.5, panel a). An alternative belief (one given little credence by the reasoner in this example) is that the ideal level represents a cut-off on both sides of a range (as in Figure 1.5, panel b). Where the reasoner adopts a weak sampling assumption in this example (Figure 1.5 – red line), no amount of data will shift them from their prior belief. More realistically, where the reasoner adopts a conservative sampling assumption, a strong prior belief in a particular representation will still be difficult to shift (Figure 1.5 – green line).[18] Under a stronger sampling assumption, however, the strong prior evidence is outweighed by a relatively small sample – the reasoner effectively learns a new representation for the inductive problem at hand (Figure 1.5 – blue line).

[18] The green line represents a mixed sampling assumption according to Equation 1.13. The parameter setting $\theta = 0.02$ reflects the value fitted to a conservative reasoner reported in Navarro et al. (2012).

The effect of a strong prior bias under different assumptions



(c) Learning simulation

Figure 1.5: A simulated learning scenario. The learner attempts to infer the "ideal level" of some quantity, and has a strong prior bias (10,000:1 in favour)) that the consequential region takes the form of a partition (as in **panel a**), rather than a range with an upper and lower bound (as in **panel b**). The simulation (**panel c**) shows the effect that this bias has under different sampling assumptions. Under a strong sampling assumption (blue line), just twelve samples are enough to drive the learner to a different view of the consequential region. In contrast, under weak sampling the learner cannot learn the alternative representation, regardless of the number of samples observed within a restricted range. Lastly, a learner who adopts a conservative sampling assumption will eventually revise his belief, but it may take considerably more data to do so – the green line depicts the learner's response having observed a sample ten times larger in size than the one shown (but within the same range).

While the scenario is somewhat simplistic it is not entirely unrealistic – there is reason to believe that people do recruit a variety of background information to better interpret the data at hand, including abstract overhypotheses (Kemp et al., 2007) and property-specific explanations (e.g., E. E. Smith et al., 1993). Chapter 2 explores this idea in the context of a category learning experiment. The basic intuition is that people with a stronger sampling assumption should be more likely than those with a weaker assumption to shift their representation of the target category as new data arrives.

There is precedence for the idea in the literature. Sanjana and Tenenbaum (2003) demonstrated how a strong sampling assumption and structured priors trade-off in category-based induction tasks. Using empirical similarity ratings for various categories of mammal (collected by Osherson et al., 1990) a taxonomic tree was created via standard agglomerative clustering algorithms (Duda, Hart, & Stork, 2001). A Bayesian generalisation model was specified (following Tenenbaum & Griffiths, 2001a) where each hypothesis represented a collection of one, two or three disjoint clusters (sub-trees). The prior was weighted in favour of fewer clusters, while the strong sampling likelihood function favours smaller hypotheses. As in the simple example above, this allows the two aspects to trade-off as sample size increases. As a consequence, simpler representations (fewer clusters) may be abandoned in favour of more complex ones in the face of sufficient additional evidence, but only if the size principle applies (even if only in diluted form). Sanjana and Tenenbaum (2003) found that behavioural data from three experiments were better captured by the Bayesian model than the best performing alternative, Osherson et al.'s (1990) similarity coverage model.

Another way of interpreting the effect of the size principle, is to say that it changes the evidentiary value of similarity (non-diversity) amongst sampled items. For example, as we saw in Figure 1.5, additional exemplars coinciding within the learner's region of interpolation are ignored under weak sampling, while the same "coincidence" is leveraged to drive a change in representation under strong sampling. In that example, the non-diversity observed involved the very quantity that was the basis for generalisation. But the same principle should apply when attempting to identify the relevant basis on which generalisation should proceed. For example, consider the following inductive argument: WALRUS → ELEPHANT, and compare it with a second: {WALRUS, WARTHOG} → ELEPHANT. In the first argument, the premise and conclusion categories share a number of features in common, any of which might serve as the basis for generalisation. In the second argument, the presence of a rare feature (tusks) in both premise categories seems more suggestive. Does the perception of a seemingly meaningful coincidence of this kind change the way that evidence is interpreted? And if so, is the effect unavoidable (as consequence of a feature weighted similarity comparison, for example), or does it depend on what people assume about the way the argument premises were sampled? I explore the effects of such coincidental similarity in Chapter 5.

Sampling with intent

Thus far, in discussing strong and weak sampling I have placed little emphasis on whether the data represents first-hand evidence – that is, "naturally" occurring data directly observed and interpreted by the reasoner themselves – or second-hand evidence supplied by another reasoner. But the distinction is an important one. Certainly the stakes are much higher when it comes to data that is social in origin. Limited physical access to data and a limited capacity to interpret it, means that we are critically reliant on second-hand evidence for much of our everyday reasoning. Given the lack of comparable constraints on the generative process behind the data we receive from others, it seems reasonable that people might adopt different assumptions when it comes to assessing its evidentiary weight. Whether people adopt an inferential stance that is *less* or *more* conservative as a result, is likely to depend on the context in which communication takes place, and may itself be the subject of further inference.

Another relevant matter that I have not yet addressed concerns how we sample information when providing it to others. How do we select examples to communicate a concept? How do we choose what to say when we refer to an object? These are examples of generative processes of which we as reasoners actually have first-hand experience (post infancy), so answers to questions such as these may offer insight into how our assumptions are shaped. And given that we experience both the production and comprehension of data, the possibility of bi-directional influence emerges. The study I describe in Chapter 6 focuses on aspects of how this bi-directional influence plays out when the spectre of deception is present. What follows then is a brief overview of the work on which Chapter 6 builds and an outline of the issues involved.

To begin on the production side, consider how a teacher might attempt to communicate a concept to a learner. Assuming an exhaustive demonstration is either impractical or impossible (because the number of possible exemplars is large or unbounded), what criteria should the teacher apply when sampling information to provide? A simple criteria would be to randomly sample information drawn from the concept. If this were the only criteria that applies when teaching concepts, then it alone would be sufficient basis to justify a strong sampling assumption on the part of the learner. However, equipped with a basic understanding of what it means to learn a concept from examples, the teacher might seek instead to give a better set of examples than one that might be chosen at random. Adopting such a strategy has obvious benefits compared with random sampling: allowing the teacher to convey the concept equally effectively but with fewer examples, or equally efficiently but more effectively. The question arises as to how to do better than random sampling.

Intuitively, a good set of examples and a set of good examples aren't necessarily the same thing. Tenenbaum and Griffiths (2001b) demonstrated this empirically, by asking people to rate the representativeness of different sets of birds. While robins individually were rated as more representative than other birds, three robins together (non-diverse sample) were rated as less representative than a robin, an ostrich, and a penguin (diverse sample), and a robin, a seagull and an eagle (intermediate sample), which rated highest overall. From a computational perspective, Tenenbaum and Griffiths (2001b) demonstrated a principled basis for perceived representativeness by showing how it corresponds to a standard Bayesian measure for the weight of evidence – the log likelihood ratio (Good, 1960; Klayman & Ha, 1987; McKenzie & Mikkelsen, 2007). Data sampled according to such a principle of representativeness should be more helpful to the learner than data that was randomly sampled from the concept. A random sample that is coincidentally non-diverse will lead the learner to generalise too narrowly, while a sample that is overly diverse (above average) may result in over-generalisation.

Empirical research confirms that both adults and children do sample systematically from the concept when teaching by example. A study by Avrahami et al. (1997) examined how adults teach linearly separable concepts involving confusable stimuli. Given the choice of both positive and negative examples, people favoured positive examples of the category and avoided using examples close to the boundary. Rhodes et al. (2010) studied teaching by example with 6 year olds using a property induction task. Children were asked to select between two sets of three examples to teach another young child about a particular property (e.g. a four-chambered heart). The goal was to indicate which animals of a particular kind (dogs, for example) had the property in question. The children showed a significant preference for the diverse set (e.g. golden retriever, Dalmatian and collie) over the non-diverse set (three Dalmatians), demonstrating that children from an early age understand how some sets of examples are better than others at demonstrating the breadth of a concept.

Interestingly, Rhodes et al. (2010) found that pedagogical context matters in this regard. When children were asked to act as a scientist and choose three examples to test whether all dogs (for example) have the given property, their preference for the diverse examples was at chance. Studying adult performance using the same tasks, Rhodes et al. (2010) found that adults preferred diverse examples in both the teaching and discovery scenarios. On the basis of their own research, Csibra and Gergely (2006) propose that children come equipped at an early age with the cognitive biases necessary to support pedagogical exchanges. Rhodes et al. (2010) suggest that attention to sample composition (i.e. sampling assumptions) is supported by different cognitive processes in pedagogical and non-pedagogical contexts.

On the other hand, most models of category and concept learning proposed in the literature make no account for this, yet successfully account for human performance across a range of learning tasks. Typically, such models assume implicitly (e.g., Kruschke, 1992; Love, Medin, & Gureckis, 2004; Nosofsky, 1986) or explicitly (e.g., Anderson, 1991; Fried & Holyoak, 1984; Tenenbaum & Griffiths, 2001a) that examples are selected by the teacher (or "the world") at random, incorporating the equivalent of weak or strong sampling. The contrast between the successful application of random sampling and the importance of pedagogical context raises an interesting question. While the idea that people do sample information systematically is intuitively obvious, when do "better than random" sampling assumptions come into play, and where do they make a difference?

Shafto and Goodman (2008; see also Shafto et al., 2014) introduce a Bayesian model of concept teaching and learning that captures the relationship between evidence production and comprehension. The model makes two key assumptions. The *rational teacher* assumption is that teachers choose examples designed to help the learner, according to the following:

$$P_{\text{TEACHER}}(x|h) \propto (P_{\text{LEARNER}}(h|x))^{\alpha} \tag{1.14}$$

which formally captures the teacher's goal to increase the learner's belief in the true hypothesis (the concept to be demonstrated). Here x denotes a candidate example, h denotes the concept to be demonstrated, and α reflects the extent to which the teacher consistently chooses in an optimal fashion. High values of α denote (near) optimal choosing, while $\alpha = 0$ implies that the teacher samples at random. A teacher who samples in this way will select examples of the concept that are representative in much the same sense that Tenenbaum and Griffiths (2001b) describe. The *rational learner* assumption is that the learner updates their beliefs according to Bayesian inference:

$$P_{\text{LEARNER}}(h|x) \propto P_{\text{TEACHER}}(x|h)P(h) \tag{1.15}$$

(cf Equation 1.7), with the expectation not just of positive examples of the concept (as in strong sampling), but representative ones (i.e. the likelihood function reflects the rational teacher assumption).

The two equations (1.14 and 1.15) are linked, reflecting the inter-dependence of optimal teaching and learning. Computationally, the model specifies that the optimal teacher and learner jointly arrive at a solution that satisfies the two equations, but it does not specify *how* they would do so. One intuitive way to think of how this might be achieved is to consider that a successive series of "if she thinks, I think, she thinks..." thoughts might occur to both parties. Reasoning reciprocally and recursively in this way, the samples chosen by the teacher and the inferences drawn by the learner will approach a simultaneous and optimal solution to both equations. A related Bayesian framework, the Rational Speech Act (RSA), models the interdependency between the production and comprehension of data (in this case words and sentences) in a similar way (M. C. Frank & Goodman, 2012; N. D. Goodman & Frank, 2016). The level of "he thinks, she thinks..." reasoning is made explicit in the RSA through nested instantiation of pragmatic listener and pragmatic speaker assumptions. Empirical evidence to date suggests that depth of reasoning of this form is typically limited (e.g., Colman, 2003; Franke & Degen, 2016; Stiller, Goodman, & Frank, 2015; Vogel, Potts, & Jurafsky, 2013). Nonetheless, even at a limited depth, this form of reflective reasoning about a counterpart has the potential to establish bi-directional influence on behaviour. I examine this form of reasoning in Chapter 6 in a scenario where the interests of two parties (a sender of information and a receiver) are opposed, potentially heightening the incentive for each person to try to "out-think" the other.

Issues of deep recursion and optimality aside, (Shafto et al., 2014) examined the predictions of the pedagogical sampling model across three experiments involving rule-based, prototype and causally-structured concepts respectively. On the teaching side, the results suggest that people choose helpful examples, well predicted by the model. On the learning side, people drew stronger inferences from the data when presented in a pedagogical context (broadly in line with model predictions) than when the same data was presented in a non-pedagogical context. The work reinforces the benefits of stronger sampling assumptions, allowing the reasoner to learn rule-based concepts faster (from fewer examples), and to learn distributional information more accurately.

In teaching concepts by example, the challenge is to promote generalisation beyond the limited set of examples provided. For the speaker in regular conversation, there is often a different challenge. Both speaker and listener have cognitive capacity constraints that put downward pressure on the length of utterances. Conversational maxims which implore the speaker to say no more than is required (Grice, 1989), are predicated on this notion. The desire for parsimony can often lead speakers to convey their meaning by means of an ambiguous utterance. Indeed, the meaning of much of everyday conversation is not to be found in a literal interpretation alone. The RSA model emphasises the importance of "informative" sampling in bridging the gap between what is said and what is meant. According to the model, the speaker chooses her words intentionally to promote belief in the meaning she wishes to convey, and the listener assumes that she does so. Without some form of bi-directional inference of this nature, a variety of ambiguous expression in everyday speech would not be resolvable. By modelling pragmatic language understanding as Bayesian inference based on informative sampling, the RSA has successfully captured important examples of ambiguous language use, including identifying referents (M. C. Frank & Goodman, 2012; Qing & Franke, 2015), and interpreting scalar implicatures (N. D. Goodman & Stuhlmüller, 2013). I explore the role of sampling assumptions in interpreting ambiguous information in Chapter 6, which examines how people may exploit the "ambiguity reduction" inherent in communicative inference to mislead by implication.

In summary, these applications of the RSA and the pedagogical sampling model demonstrate both the value to the reasoner of adopting an informative sampling assumption, and the variety of situations in which it applies. But widespread reliance on a strong inferential assumption can bring problems of its own. For example, Bonawitz et al. (2011) found that because learners who adopt an informative sampling assumption could make efficient use of small samples, they were less inclined to look for further evidence. This raises a possibility that I explore in Chapter 6 – that speakers wishing to conceal the truth may use this fact to their advantage through limited (yet informative) disclosure.

Whether an informant lacks knowledge, has alternative motives, or is simply not sufficiently motivated, there are often good reasons to exercise a more conservative inferential approach. However to do so uniformly is to give up the potential benefits which I have just been outlining. What do people do to avoid this inferential dilemma? Sperber et al. (2010) suggest that people have a suite of cognitive mechanisms for *epistemic vigilance*, which serve to limit the risks of accidental or intentional misinformation. Research shows that even young children draw inferences about their informants based on perceived expertise, familiarity, and from more sophisticated evidence such as group consensus (for detailed discussion see Eaves & Shafto, 2012). A computational basis for this kind of screening mechanism has been explored using Bayesian models proposing various forms of *joint inference* (e.g., N. D. Goodman & Frank, 2016; Gweon et al., 2010; Shafto, Eaves, Navarro, & Perfors, 2012).

The idea behind this kind of joint inference is that the alternative hypotheses entertained by the reasoner reflect not only their uncertainty about some question of interest (the meaning of a word, or the extension of a property, for example), but also about aspects of their informant that they do not take for granted. The reasoner uses the data to update their beliefs about all such aspects. In essence, if people are capable of this kind of joint inference then it means they can test alternative assumptions "on the fly", transforming what are effectively context-based sampling presumptions into more fully-fledged sampling theories subjected to data-based scrutiny. Such a capability might go some way to explaining if and how sampling assumptions are bootstrapped during infancy or early childhood (Gweon et al., 2010). In Chapter 6, I explore the question from the production side, essentially asking whether speakers act as if people make content-based sampling assumptions. Chapter 5 also examines this issue. By systematically varying both sampling cover story and sample contents, the experiment I report there offers a glimpse at the strength of sampling assumption people infer on the basis of differences in the data alone.

1.5 Overview of new research

Throughout this chapter, I have attempted to highlight the ubiquity and utility of the human capacity to reason beyond the data. And I have begun to make a case for why our sampling assumptions, by influencing the way we view the evidence in the data, might play a central role in underwriting our capacity for inductive reasoning, driving the strength and quality of inference as a result. I conclude this chapter with a short sketch of the research I have conducted in order to investigate the strength of this claim.

To some degree it is fair to say that the study of sampling assumptions and their effect on inductive inference and inductive meta-inference is in its infancy. So there was (and remains) no shortage of important, interesting and open questions to consider. The questions I have chosen to pursue represent less a linear progression of studies providing a single continuous thread of evidence, and more a collection of distinct but related research threads. Nonetheless, the studies I present in the chapters that follow were each designed to speak to the central issue of concern here: that is, how we form and use our theories of what lies behind the data in order to reason beyond it.

The problem of where to draw the line

The first study I present across Chapters 2–4, examines of the role of sampling assumptions in category learning. The ability to acquire categories from labelled examples (that is, via *supervised learning*) is a vital part of our cognitive toolkit. This is particularly so during development, supporting as it does the transfer of knowledge to previously unobserved category members. Acquisition of categories and category boundaries involves one of the fundamental inductive reasoning challenges, namely how far beyond the data to generalise.

I begin in Chapter 2 by examining how generalisation of a single learned category to novel items, is affected by the sampling assumptions that people adopt. By employing an experimental design where the diversity of examples is held constant but sample size is increased, I attempt to isolate the effect of category frequency on the category representation learned. In so doing, the experiment addresses an important question. If the learner's beliefs do reflect an uncertain and graded notion of category membership (as in Rips, 1975; Rosch & Mervis, 1975, for example), then how does that graded nature change in response to category frequency, and in particular, how does this adaptation depend on people's theories of how the data was sampled?

Building on this work, the experiment described in Chapter 3 introduces the complication of a second category. The second category has the potential to change the kinds of category representation that people learn, depending for example, on whether people believe they are learning two independent or mutually exclusive categories. Further, an additional source of information arises in the form of category base rates. By providing different numbers of exemplars across two categories of stimuli, I examine whether and how sampling assumptions affect the evidentiary value of base rate information. The central question addressed is how all of this affects the category boundaries that people infer and whether the impact of sampling assumptions changes in some way with the introduction of a second category.

As well as illuminating the effect of sampling assumptions on category learning, the experiments in Chapters 2 and 3, are important in another way. Taken together, the work has the potential to shed light on an interesting discrepancy that occurs between empirical results discussed in the categorisation literature and the adjacent generalisation literature. In generalisation tasks, as exemplified by many of the studies refereed to in section 1.4 (e.g., Navarro et al., 2012; Sanjana & Tenenbaum, 2003; Xu & Tenenbaum, 2007a), increasing sample size typically has a tightening effect on generalisation. In contrast, the typical effect of increased exemplar frequency in categorisation tasks is a widening of generalisation (e.g., Nosofsky, 1988, but see Hendrickson et al., 2019 for an extended discussion of this issue). Despite the differences between the paradigms, this discrepancy is somewhat puzzling, particularly if common mechanisms for inductive reasoning underlie performance on both kinds of task. One obvious difference is in the number of categories involved. Tightening of generalisation is typically observed in tasks where a single concept is being considered. Whereas the kind of categorisation task where widening of generalisation is observed usually involves two or more categories. Why might the number of concepts or categories make such a difference? Chapters 2 and 3 offer insight into this question by exploring the idea that the contextual shift between one-category and multi-category learning changes the sampling assumption implicit in the task, and that the change in patterns of generalisation reflect this.

Although it relies on the same experimental framework as employed in Chapters 2 and 3, I leave my sketch of Chapter 4 until last, because it gets at a different aspect of sampling assumptions than the rest of the work I describe.

The problem of representation

The work described in Chapter 2 also examines another important challenge facing the learner, concerning the appropriate mental representation to adopt when learning a category. If the mental representation that the learner adopts is well matched to the concept being learned, the learner may require fewer examples to acquire it, or generalise more accurately from a fixed set of examples (e.g., Attneave, 1957; Posner & Keele, 1968). A variety of category representations have been modelled in the literature including prototypes (J. D. Smith & Minda, 1998), exemplars (Nosofsky, 1986), decision boundaries (Ashby & Townsend, 1986), independent regions (Navarro, 2006), and so on. And the idea that learners adapt their mental representation of a category "on the fly" has also been studied. For example, there is evidence to suggest a change from prototype to exemplar representation (Griffiths, Canini, Sanborn, & Navarro, 2007), a change from exemplar to prototype (Homa, Sterling, & Trepel, 1981) and a mixture of representations that vary throughout learning (Vanpaemal & Navarro, 2007) or across individuals (Kalish & Kruschke, 1997).

The question thus arises as to where sampling assumptions fit in. If representations and sampling assumptions both affect learning, then how do the two interact? As discussed in section 1.4, one consequence of adopting a stronger sampling assumption is that it offers the learner the potential to overcome strong prior biases and acquire alternative category representations from data – something which may be difficult or impossible to learn under a more conservative sampling assumption. The experiment described in Chapter 2 represents an empirical test of the idea that the sampling assumptions people adopt influence the category representations they employ, and the two elements combined have a significant impact on generalisation performance.

The problem of relevance

A fundamental challenge of inductive inference is to reason about problems which are under-determined, both by the data at hand and by the sum of past experience. While data is seldom in short supply (the stimulus environment and our memory is full of it) evidence often is. Short of evaluating the evidentiary weight of all available data, it makes sense to prioritise the use of the most relevant data. Relevance theory (Wilson & Sperber, 2004) holds that central to much of human cognition are processes that seek to maximise the *relevance* of the information available. The relevance of any information reflects a tension between the cognitive effect that it affords and the cognitive effort required to employ it. In computational terms, relevance can be thought of as something akin to evidentiary weight, but with a built-in penalty for representational and computational complexity. If the relevance of information is related to its evidentiary weight, then the question arises as to whether the perception of relevance may be impacted by sampling assumptions. This question is the basis for the study described in Chapter 5.

In particular, the experimental investigation I describe looks at how people reason in a category-based induction task where the basis for generalisation is not immediately clear. Blank properties are used in order to promote the reasoner's search for relevant information. For inductive arguments of the kind used, the addition of further positive examples having the property in question typically acts to strengthen the argument, an effect known as *premise monotoncity* (Osherson et al., 1990). But violations of this effect, that is *premise non-monotonicity* have also been observed (e.g., Medin et al., 2003). Which of these phenomena people exhibit depends upon the perceived relevance amongst the premise items. I test the conjecture that the process of determining the relevant *features* of the data to use for the projection of novel properties is itself dependent upon people's sampling assumptions. Although it is the main focus of the study, I also gain the first glimpse of whether the data itself helps to shape people's sampling assumptions. The empirical results I report suport a claim for the importance of sampling assumptions in shaping the strength and *direction* of inductive inference.

The problem of calibration

Across the three major studies I present, the inferential task for the reasoner gets progressively more difficult. In Chapters 2–4, the method of sampling and the basis on which generalisation should proceed are both made clear. In Chapter 5, the reasoner must infer the relevant basis for generalisation themselves, but the method by which data was sampled is still clear ("no cover story" conditions notwithstanding). In the study of *deception without lying* that I present in Chapter 6, reasoners face arguably their most difficult challenge – calibrating their own inference with the inferences of another. Given the extent to which our everyday reasoning relies on socially generated data, and frequently involves reasoning well beyond it, the need to calibrate our sampling assumptions based on contextual and content-based cues becomes all the more important.

Through a combination of computational modelling and behavioural experiments, inspired by previous work on pedagogical learning (Shafto & Goodman, 2008; Shafto et al., 2014) and rational communication (M. C. Frank & Goodman, 2012; N. D. Goodman & Stuhlmüller, 2013), I investigate the issues of calibrated inference from two related perspectives. On the comprehension side, people play the role of a *receiver* of information who must draw inferences from information provided. The receiver's challenge lies in figuring out just how the information was sampled. In regular communication, as opposed to pedagogical settings for example, an assumption of helpful, representative or simply unbiased sampling is not always appropriate. Reasoners must recruit what cues they can to infer the extent that reasoning beyond the data is warranted. On the production side, people play the role of the *sender* who selects the information in the first place. The sender's goal is to do what they can to conceal the truth. In deciding how best to go about this, the sender's challenge is to infer how the data they select will be interpreted. Having people actively sample information to provide another affords the opportunity to examine people's own intuitive understanding of how people reason from data and the factors that affect the process. For example, what cues do people think that receiver's will be sensitive to in deciding whether or not to reason beyond the data provided?

By framing twin experiments in a context where the goals of interlocutors are not aligned, I gain additional diagnostic capacity to determine whether people make nuanced sampling assumptions that take into account the assumptions of their counterpart, or simply reason from the constraints implicit in the sampling process. Thus, the study offers the opportunity to tease apart the effects of how sampling constraints, perceived intent and the content of data itself play a role in how people reason when communicating.

Do sampling assumptions influence learning or reasoning?

Each of the studies introduced thus far deal with "what" style questions that get at the computational problem that the reasoner might be attempting to solve. What is at issue in these studies is not an account of *how* the reasoner goes about solving the computational problem, but rather the nature of the computation itself - what are the inputs and outputs, what are the computational abstractions involved and how are all these things related. In psychological terms this amounts to exploring the things that people's sampling assumptions are sensitive to, and the direction and magnitude of the effect that such sensitivities have on the outcomes of inference.

The study I present in Chapter 4, though something of a departure from this computational theme, is nonetheless a question that is begged by the results of each of the other experimental investigations that I have conducted as well as related findings in the literature (see section 1.4, for example). If people's sampling assumptions can be said to effect the inferences that they draw, then what is the nature of that effect? Is it a *reasoning effect* that comes into play when drawing an inference on the basis of previously observed data? Or is it a *learning effect* which effects the mental representation of data when it is first encountered and encoded in memory? Put another way, if people's sampling assumptions determine how they evaluate data as evidence, then is the evidence evaluated upon *encoding* or *retrieval* of the data? Of course it may be that either effect is possible, or that both kinds of effects may impact a single inference.

As I mentioned in section 1.3, the issue connects with more general questions raised in the literature. For instance, whether generalisation gradients are a product of learning (as suggested by Hull, 1943) or reasoning (per Razran, 1949). Or whether representational abstractions are formed during learning or extracted later when a generalisation decision is required (a question raised by Posner & Keele, 1968). To the best of my knowledge and despite these important implications, this learning/reasoning distinction has not been directly addressed in the literature, at least in as far as it relates to sampling assumptions. Thus, although the experiment described in Chapter 4 should be regarded as a proof of concept, it nonetheless complements the other new work I present in the chapters that follow.

Study I

WHEN THE BASIS FOR INDUCTION IS UNCLEAR

STATEMENT OF AUTHORSHIP

TITLE OF PAPER	Representational and sampling assumptions drive individual differences in single category generalisation
PUBLICATION STATUS	Published
PUBLICATION DETAILS	**K Ransom**, A Hendrickson, A Perfors, and DJ Navarro (2018). Representational and sampling assumptions drive individual differences in single category generalisation. In C Kalish, M Rau, J Zhu and T Rogers (Ed.) *Proceedings of the 40th Annual Conference of the Cognitive Science Society.*

Principal author

NAME OF PRINCIPLE AUTHOR (CANDIDATE)	Keith Ransom
CONTRIBUTION TO THE PAPER	Designed and ran experiments, performed data analysis, wrote manuscript and acted as corresponding author.
OVERALL PERCENTAGE (%)	75%
CERTIFICATION	This paper reports on original research I conducted during the period of my Higher Degree by Research candidature and is not subject to any obligations or contractual agreements with a third party that would constrain its inclusion in this thesis. I am the primary author of this paper.
SIGNATURE	
DATE	18/08/19

Co-author contributions

By signing the Statement of Authorship, each author certifies that:

1. the candidate's stated contribution to the publication is accurate (as detailed above);
2. permission is granted for the candidate in include the publication in the thesis; and
3. the sum of all co-author contributions is equal to 100% less the candidate's stated contribution.

NAME OF CO-AUTHOR	Andrew Hendrickson
CONTRIBUTION TO THE PAPER	Supervised development of work, helped with experimental design and editing of the manuscript.
SIGNATURE	
DATE	18/08/19

NAME OF CO-AUTHOR	Amy Perfors
CONTRIBUTION TO THE PAPER	Supervised development of work, helped with experimental design and editing of the manuscript.
SIGNATURE	
DATE	18/08/19

NAME OF CO-AUTHOR	Danielle Navarro
CONTRIBUTION TO THE PAPER	Supervised development of work, helped with experimental design and editing of the manuscript.
SIGNATURE	
DATE	18/08/19

2 IS THIS A DAX I SEE BEFORE ME?

Human activity requires an ability to generalise beyond the available evidence, but when examples are limited – as they nearly always are – the problem of how to do so becomes particularly acute. In addressing this problem, Shepard (1987) established the importance of *representation*, and subsequent work explored how representations shift as new data is observed. A different strand of work extending the Bayesian framework of Tenenbaum and Griffiths (2001a) established the importance of *sampling assumptions* in generalisation as well. Here we present evidence to suggest that these two issues should be considered jointly. We report two experiments which reveal replicable qualitative patterns of individual differences in the representation of a single category, while also showing that sampling assumptions interact with these to drive generalisation. Our results demonstrate that how people shift their category representation depends upon their sampling assumptions, and that these representational shifts drive much of the observed learning.

2.1 Introduction

Suppose that, upon encountering a wallaby for the first time, I am reliably informed that *wallabies are dax*. What should I infer to be the extension of the property *dax*? If I know that *dax* is a biological property I might generalise to other macropods, marsupials, or mammals. Alternatively, if *dax* describes a behaviour I might instead generalise to other hopping or grazing animals. As this thought experiment suggests, human category representations are structured and complex; multiple systems of categories are relevant to a single domain and different systems of knowledge are relevant in different contexts (Heit & Rubinstein, 1994; Ross & Murphy, 1999).

Although there is some work investigating how people acquire multiple systems of categories (Shafto, Kemp, Mansinghka, & Tenenbaum, 2011) and learn which representations are relevant to inductive problems like this (Austerweil & Griffiths, 2010), very little is known about individual differences in representation. Do such differences exist, and can they be measured? When people learn based on new data, do their representations shift? If so, how and why? Do their assumptions about how the data were generated drive any of this? These are the questions we focus on in this paper.

Representation and generalisation

The problem we consider is ostensibly a simple one: learning how to generalise along a single stimulus continuous dimension. Stimulus generalisation in this situation often resembles an exponential decay as a function of distance along the relevant dimension, but only when formulated with respect to the proper stimulus representation (Shepard, 1987). When adapting Shepard's analysis into an explicitly Bayesian framework, Tenenbaum and Griffiths (2001a) noted that generalisation from multiple examples allows for many different possible stimulus representations. Indeed, there are many different assumptions a learner might make about category representation. These include exemplar models (Nosofsky, 1986), prototype models (J. D. Smith & Minda, 1998), decision boundaries (Ashby & Townsend, 1986), critical regions that mimic prototype models if the regions are connected (Tenenbaum & Griffiths, 2001a), or exemplar models in which each item corresponds to a region (Navarro, 2006). Additionally, these representations are not fixed and stable. Evidence from category learning has shown that human learners tend to "grow" category representations as they see additional items, with a shift during learning from prototype to exemplar representations (Griffiths et al., 2007; Love et al., 2004), or from exemplar to prototype (Homa et al., 1981), or a mixture of representations across individuals (Kalish & Kruschke, 1997).

Sampling and generalisation

An adjacent literature on inductive generalisation has revealed that what the learner assumes about how this data came to be the data has a substantial influence on the inferences people draw. These *sampling assumptions* affect inferences in concept learning tasks (Navarro et al., 2012), property induction tasks (Ransom, Perfors, & Navarro, 2016), and word learning problems (Xu & Tenenbaum, 2007a).

While there are many possible sampling assumptions that one might adopt (e.g., Ransom, Voorspoels, Perfors, & Navarro, 2017; Shafto et al., 2014), much of the literature has focused on two simple possibilities. A helpful teacher is likely to choose positive examples that belong to the relevant category (known as strong sampling), whereas a random sampling process selects exemplars independently of the category label (known as weak sampling). The difference between the two leads to a variety of differences in how people generalise: most notably, people tend to *tighten* their generalisations with additional data if they are assuming strong sampling, but don't if they aren't (e.g., Ransom et al., 2016; Xu & Tenenbaum, 2007a).

SAMPLING AND REPRESENTATION?

If both representation and sampling assumptions shape generalisation, how do they fit together? The literature on sampling assumptions typically assumes a fixed stimulus representation, and the literature on stimulus representation has given little consideration to the manner in which exemplars are chosen. In this paper, we present empirical evidence suggesting that these two problems should be considered together. We report results from two experiments involving a simple inductive generalisation task that manipulates the sampling assumptions across conditions. We find evidence for individual differences in category representation, with different participants appearing to represent categories in different ways. Moreover, there appears to be an interaction between people's representations and the degree to which they are sensitive to the sampling manipulation. Observations selected by a helpful teacher are more likely to cause people to *shift* their mental representation of the category in a consistent direction than if the same observations are selected at random. In fact, these representational shifts seem to account for the largest share of learning in the task.

2.2 EXPERIMENT 1

Experiment 1 is a single category generalisation experiment that, within the same experimental framework, combines manipulations of sample size (as in Navarro et al., 2012; Vong et al., 2013) and sampling cover story (as in Ransom et al., 2016; Xu & Tenenbaum, 2007a). As a post-hoc analysis, we use people's responses across all test items to identify clusters of people who generate similar patterns of generalisation. These patterns are then used as predicted outcomes in Experiment 2, where they are explicitly connected to representational clusters. Furthermore, the assignment of individual behaviour to clusters is tracked during learning, in order to determine whether representational shifts correspond to learning outcomes.

METHOD

Participants

603 people participated in this experiment via Amazon Mechanical Turk, where they were paid $1.30US for the 5-10 minute task. 45% were female, 93% were from the US, and median age was 32 (range: 19 to 77).

Design

People were randomly assigned to one of three conditions that varied the number of category exemplars ("Wuggams") as well as the manner in which they were sampled. In the

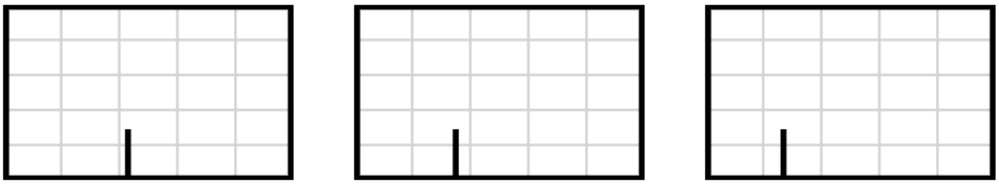


Figure 2.1: **Example stimuli**. Items varied only in the position of the short black vertical line along the bottom edge of the rectangle.

FOUR condition ($N = 194$) participants were shown four exemplars with no explanation offered for how these examples were chosen. Participants in the TWELVE HELPFUL ($N = 200$) and TWELVE RANDOM ($N = 209$) conditions were also shown the same four exemplars with no explanation, but were then subsequently shown eight more exemplars for which an explanation was given. In the TWELVE HELPFUL condition people were told that the additional examples had been intentionally chosen to help them understand the category, whereas people in the TWELVE RANDOM condition randomly selected additional items themselves.

Stimuli

Stimuli consisted of a black rectangular frame drawn against a white background, with a vertical black line inside attached to the bottom edge (see Figure 6.3). The rectangle was sized to occupy 14% of the available horizontal viewport, with an aspect ratio of 5:3, and the vertical line extended 29% of the rectangle's height. To assist with stimulus discriminability, four evenly spaced light grey vertical and horizontal lines were drawn within the rectangle. Stimuli varied along a single dimension, corresponding to the horizontal position of the vertical line within the rectangle (referred to later as the *stimulus value*).

The full set of training stimuli included 12 examples with stimulus values ranging from 21% to 43% in increments of 2%. People in the TWELVE HELPFUL and TWELVE RANDOM conditions saw all 12 examples, while those in the FOUR condition saw four, including the two extreme examples (at 21% and 43%) plus two random others in between. The test stimuli consisted of 19 items with values ranging from 5% to 95% in increments of 5%.

Procedure

The experiment consisted of a training phase where people were shown examples from the target category, followed by a test phase where they were asked to decide whether previously unseen items were in that category.

Training. Participants were told that the purpose of the experiment was to see how people judged whether or not unfamiliar objects were in the same category as known examples. In the FOUR condition the instructions stated:

> So, we'll start by showing you some objects that all belong to the same category («`Wuggams`»).

at which point four training examples were displayed simultaneously on-screen. Participants in the the other two conditions were given the same introduction. However, after the initial examples were shown those in the TWELVE RANDOM condition were further informed:

> **The computer has assigned you to experiment group «`J8`»** so we're going to let you pick an additional «`8`» items at random from our collection, and let you see any «`Wuggams`» that you find.

Following this a 6 × 5 arrangement of icons resembling packing boxes was displayed on screen, and people were asked to select eight boxes one by one. After clicking on an icon the image was replaced with that of an open box, people were informed that they had found a «`Wuggam`» inside, and one of the training examples was added to the display.

The TWELVE HELPFUL condition proceeded along similar lines, but people were instead told:

> **The computer has assigned you to experiment group «`K8`»** so we're going to help you by showing you an additional «`8`» «`Wuggams`» **chosen by a helpful teacher** to give you a good idea of the full range of «`Wuggams`».

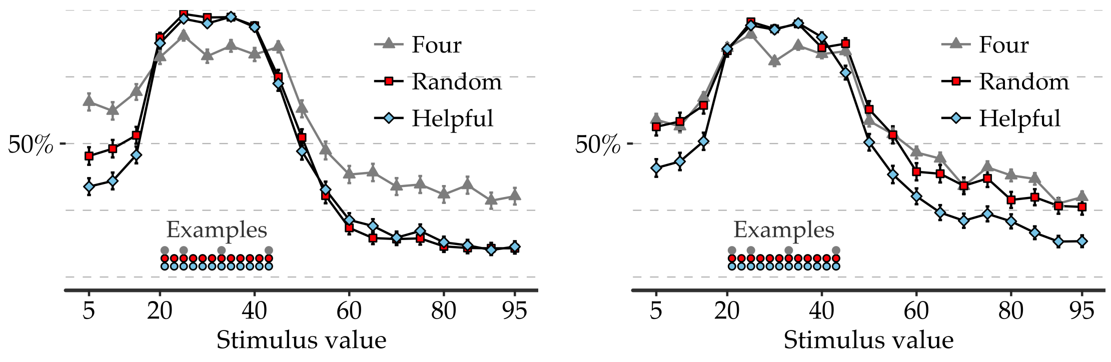
After which, the array of boxes was displayed with eight of the boxes already opened. Simultaneously, the display was updated with the eight additional examples. In all conditions the on-screen presentation order was randomised.

Testing. To minimise any memory effects, the training examples remained on screen during testing, along with a reminder of how the exemplars were selected. Participants in all conditions were shown the 19 test stimuli one at a time in random order; this sequence was repeated four times. The test query was a simple yes or no question, "Do you think this object is in the «`Wuggam`» category?".

RESULTS AND DISCUSSION

The results are shown in Figure 4.3(a), which plots the proportion of trials on which each test item was assigned to the Wuggam category in each condition. There is a clear effect of sample size: people who saw 12 examples generalised to a narrower range of test items than those who saw 4. A Bayesian ANOVA reveals strong evidence ($\mathrm{BF}_{10} > 10^6$) for a model that includes effects of stimulus value, sample size and an interaction, tested against a null model that includes only the effect of stimulus value.[1] However, the cover story

[1] Model comparisons included a random intercept for each subject, and were fit using default priors (Liang, Paulo, Molina, Clyde, & Berger, 2012; Rouder, Morey, Speckman, & Province, 2012) from the BayesFactor package (version 0.9.12-2) in R (version 3.4.3).



(a) Experiment 1.

(b) Experiment 2.

Figure 2.2: Performance on a one category generalisation task as a function of sampling procedure (manipulated between subjects) and sample size (manipulated between subjects in Experiment 1 and within subjects in Experiment 2). The graphs show the proportion of positive responses to the question: "Do you think this object is in the «Wuggam» category?" for each of the test stimuli. The performance of people who saw four examples of the target category (grey line) is contrasted with two groups of people who saw 12 examples (black lines). In Experiment 1, people tightened their generalisations as more data is observed, but the sampling manipulation had little effect; whether people actively sampled the additional examples at random (red squares) or were told that the items had been selected by a helpful teacher (blue diamonds), they generalised less when they saw 12 examples rather than 4. In Experiment 2, where the wording of the sampling manipulation was slightly adjusted, tightening with increased sample size occurs, but only in the HELPFUL condition.

appeared to have little to no effect, with modest evidence favouring the null hypothesis ($BF_{01} = 10$) that generalisation patterns were the same in both 12-item conditions.

The one exception to this pattern is the three test items to the far left of Figure 4.3(a). Visual inspection suggests that participants in the TWELVE HELPFUL condition were somewhat less willing to generalize to these items than were people in the TWELVE RANDOM condition. This asymmetric pattern is not predicted by "standard" implementations of the Bayesian generalisation model (e.g., Navarro et al., 2012; Vong et al., 2013). However, it is consistent with a shift in the proportion of people using a single decision boundary, which should not fall off on the far (left) side of the observed exemplars.

To examine this possibility we conducted a post hoc clustering analysis of generalisation curves at the individual subject level. This analysis, which was based on a Dirichlet process mixture model, automatically identified 11 different "patterns" of generalisation curves. Nine of the 11 patterns accounted for 98% of the data; and of these nine, three were minor variants of the others.[2] The remaining six patterns (illustrated in Figure 2.3)

[2] We used the `BayesianGaussianMixture` class from the `scikit.learn` module (v0.19.1) under Python 3.6.3. The concentration parameter for the Dirichlet process was set to 1, the multivariate Gaussian distribution assumed a diagonal covariance structure, and the random seed was set to 1. Each generalisation pattern was encoded as a point in 19-dimensional space with each dimension corresponding to a stimulus

form the core of the analysis in Experiment 2, and cover 85% of the data from that experiment. We turn to it next.

2.3 Experiment 2

Participants

404 people participated in this experiment via Amazon Mechanical Turk, where they were paid $1.50US for the 10-15 minute task. 48% were female, 94% were from the US, and median age was 32 (range: 18 to 71).

Design, stimuli & procedure

Experiment 2 was a preregistered[3] replication and extension of Experiment 1. The two experiments were identical except for three key differences. First, we adopted a within subject manipulation of sample size. Regardless of condition, participants were shown four exemplars with no sampling explanation given and then tested. They were then shown an additional eight exemplars – either within a Helpful (N=205) cover story or a Random one (N=199) – and then tested a second time. Testing each person twice allows us to assess how their representation changed based on four examples or twelve.

Second, at the end of each test phase participants were asked to identify the strategy they used, selecting one of the six options listed in Figure 2.3(c). This data is useful for determining whether their reported strategies correspond to the generalisation patterns our model assigns to them.

Third, the cover story in the Random condition was altered slightly in order to leave open the possibility that some boxes might not contain Wuggams. People were told that "some of the boxes are stuck and won't open; in that case just try another." Each person sampled 11 boxes but saw only 8 «`Wuggams`» in total; the other three times (when the box remained closed) occurred in a random order with the constraint that the first and last item was always a «`Wuggam`».

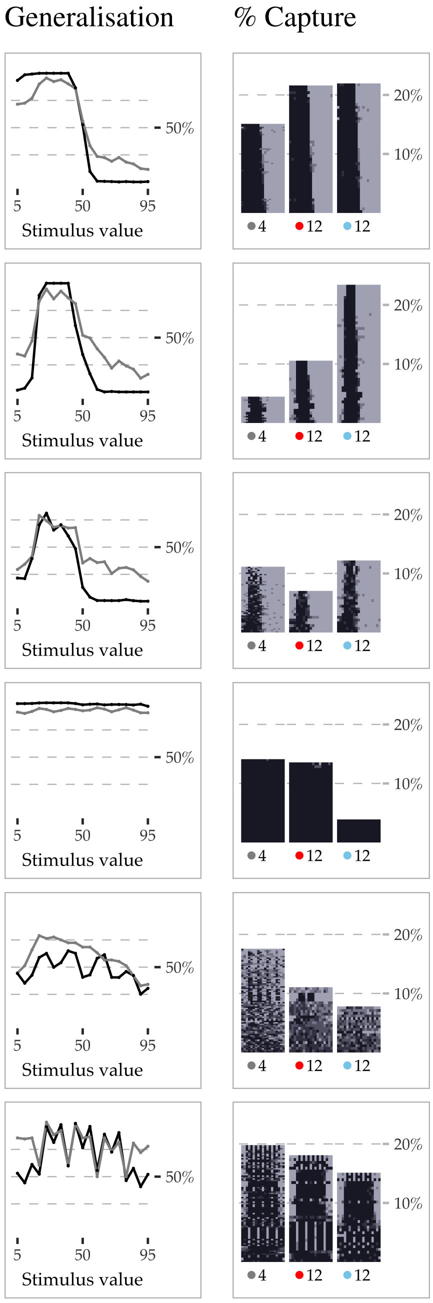
Results and discussion

Generalisation

Generalisation patterns in Experiment 2 partially replicated the results from Experiment 1, as shown in Figure 4.3(b). As before, we find a clear effect of sample size ($BF_{10} > 10^6$),

value included in the test items and the value along each dimension corresponding to the probability of generalising the category label to that test stimulus. Supplemental materials describing details of the model and all 11 patterns are here: https://tinyurl.com/RPNH18

[3] https://aspredicted.org/3tq89.pdf



Representation

PARTITION

An item is a «Wuggam» if the **short black line is on one side of a certain boundary**. I use the examples to work out where that boundary is, and which side the «Wuggams» are on.

RANGE

An item is a «Wuggam» if the **short black line appears within a certain range from side to side**. I use the example «Wuggams» to work out where that range starts and ends.

BY EXAMPLE

An item is a «Wuggam» if the **short black line is in the same position as one of the examples given**. I use the examples given to check for a match

ALL

An item is a «Wuggam» if it **contains a short black line**. No matter where the line is, I mark all such items as «Wuggams».

ANY

Any item could be a «Wuggam», there's no way to tell, so I just guess.

OTHER

None of the above. An item is a «Wuggam» if...
(*a free text box followed*)

(a) (b) (c)

Figure 2.3: A graphical depiction of individual differences in generalisation in Experiment 2. The panel columns represent: (a) Aggregate generalisation curves for people grouped by data driven pattern definition (black lines) and by response to self report question (grey lines). (b) The proportion of people allocated to a given pattern. The three bars from left to right represent people after seeing four examples, and after seeing 12 examples in the RANDOM (red) and HELPFUL (blue) conditions respectively. The rows of pixels within each bar constitutes a grey-scale representation of the generalisation data of individuals in that pattern and condition (see main text for detail). Both sample size and sampling assumption impact people's representation of the target category. (c) The response options for the questions that asked people about their response strategy (title added). There is a one-to-one mapping between the patterns shown and the representation associated with each response option.

but unlike Experiment 1 we also find an effect of the sampling manipulation. On an aggregate level, people in the HELPFUL condition tightened their generalisations ($BF_{10} > 10^6$) whereas those in the RANDOM condition did not ($BF_{01} = 31$). This suggests that the changed wording in the RANDOM condition, which provided a mechanism for potentially seeing a non-«Wuggam», helped to make the sampling cover story believable.

Representational analysis

Our primary question was whether people used different representations and whether their representations shifted in different conditions or with extra data. To address this, we used the six main generalisation patterns identified in Experiment 1, shown in Figure 2.3(a). They are each suggestive of qualitatively different mental representations: a one-sided decision boundary (Partition), a two-sided Range, several different kinds of non-contiguous regions (By Example, Any, Other), and an assignment of All test items to the category. Each participant at each test phase in Experiment 2 was then separately assigned to the most similar pattern using the model derived from the results of Experiment 1.

The results of this analysis are displayed in Figure 2.3. First, we note that the six patterns identified by our model are indeed roughly equivalent to the six self-report options offered during the test phase (shown in the column (c)).[4] This is clear when we compare the black lines in panel (a) on the left (which plot the average response for all people assigned to the relevant pattern) to the grey lines in the same panels (which plot the average generalisation curve for all people who chose the relevant self-report option). In most respects, the grey and black curves mirror each other very closely, illustrating that the data-derived patterns (based on classifications) and self-reported strategy are very similar.

Although the six patterns shown in Figure 2.3 are quite dissimilar to one another, there is a remarkable degree of within-pattern homogeneity, especially with respect to the first four patterns: most people assigned to a pattern do genuinely appear to be closely matching that pattern. This can be seen in Figure 2.3(b), which depicts a compressed grayscale representation of the raw responses for every participant within a pattern. Each panel shows three bars corresponding to one of the three possible conditions (4 exemplars, 12 exemplars RANDOM and 12 exemplars HELPFUL). The height of each bar captures how many people's generalisations best matched that pattern (thus, for instance, many more people matched the Range pattern in the HELPFUL condition than any other). Within each bar, every row of black pixels displays the responses of a single participant:

[4] Alignment of the self-report to the model-identified patterns was done based on our qualitative assignment, but we also performed all analyses using assignments based on RMSE fit (which differ from the qualitative assignments for 2 of the 11 clusters), or using the (somewhat noisy) self-report data directly. In all cases the conclusions are the same. Even collapsing Partition and Range into a single representation and the remaining representations into another produces a qualitatively similar pattern of results.

each row consists of 19 cells, each colour coded to represent the probability of assigning the relevant test item to the «Wuggam» category. For instance, an all black row occurs if all items are assigned to the «Wuggam» category, whereas a grey bar with a patch of black in the middle would represent a generalisation pattern where only the middle group of test items were labelled as «Wuggams».

Representational shifts

We are now in a position to address the central questions motivating this experiment. To what extent are changes in generalisation driven by a change in people's *representation* of the underlying category structure (e.g., shifting from Partition to Range), as opposed to learning the parameters of a representation (e.g., learning where the boundary in a partition lies)? Do sampling assumptions have an effect on how people shift their representations?

To investigate this, note that the panels in column (b) of Figure 2.3 are effectively bar charts, displaying the proportion of participants assigned to each of the six patterns, broken down by experimental condition. Visual inspection reveals marked differences as a function of sample size: when only 4 exemplars are observed, people are most likely to be assigned to the Any or Other patterns, whereas by the time 12 exemplars are observed the generalisation patterns are closer to Partition, Range or Other. Similarly there is evidence of a sampling effect: helpful sampling guides learners towards a Range representation whereas random sampling does not. A Bayesian contingency test (3 conditions × 6 patterns) finds strong evidence ($\text{BF}_{10} > 10^5$) for a difference in pattern assignments across conditions.

Looking more closely at these data, we can examine whether the sampling conditions each had a different impact on how people shifted their representations. To do so, we used the representation label assigned to each generalisation pattern (see Figure 2.3). If people were assigned to one pattern after seeing four examples and a different pattern after seeing 12 examples, *and* those patterns had different representation labels, then and only then would they be considered as having undergone a representational shift. Figure 2.4 plots the results of this analysis. It is clear that for those people who believed that examples were selected by a helpful teacher, additional exemplars led to an overall narrowing of generalisation, largely as a consequence of a representational shift.[5] A Bayesian ANOVA

[5] At first glance, Figure 2.4 appears to reveal a difference in generalisation between people in the RANDOM and HELPFUL groups at the point when only four examples have been observed and for which no sampling explanation was offered. But this difference is not reflective of the two conditions as a whole; rather it occurs only when the data is conditioned on representational shift. It reflects the fact that helpful sampling was interpreted more consistently than random sampling. Those people in the HELPFUL condition who already generalised narrowly after seeing only four examples, were less likely to narrow further upon observing additional examples, and thus more likely to maintain a stable representation; conversely, those who generalised more widely at first were more likely to change representation. This selection effect is not the case for those in the RANDOM group where representational shift was less consistent in direction.

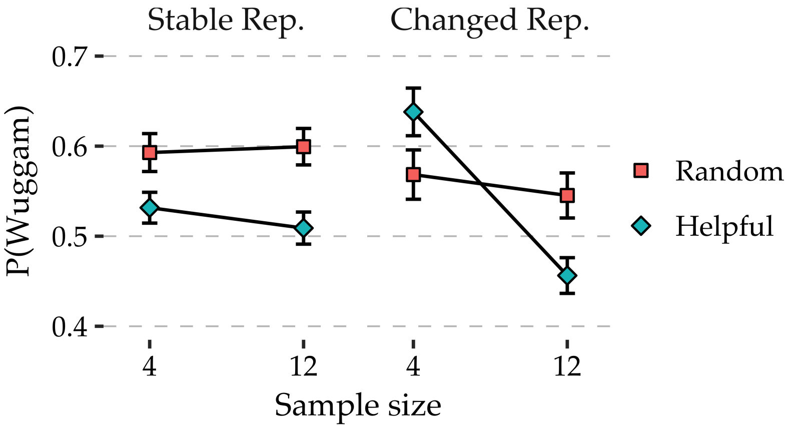


Figure 2.4: The mean effect of additional examples on the marginal probability of generalising the learned category to novel stimuli, as a function of sampling assumption and representational shift. Over half of the participants in Experiment 2 (N=119, from the RANDOM condition and N=111, from the HELPFUL condition) maintained a stable representation of the underlying category in response to observing an additional 8 examples, and showed little change in generalisation overall. Likewise, for people in the RANDOM condition who did undergo a representational shift (N=80). But for many people in the HELPFUL condition (N=94), the additional examples led to a representational shift resulting in a significant and consistent contraction in generalisation overall.

reveals strong evidence ($\mathrm{BF}_{10} > 10^6$) in favour of a model that includes effects of sample size, sampling condition, representational shift and interactions, tested against a null model which includes only the participant as a random nuisance parameter.

2.4 General discussion

The present work examines how people generalise a concept on the basis of learned examples. In a single experimental framework, we jointly considered two important considerations known to shape such generalisation: namely, people's assumptions about how the data was sampled, and their representation of the concept they seek to generalise.

In an initial between-subject experiment we found an effect of sample size consistent with other inductive generalisation tasks of this kind (e.g., Navarro et al., 2012; Vong et al., 2013). While there was no aggregate effect of sampling assumption, a post hoc analysis of individual responses revealed common patterns of generalisation suggestive of mental representations explored in the literature. These included non-contiguous regions (Nosofsky, 1986), a one-sided decision boundary (Ashby & Townsend, 1986), and a two-sided connected region (Tenenbaum & Griffiths, 2001a). This analysis also suggested that people's sampling assumptions might play a role in determining their representation of the category, a hypothesis we tested in a second pre-registered experiment.

The second experiment was based closely on the first but was within-subjects and involved a random sampling cover story that was slightly modified to be more suggestive of weak sampling. It replicated the effect of sample size and also found an effect of the revised sampling manipulation. Moreover, by linking response patterns identified in the first experiment to people's responses in the second, we found that observing additional examples causes some people to undergo a change in their mental representation. This shift drove much of their change in generalisation, and the nature and consistency of the change critically depended upon people's sampling assumptions.

In many ways our results are consistent with previous work finding that people tighten their generalisations when strong sampling holds but fail to do so when it does not (Ransom et al., 2016; Voorspoels et al., 2015; Xu & Tenenbaum, 2007a). This work has attributed such tightening to the operation of the size principle, which favours smaller hypotheses in a fixed (researcher defined) hypothesis space (Tenenbaum & Griffiths, 2001a). However, our results suggest that while the size principle may still be at work in some fashion, the truth may be more complex. Learning may in fact be operating on (at least) two levels in an hierarchical space, one with different representations (hypothesis spaces) at the top level and fixed hypotheses within each representation at the lower level. Behaviour that on aggregate looks like generalisation according to the size principle may actually reflect individuals shifting their representations more than individuals tightening their generalisations within the same representational space. An interesting line for future research would be to attempt to account for this behaviour using a hierarchical model that learns on both of these levels.

2.5 Conclusion

"Is this a dagger which I see before me...?"
– Macbeth

"That's not a knife. That's a knife."
– Crocodile Dundee

On the question of how best to classify sharp pointy things, great literary protagonists differ. And life, in this respect, may imitate art. Individual differences in representations, known to be driven by data, may be driven by sampling assumptions as well. By taking such differences seriously we have begun to understand that relationship; we hope that further research in this direction will continue to yield richer insights.

2.6 Acknowledgments

This work was supported by an Australian Government Research Training Program Scholarship (KR) and ARC Discovery Grant DP180103600.

2.7 Appendix A

This appendix describes data and analyses supplementary to the main text. It reproduces the material referred to in the original publication (Ransom, Hendrickson, Perfors, & Navarro, 2018) via an online supplement.

As discussed in the main text, we analysed the data from an initial experiment and a pre-registered replication and extension to that experiment. Experiment 1 involved a between subjects manipulation of both sample size and sampling related cover story. Experiment 2 manipulated the same factors, but the sample size manipulation was within subjects.

The data set

Each participant in Experiment 1 saw 19 distinct test stimuli a total of four times each. On each trial, participants would respond whether or not the given stimuli was in the target category. Thus, each set of responses we analyse for Experiment 1 is modelled by a 19 element tuple; each element in the tuple corresponds to one of the 19 test stimuli, and the value of the element represents the proprtion of times that the participant responded that the given stimuli was in the target category. Because the sample size manipulation in Experiment 2 was within subjects, each participant in that experiment contributed two sets of responses to the data set.

Cluster analysis – implementation details

Our analysis involved fitting a clustering solution to the data from Experiment 1 in order to determine whether participants' response data could be meaningfully separated into distinct patterns of responding. We modelled the data as a mixture of multivariate Gaussian distributions in 19-dimensional space, placing a Dirichlet process prior over the number of clusters. Having found a clustering solution by fitting the data from Experiment 1, we used the clusters obtained (each of which represents a differently parameterised 19-dimensional Gaussian) to analyse the response data from Experiment 2 and assign each set of responses to the cluster which best captures it.

In practice, we implemented this procedure using the `BayesianGaussianMixture` class from the `scikit.learn` module v0.19.1, under Python 3.6.3. The concentration parameter for the Dirichlet process was set to 1, the multivariate Gaussian distribution assumed a diagonal covariance structure, and the random seed was set to 1. There are a number of other minor parameter settings specific to the module used. These can be found in the python code shown in Listing 2.1 which we include for completeness. See http://scikit-learn.org/stable/modules/generated/sklearn.mixture.BayesianGaussianMixture.html for a full description of the `BayesianGaussianMixture` API and the various parameters used.

```python
import numpy as np
from sklearn import mixture

def initModel(random_seed = 1,
              nClusters = 20,
              concentration = 1,
              covariance = ’diag’):
    return mixture.BayesianGaussianMixture(
        covariance_type = covariance,
        weight_concentration_prior_type = "dirichlet_process",
        n_components = nClusters,
        init_params = ’random’,
        max_iter = 10000,
        tol = 1e-6,
        n_init = 100,
        random_state = random_seed,
        weight_concentration_prior = concentration)


def runModel(proportions1, proportions2):
    # ’proportions1’ represents response proportions for Experiment 1
    # from N=603 people, for each of the 19 test stimuli.
    assert np.shape(proportions1) == (603, 19)

    # ’proportions2’ represents response proportions for Experiment 2
    # from N=404 people, for each of the 19 test stimuli.
    # Note that in Experiment 2, each person was tested twice.
    assert np.shape(proportions2) == (404 * 2, 19)

    # Initialise the DPGMM model and fit to Experiment 1 responses.
    bgm = initModel()
    bgm.fit(proportions1)

    # Return the cluster weights and means and the predicted cluster
    # labels for all participants across both experiments.
    return
        bgm.weights_,
        bgm.means_,
        bgm.predict(proportions1),
        bgm.predict(proportions2)
```

Listing 2.1: Python code used to configure the Dirichlet Gaussian Mixture Model. The model was used to derive the clustering solution for participant responses described in the main text.

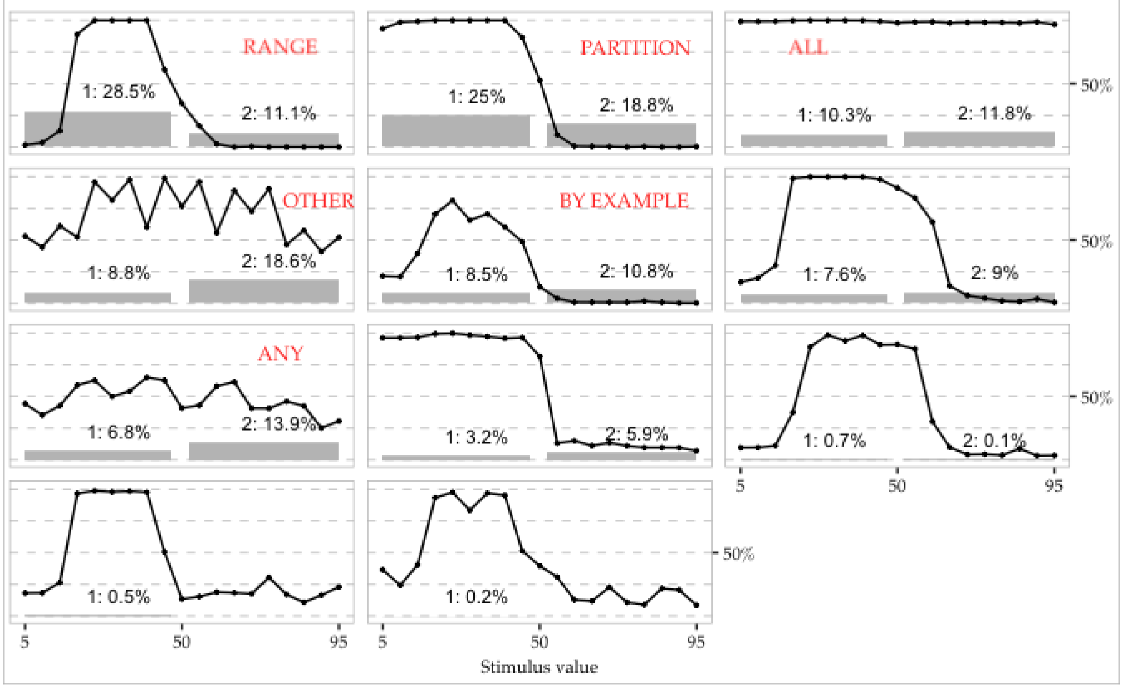


Figure 2.5: The complete clustering solution for participant response data derived from responses gathered in Experiment 1. Each panel shows the proportion of positive responses to the question: "Do you think this object is in the «Wuggam» category?", for each of the 19 test stimuli aggregated across all participants assigned to the given cluster. The two grey bars in each panel show the total proportion of response data assigned to the given cluster for Experiment 1 (left bar) and Experiment 2 (right bar). The analysis reported in the main text is focused on the six labelled clusters that capture the most data from Experiment 2. The labels correspond to those used in Figure 2.3 in the main text.

Clustering solution

Running the clustering model as described above, fitted 11 clusters to the data from Experiment 1. Subsequently, when these clusters were used to group response patterns from Experiment 2, only 9 of the 11 clusters received assignments. In the main paper, we focus our analyses on those 6 clusters that accounted for the largest number of participants from Experiment 2. For completeness, Figure 2.5 shows all 11 clusters found, along with the proprotions of response patterns that each cluster accounted for.

3 WHERE TO DRAW THE LINE

In Chapter 1, I presented an account of theoryful inductive reasoning that views generalisation and other forms of inductive inference as an inherently interpretive act. Under such an account, data encountered is not evidence *per se* but prospective evidence that must be interpreted in light of the learner's assumptions, including their sampling assumptions. And the act of adopting a sampling assumption can be viewed as one part of the process of theory selection – the part which involves reasoning about what can be learned from the data by determining what it is representative of.

In Chapter 2, I begun to explore some of the issues behind this form of reasoning. The study presented there involved a one-category generalisation task in which alternative explanations were given for how the training items were sampled. The experimental context aimed at stripping down much of the complexity of inductive reasoning. By employing "knowledge poor" stimuli and making the basis for generalisation clear, the intent was to greatly simplify the learner's theory selection process. In this way, any difference in patterns of generalisation between conditions might reasonably be attributed to people's theories about the sampling process. Further, by restricting the diversity of training stimuli, the aim was to isolate how the "evidence" of additional non-diverse samples is interpreted in light of different assumptions. The results suggest that people interpret such additional data more consistently when they have a clearer theory regarding its origin. Typically, when people believed that the extra information was representative of the to-be-learned concept they restricted their generalisation around the examples provided. When instead the origin of the data was unclear, as when some form of censoring was evident, the effect on generalisation was less predictable.

In this chapter, I extend this investigation to consider what happens when two categories are involved. How does the presence of a second category alter the patterns of generalisation observed? In particular, how does it change the way that additional non-diverse samples are interpreted? The work described here forms part of a broader investigation by Hendrickson et al. (2019)[1], which sought to understand how the inferential problems of categorisation and generalisation are related. A common technique used to study generalisation problems is to employ a one-category learning task. And a common method used to study categorisation is the two-category learning task. Viewed from the perspective of these tasks the problems seem closely related indeed. Nonetheless, the addition of a second category has significant potential to change the way that people

[1] The material that follows is based on my own contribution to the article (Experiment 3, Hendrickson et al., 2019, pp. 93–97)

reason. Notably, the one- and two-category judgment tasks may be distinguished by their decision complexity (Seger & Peterson, 2013). A test stimulus in the two-category task may represent neither, one or both of the categories being learned. In the case where both (or neither) category labels might reasonably apply to a novel item, the focus of the question changes to selecting which label represents the better fit. As a consequence people may recruit additional information when making a forced choice judgment between the two labels. Category base rates represent an obvious potential source of evidence when making such decisions. But just what it is that base rates are evidence of, is a question that the reasoner must address. A particular focus of this chapter is how that interpretation changes based on the context in which observations are made.

To my knowledge, the only previous attempt to investigate this question using a multiple-category design is a study by Vong et al. (2013) which found evidence to suggest that sampling assumptions can shape generalisation from multiple categories. The present study differs from that work in two key ways. Firstly, the experimental task employed here is more consistent with the kind used in the category learning literature. The cover story used in Vong et al. (2013) indicated that the data to be generalised were temperature observations at which two different kinds of bacteria were found alive. The temperature readings were depicted schematically, as a distribution of data points along a single dimension, following the paradigm adopted in Navarro et al. (2012). Here, as in Chapter 2, although the stimuli used vary in only a single dimension, it is the (two-dimensional) stimuli themselves that are displayed to participants. Presenting the stimuli as "objects" (albeit artificial ones) falling into two categories may have consequences in terms of how people interpret the evidence. Category labels, particularly when two or more are given to a seemingly related class of objects, typically exist for a reason. But that reason is not always clear, and as a consequence people might be expected to entertain different theories regarding just what it is that the categories might usefully distinguish. Indeed, as the analysis in Chapter 2 revealed, people did consider a variety of representations for the stimuli used. Secondly, the key classification question I use (following Hendrickson et al., 2019) is a neutral one: "Which category do you think this example belongs in?". In the two-category study of Vong et al. (2013) participants were asked only about "category A". It is conceivable that this framing encourages people to focus exclusively on that category, thereby increasing the tasks' resemblance to a one-category generalisation task.

3.1 Introduction to the experiment

The design of the two-category experiment closely resembles that of the one-category experiment described in Chapter 2 (Experiment 1). There is a single Four exemplar condition, where four examples of both categories A and B are provided without explanation, as well as two variants of a Twelve exemplar condition involving an

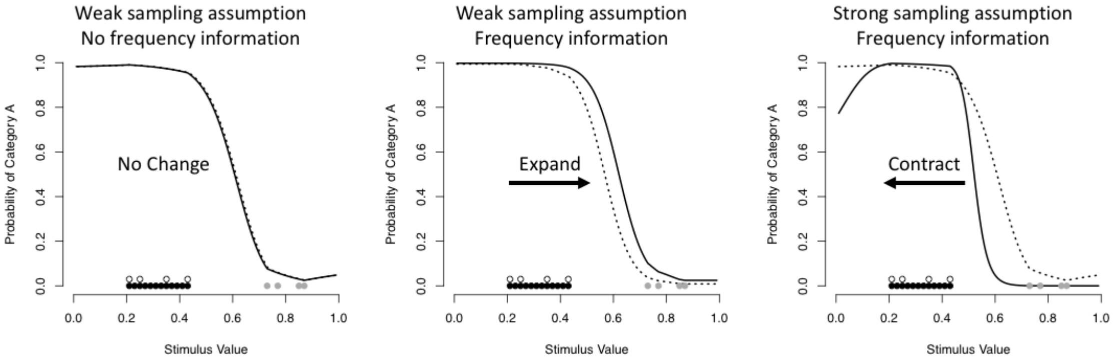


Figure 3.1: **Three possible patterns of results in the two-category learning experiment**. Each panel depicts a category boundary based on four observations of each of two categories (dashed line), and illustrates how that boundary may change as a result of seeing additional examples of a single category (solid line). The panels illustrate how three different assumptions lead to different qualitative effects. **Left panel**: Assuming that additional items are weakly sampled and that category base rates are uninformative, leads to no change in generalisation – the additional (non-diverse) items are effectively ignored. **Middle panel**: Under the assumption that additional items are weakly sampled but base rates are nonetheless informative, the boundary shifts towards the less frequent category. **Right panel**: On the assumption that the additional items are strongly sampled in a way that renders category base rates uninformative, generalisation of category A contracts, pulling the decision boundary towards it. The plots are reproduced from Hendrickson et al. (2019), where the corresponding prediction models are described in further detail.

additional eight examples of category A. The only difference between the two latter conditions concerns the cover story explaining how items are sampled – the set of training items used is identical. By suggesting that items are selected at random, independently from the category to which they belong, the TWELVE RANDOM condition is designed to induce a form of weak sampling. The TWELVE HELPFUL condition in contrast, is designed to induce a strong sampling assumption by encouraging the belief that items are chosen from a specific category by a helpful teacher.

In keeping with the goal of the experiment, the two cover stories used in the TWELVE exemplar conditions are also designed to suggest different explanations as to *why* there are more examples of one category than another. By purporting to select items at random, the TWELVE RANDOM condition is intended to promote the belief that the different sample sizes are reflective of the true category base rates. In contrast, people in the TWELVE HELPFUL condition are led to believe that the number of exemplars provided is simply a constraint imposed on the selections made by the helpful teacher and therefore not reflective of the true base rates.

By comparing performance across these three conditions, it is possible to examine the effect of additional exemplars on generalisation and to determine whether the nature of the effect changes depending on the learner's assumptions. In addition, by comparing people's behaviour against the predictions of a theoretically motivated computational model, further insight into the effects may be gained. Hendrickson et al. (2019) present a straightforward adaptation of the Bayesian generalisation model (Tenenbaum & Griffiths, 2001a) that is suitable for analysing the two-category decision task. The model assumes that underlying response strength is determined for each category independently, following the Bayesian generalisation model. Two plausible *Luce choice* decision rules (Luce, 1959) are suggested, one which incorporates inferred category base rates, and one which does not. Using the model to examine the effect of additional category A exemplars, each of the three qualitatively different patterns – no change, contraction, or expansion in generalisation – are all plausible predictions, as illustrated in Figure 3.1.[2]

Given the findings I report in Chapter 2, where a similar cover story was applied, one reasonable possibility is that the TWELVE HELPFUL sampling cover story is sufficient to lead people to believe that each of the eight additional category A items were strongly sampled from the category. In this case, people should consider that the sample distribution for category A is representative of the true distribution and should tighten their generalisations toward category A, relative to people in the FOUR condition (as in Figure 3.1, right panel).

A plausible alternative is that observing items drawn from multiple categories significantly increases the learner's uncertainty regarding the nature of the sampling distribution and its relationship with the distribution of interest (the joint distribution of labels and objects). Thus it is plausible that the two category task induces a weak sampling assumption regardless of the cover story. But uncertainty regarding the sampling distribution need not uniformly impact conclusions regarding the relevance of stimulus diversity and base rate information. In which case, the following two effects are predicted. Firstly, those people who observe twelve randomly sampled category A exemplars should be more likely to attribute the difference in sampling frequency to a genuine difference in category base rates, rather than ignore it as uninformative. As a consequence, these people should show a shift in the boundary away from category A when compared with people who see only four category A exemplars (as in Figure 3.1, middle panel). Secondly, people who believe that the eight additional exemplars were restricted to come from category A (but in a way that need not reflect category diversity) should show no change in their generalisations toward category A relative to people in the FOUR condition (as in Figure 3.1, left panel).

[2] There are of course alternative explanations for each of the three qualitative patterns, based on alternative sampling assumptions, for example. However, the three explanations given arguably represent the most theoretically motivated explanations.

3.2 Method

Participants

I recruited 364 participants for this experiment via Amazon Mechanical Turk. Of these, 20 people were excluded from participation, having taken part in previous experiments employing the same stimulus materials. No results were collected from 31 people who failed to complete the experiment. A further 15 people were excluded from further analysis for failing to reach a predefined accuracy threshold.[3] Data from the remaining 298 participants were included in all subsequent analyses. Participants ranged in age from 18 to 68 (median age: 32), 39% were female, and 98% of participants were from the USA. Participants were paid $USD 1.25 for taking part in the 7 minute experiment.

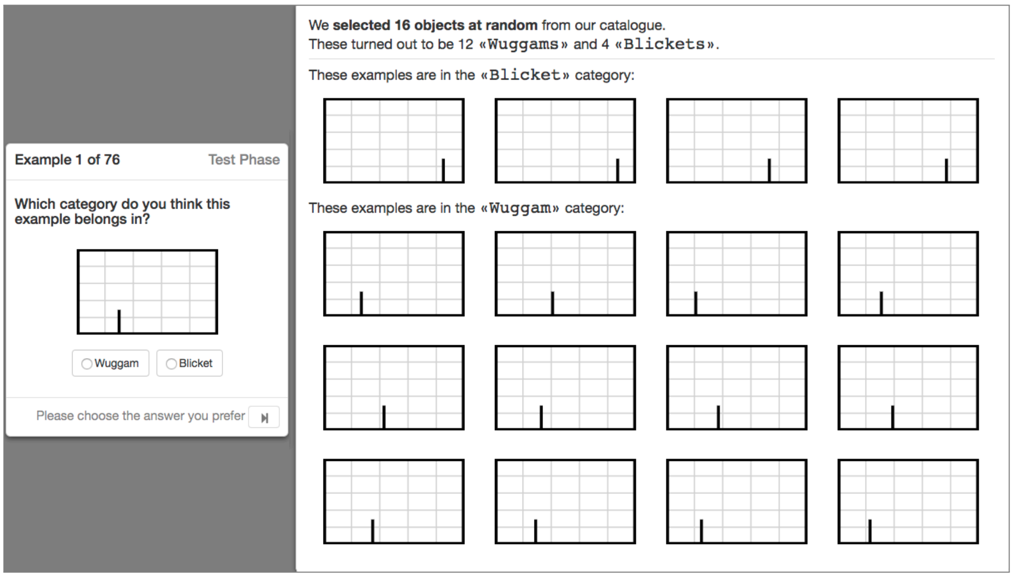
Design

People were allocated at random to one of three conditions. People in the Four condition ($N = 96$) were shown four exemplars from category B and four from category A, with no explanation offered for how these examples were chosen. Likewise, participants in the Twelve Helpful condition ($N = 99$) saw four exemplars from each category for which no explanation was offered. However, in addition they saw a further eight exemplars from category A which, they were told, had been selected from the category by a helpful teacher. In the Twelve Random condition ($N = 103$), people were told that 16 examples had been chosen for them at random. The "random" selection always consisted of the four category B exemplars, and twelve exemplars from category A.

Stimuli

The stimuli were identical in general form and appearance to those described in Chapter 2, varying along a single dimension, corresponding to the horizontal position of the vertical line within the rectangle (referred to here as the *stimulus value*). In related experiments using the same stimuli (Hendrickson et al., 2019) the height of the line rather than its horizontal position was varied, and stimuli values reflected horizontally or vertically were used for randomly selected participants. There was no material impact of these variations upon judgments, so for simplicity of design, testing and exposition a single variation is used in this experiment. The full set of training stimuli included 12 examples of category A with stimulus values ranging from 21% to 43% in increments of 2%. For

[3] This threshold was adopted (prior to inspection of the data) from a related study described in Hendrickson et al. (2019, Experiment 1) in order to facilitate the comparison of results between the two experiments. Briefly, the criteria excludes those people who classified test items as category B, despite falling within the interpolation range of category A, on at least 20% of related trials.



We **selected 16 objects at random** from our catalogue.
These turned out to be 12 «Wuggams» and 4 «Blickets».

These examples are in the «Blicket» category:

These examples are in the «Wuggam» category:

Example 1 of 76 Test Phase

Which category do you think this example belongs in?

Please choose the answer you prefer

Figure 3.2: **Sample stimulus display in the two-category learning experiment**. The four training stimuli from Category B are displayed in the top row, while the lower three rows show the twelve Category A stimuli. A reminder of how the stimuli were sampled for the given condition (TWELVE RANDOM in this example) is displayed at the top of the screen. The panel to the left of the screen shows the test exemplar for a single trial displayed above the two response buttons.

category B, the training set comprised four stimuli with values of 73%, 77%, 85%, and 87%, respectively. The complete set of stimuli can be seen in Figure 3.2.

People in the TWELVE conditions saw all 12 category A examples, while those in the FOUR condition saw four, including the two extreme examples (at 21% and 43%) plus two others randomly selected from the remaining 10 examples. People saw the same four category B stimuli across all conditions.

PROCEDURE

Following the same basic procedure described in Chapter 2, the experiment consisted of a training phase where people were shown examples from two target categories, followed by a test phase where they were asked to decide which category novel items belonged to. Participants in all conditions were told that the purpose of the experiment was to see how well they could judge between two categories of similar looking objects.

Training. At the commencement of the training phase, participants were informed how examples would be selected. This explanation differed across the three conditions. People in the FOUR condition were told simply:

> We'll start by showing a few examples of each category, taken from our catalogues.

at which point the four category B and four category A exemplars were simultaneously displayed on-screen. Participants in the TWELVE HELPFUL condition were given this same introduction. However, after the initial exemplars were displayed, they were informed:

> **The computer has assigned you to experiment group «K8»**, so we're going to help you by showing you an additional «8» «Wuggams» **chosen by a helpful teacher** from our Wuggam catalogue.

After pressing a button to progress in the experiment, the additional eight category A exemplars were simultaneously displayed below the original exemplars. In the TWELVE RANDOM condition, people received the following explanation:

> **The computer has assigned you to experiment group «J16»**, so we'll start by **selecting «16» objects at random** from our catalogue. We'll classify the objects on-screen for you so that you have some examples to work with.

Following this explanation, participants in the TWELVE RANDOM condition were shown all 16 exemplars simultaneously. All subsequent instructions were identical across conditions.

Testing. The training stimuli remained on-screen during the testing phase, and were annotated with a reminder of how the stimuli were chosen. Participants in all conditions were tested on the 19 test stimuli one at a time in random order; this sequence was repeated four times. The test query asked: "Which category do you think this example belongs in?". People responded by clicking one of two on-screen buttons, named after the categories (see Figure 3.2).

3.3 RESULTS

The overall pattern of responses is consistent across all conditions. A clear boundary emerges, with people more likely to indicate that lower stimulus values were in category A, while higher values were in category B (Figure 3.3).

The TWELVE RANDOM and TWELVE HELPFUL conditions used different cover stories to explain how the training items were sampled. The first question of interest is whether people generalised differently on the basis of the explanation they received? I examine this by comparing both of the TWELVE conditions to the FOUR condition. People in the TWELVE RANDOM condition (top row of Figure 3.3, responses shown in black) are more likely to classify test stimuli as belonging to category A than people in the FOUR condition (response shown in gray). In contrast, people in the TWELVE HELPFUL condition (bottom row), who are told that the eight additional exemplars had been selected

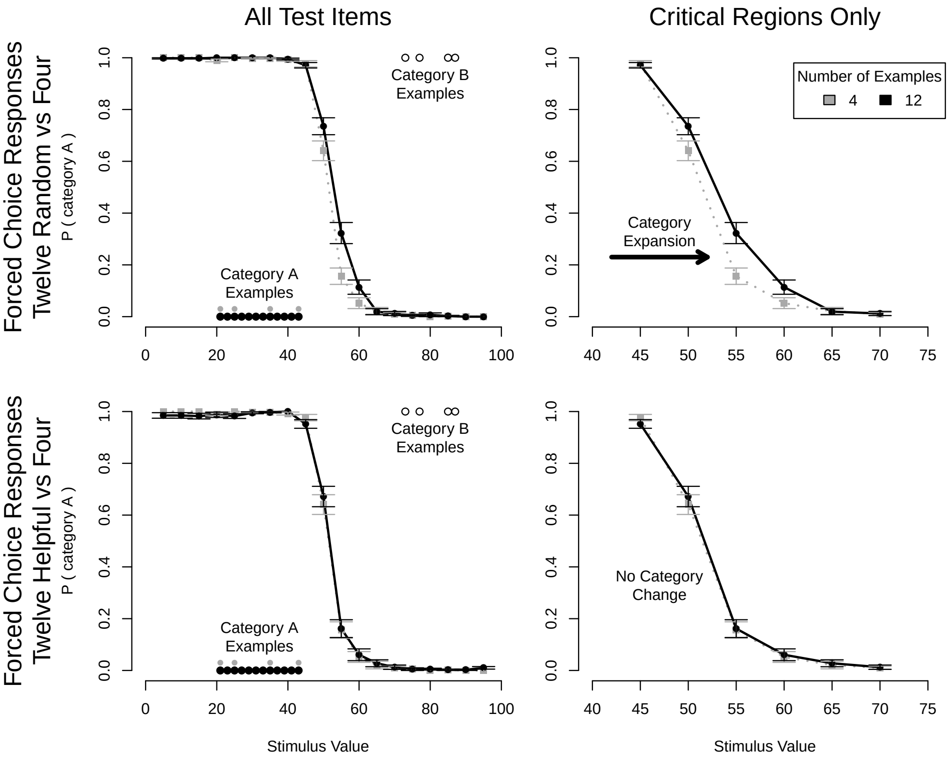


Figure 3.3: **Human performance in the two category learning experiment**. The graphs show the proportion of responses selecting category A in response to the question: "Which category do you think this example belongs in?" for each possible stimulus value at test. The graphs on the left show responses over the entire range of stimulus values; those on the right show responses for stimuli in the range between the two categories. Across all conditions, people saw four category B exemplars along with either four or twelve category A exemplars. Each row contrasts the performance of people who saw four category A exemplars (shown in gray) with one of the groups that saw twelve (shown in black): Twelve Random in the top row, and Twelve Helpful in the bottom row. The effect of additional category A exemplars differs depending on condition. People who were told that all exemplars are selected at random from a collection of objects were more likely to assign items to category A. In contrast, those who were told that an additional eight exemplars had been chosen by a helpful teacher, exhibited near identical responses to people who saw only four exemplars.

		Bayes Factor (relative to VALUE ONLY)		
Contrast	Best Model	VALUE ONLY	VALUE + CONDITION	INTERACTION
RANDOM VS. HELPFUL	INTERACTION	1 : 1	1.3 : 1	**7.2 : 1**
FOUR VS. TWELVE RANDOM	INTERACTION	1 : 1	2.5 : 1	**190 : 1**
FOUR VS. TWELVE HELPFUL	VALUE ONLY	**1 : 1**	0.11 : 1	2.8×10^{-4} : 1

Table 3.1: Comparison of how well three different linear regression models capture the results from selected conditions of the two category learning experiment. **Top row**: Contrasting the results from the TWELVE HELPFUL and TWELVE RANDOM conditions shows that the INTERACTION model, which includes both stimulus value and the condition (random instructions or helpful instructions), as well as an interaction term, best captures the data. **Middle row**: Similarly, the contrast between the FOUR and TWELVE RANDOM conditions is also best explained by the INTERACTION model which includes differences between the FOUR and TWELVE RANDOM conditions. **Bottom row**: In contrast, analysis of data from the FOUR and TWELVE HELPFUL conditions shows that the VALUE ONLY model, without any difference between the FOUR and TWELVE HELPFUL conditions, best captures the data. Bayes factors are expressed as odds ratios against the VALUE ONLY model reported to two significant figures.

from the category by a helpful teacher, show a similar pattern of response to people in the FOUR condition.

In order to quantify these effects, three contrasts were created based on subsets of the data. For each contrast, I calculate posterior odds for three different linear models. In the VALUE ONLY model, predictions are based on the stimulus value only. In the VALUE + CONDITION model, predictions are based on both stimulus value and experimental condition. Lastly, the INTERACTION model extends the VALUE + CONDITION model with a term that models an interaction between the two predictors. All models include the participant as a random effect. To determine which of the three models best captures the data for each contrast, Bayes factors for each model are constructed relative to the VALUE ONLY model.

In the first contrast, I compare performance between the TWELVE RANDOM and TWELVE HELPFUL conditions in order to quantify the extent to which the cover story manipulation changed the way that people interpreted observations. (top row of Table 3.1). This comparison favours the interaction model, indicating that the two conditions which differ only in their cover story, produce different generalisations.

The second and third contrasts compare performance in the FOUR condition with that of the two TWELVE conditions (middle and bottom rows of Table 3.1) in order to quantify the effect of additional exemplars on generalisation and how it varies depending on cover story. The results mirror the qualitative patterns in Figure 3.3 and show markedly different results for the two contrasts. The contrast between the FOUR and the TWELVE RANDOM conditions is best captured by the INTERACTION model that indicates a difference between conditions. This is not the case for the contrast between the FOUR and TWELVE HELPFUL conditions, which is best captured by the VALUE ONLY model, suggesting that for people who saw the the "helpful" cover story, there was no effect of additional category items.

3.4 DISCUSSION

The results of the present study indicate a shift in the boundary between two categories due to a cover story manipulation describing how training items were sampled. The finding is consistent with the notion that people interpret prospective sources of evidence such as sample variability and category base rates in ways which depend upon their theory of how the sample was produced. Compared with people who had seen four items from each category, people who saw an additional eight category A items purportedly selected by a helpful teacher, showed no change in generalisation overall. A similar comparison revealed that when people were told that all 16 items were sampled at random from an unknown distribution, this had the effect of expanding the more frequent category.

A Bayesian model of the two-category decision task (described in detail in Hendrickson et al., 2019) reveals a plausible computational account of these qualitative effects (see Figure 3.1). The "no change" found in the TWELVE HELPFUL condition matches a weak sampling model without base rates in the decision rule; this would imply that people effectively ignored both the lack of item diversity and the unequal base rates when making their decisions. The category expansion induced in the TWELVE RANDOM condition also corresponds to a weak sampling model, but one that assumes people *do* incorporate base rates into their decisions, at least when the base rate information is perceived to be reliable. In sum, the analysis supports the finding that people ignored the prospective evidence of stimulus non-diversity regardless of cover story, but ignored base rates only when they believed that the additional items were necessarily restricted to a single category. This suggests that people do not consider the representativeness of the sample as an all or nothing affair. Even when the origin of items is not entirely clear, a sample may appear representative in one way (in terms of category base rates, in this case) but not in another (sample variability with respect to a single property or dimension).

The work described in Chapter 2 as well as the present study represents an extension of the experimental investigation described in Hendrickson et al. (2019, Experiments 2 & 1, respectively). All experiments adopt the same basic training and test procedures and use a consistent set of stimuli. The important difference between the two pairs of experiments is that my own investigation has involved cover story manipulations of sampling assumptions, whilst the original pair did not. The consistency of the experimental frameworks across all four experiments thus supports some instructive comparison of findings, as summarised in Table 3.2.

				Sample seems representative?	
Experiment	#Cat.	Manipulation	Effect	Variability	Base Rates
Present	2	Random	expand	no	yes
Present	2	Helpful	no change	no	no
Hendrickson et al., Exp. 1	2	–	expand	no	yes
Ch 2, Exp. 1	1	Random	contract	yes	no
Ch 2, Exp. 2	1	Random (censored)	no change	no	no
Ch 2, Exp. 1 & 2	1	Helpful	contract	yes	no
Hendrickson et al., Exp. 2	1	–	contract	yes	no

Table 3.2: Summary of the present experiment and those described in Chapter 2, and Hendrickson et al. (2019, Expt. 1 & 2). The studies share a common experimental framework, and employ the same training and test stimuli for both category A and category B (where appropriate), supporting meaningful comparison of empirical results. Across experiments, increasing the sample size of category A from four to twelve exemplars had different effects on overall generalisation performance which was influenced by people's assumptions about the evidentiary value of the sample. When people believed that the relative lack of variability amongst sampled items was representative of the concept of interest, their region of generalisation tightened around the items in the sample. When people believed that the base rates observed were reliable, they used this evidence to expand the boundary of category A. And when neither form of evidence appeared reliable, generalisation performance was unaffected. Refer to main text for further discussion.

With respect to the two-category learning experiments, the shift in the category boundary observed between the TWELVE RANDOM and FOUR conditions in the current study closely resembles the shift observed in the corresponding FORCED CHOICE condition in Hendrickson et al. (2019, Experiment 1). This suggests two things about what people assume by default in the context of a two category learning experiment – i.e. in the absence of explicit guidance to the contrary (via cover story or sampling manipulation). Firstly, it suggests that people assume that the distribution over category labels observed in the sample is representative of the true distribution (or at least the one from which the test items are drawn). In contrast, it suggests that when more than one category of item appears in the sample, people discount the implicit negative evidence that a non-diverse sample might otherwise represent.

When learning from examples of a single category only, the default assumptions appear to be reversed. Evidence for this is suggested by the close match between the HELPFUL condition reported in Chapter 2 and the FORCED CHOICE condition in Hendrickson et al. (2019, Experiment 2). In both cases, there is significant tightening of generalisation around the range spanned by the training exemplars, consistent with the predictions of strong sampling. The lack of diversity amongst sampled items, whilst effectively ignored when two categories are presented becomes strong evidence in the one-category context. The same qualitative tightening was also found in the RANDOM condition in the first one-category experiment reported in Chapter 2, suggesting that a strong sampling assumption may have a strong prior probability in the one-category context. As Table 3.2 shows, across all related one-category experiments, only the TWELVE RANDOM condition (Chapter 2, Expt. 2), with its implication that the data was censored, appeared to weaken people's assumption that the sample variability was representative of the category. And while it seems intuitive that base rate information should be ignored by default when only a single category is presented, it is by no means a given. In a relatively small sample, the absence of a second category label (for example) may simply be evidence of a low base rate for that category.

Taken together, these findings suggest that while people will adapt their sampling assumptions when a reason to do so is made explicit, the number of categories observed in the learning context plays a dominant role in informing people's assumptions by default, and shaping their willingness to generalise beyond the examples observed. I return to further discuss these findings in a wider context in the closing chapter of this thesis.

STATEMENT OF AUTHORSHIP

TITLE OF PAPER	Exploring the role that encoding and retrieval play in sampling effects.
PUBLICATION STATUS	Published
PUBLICATION DETAILS	**K Ransom** and A Perfors (2019) Exploring the role that encoding and retrieval play in sampling effects. In A Goel, C Seifert, and C Freksa (Eds.) *Proceedings of the 41st Annual Conference of the Cognitive Science Society.*

Principal author

NAME OF PRINCIPLE AUTHOR (CANDIDATE)	Keith Ransom
CONTRIBUTION TO THE PAPER	Designed and ran experiments, performed data analysis, implemented model simulations, wrote manuscript and acted as corresponding author.
OVERALL PERCENTAGE (%)	80%
CERTIFICATION	This paper reports on original research I conducted during the period of my Higher Degree by Research candidature and is not subject to any obligations or contractual agreements with a third party that would constrain its inclusion in this thesis. I am the primary author of this paper.
SIGNATURE	
DATE	18/08/19

Co-author contributions

By signing the Statement of Authorship, each author certifies that:

1. the candidate's stated contribution to the publication is accurate (as detailed above);
2. permission is granted for the candidate in include the publication in the thesis; and
3. the sum of all co-author contributions is equal to 100% less the candidate's stated contribution.

NAME OF CO-AUTHOR	Amy Perfors
CONTRIBUTION TO THE PAPER	Supervised development of work, helped with experimental design and editing of the manuscript.
SIGNATURE	
DATE	18/08/19

4 LEARNING OR REASONING FROM EVIDENCE?

A growing body of literature suggests that making different sampling assumptions about how data are generated can lead to qualitatively different patterns of inference based on that data. However, relatively little is known about how sampling assumptions are represented or when they are incorporated. We report the results of a single category generalisation experiment aimed at exploring these issues. By systematically varying both the sampling cover story and whether it is given *before* or *after* the training stimuli we are able to determine whether encoding or retrieval issues drive the impact of sampling assumptions. We find that the sampling cover story affects generalisation when it is presented before the training stimuli, but not after, which we interpret in favour of an encoding account.

4.1 Introduction

For most of the reasoning tasks with which we are routinely faced, it is impossible to draw conclusions that are logically entailed by what we know already. Instead, we must by necessity make inductive generalisations on the basis of the limited data we have. In order to make the most of that data, it is important to accurately assess its evidentiary weight – to recognise precisely what kind of generalisations it supports. Doing this assessment accurately depends on understanding the context in which it was observed.

To illustrate why, imagine that you need to buy a present for a colleague as a part of your workplace Secret Santa. You don't know this colleague that well, but while helping them move offices you see a box containing the CDs that they listen to while at work. Sensing an opportunity to re-gift an unwanted copy of *Taylor Swift*, you take a closer look. Upon realising that almost all of their collection consists of 80s Billboard Hits, you conclude that their musical taste is dated[1] and reluctantly decide that Taylor Swift is not for them.

Suppose, instead, that you had seen the exact same data (a box of CDs) but in the context of helping your colleague move their entire music collection – many dozens of boxes worth – and that box just happened to be the only open one. Now the same data is no longer quite so representative: instead of being a carefully culled and chosen set of

[1] The fact that your colleague still uses CDs may have told you this already.

favourites, it is one of many. Thus, it tells you much less about whether your colleague would like Taylor Swift.

As this example illustrates, knowing something about why one saw the data that one did (and not some other data) enables people to make more valid inferences. Put another way, being able to reason about the generative process behind a set of observations tells people about the weight of evidence that those observations supply. These assumptions about the generative process are often referred to as the *sampling assumptions* that people bring to inference problems. Different sampling assumptions appear to drive qualitatively distinct patterns of generalisation (e.g., Hayes, Navarro, et al., 2019; Hendrickson et al., 2019), support epistemic trust (Shafto, Eaves, et al., 2012) and epistemic vigilance (Landrum, Eaves, & Shafto, 2015; Ransom et al., 2017), fuel pragmatic implicature (N. D. Goodman & Frank, 2016), and promote accelerated learning (Shafto et al., 2014).

Despite this wealth of empirical support for the utility and importance of sampling assumptions in generalisation, little is known about either how they affect the encoding and retrieval of the data, or how they affect people's mental representations. Is the evidentiary weight of data under a given sampling assumption computed only at the point at which the data is later retrieved? Or is it encoded at the time of learning, thus shaping the underlying representation from the beginning? And how is inference affected as people's memories of the data begin to fade?

Using a single-category learning task, we explore these questions here for the first time. We manipulate both the sampling assumptions people make about the training data (via cover story) as well whether that cover story is available before or after learning. As we explain in the next section, if sampling assumptions affect generalisation at retrieval, we expect no difference in performance regardless of when the cover story was revealed. Conversely, if they affect how the data are encoded, we expect different patterns of generalisation depending on when the cover story was available.

Sampling assumptions and inductive generalisation

The Bayesian generalisation approach of Tenenbaum and Griffiths (2001a) provides a useful framework for our research question. In the context of our single category generalisation experiment, we are interested in how the learner decides whether or not to extend the target category c to a novel item y on the basis of previously observed examples x. Within the framework, this decision is assumed to be probabilistic, based on the available evidence. That is:

$$P(y \in c|x,s) = \sum_{h \in \mathcal{H}_c : y \in h} P(h|x,s) \tag{4.1}$$

where s represents the learner's assumption about the process generating the data x, and $\mathcal{H}_c$ represents the set of alternative hypotheses the learner considers concerning the true extent of the category c.[2] In other words, the evidence in favour of category membership is effectively combined across all hypothetical versions of the category containing the novel item. Using a straightforward application of Bayes' rule the term $P(h|x,s)$ may be expressed as:

$$P(h|x,s) \propto P(x|h,s)P(h). \tag{4.2}$$

This formulation assumes, for simplicity, that the learner entertains a single sampling assumption (i.e. $P(s) = 1$), which we presume was given to them by a cover story describing the generative process.

It is the likelihood function $P(x|h,s)$ that is critical for our current purposes. Substituting different likelihood functions into this system of equations yields different predictions about the way that people generalise from given data. For instance, strong sampling implies a likelihood that embodies the size principle, such that each subsequent datapoint serves as evidence to further tighten one's generalisations around the data; weak sampling uses a different likelihood which implies no such tightening (Tenenbaum & Griffiths, 2001a). Thus, the likelihood may be thought of as representing different ways of calculating the weight of evidence that the data provides for the hypothesis under a given sampling assumption.

Our first question here is *when* the likelihood is calculated: when the data is first encoded, or when it is retrieved? If learners do not need to rely on their memories and the sampling cover story is available from the beginning, it is impossible to disentangle these two possibilities. However, if we manipulate when participants are aware of how the data were sampled (i.e., before or after learning), then different possibilities yield different predictions. We consider two main possibilities in detail.

Retrieval. If the likelihood is calculated upon retrieval, then encoding need only involve storing the raw data x in some form. The likelihood calculation would be shaped by whatever sampling assumption was in play during retrieval, regardless of what was assumed during learning. In this sense, the calculation would resemble the conventional or "idealised" interpretation of the Bayesian generalisation model. However, while the conventional interpretation assumes perfect recall of exemplars, a failure to retrieve some data would imply that the likelihood calculation was effectively over a reduced dataset (i.e., smaller sample size). The precise effect that this has will depend on the sampling assumption and on the particular items forgotten. For example, if the diversity of the dataset is largely unaffected by the failure to retrieve certain items, then generalisation under a strong sampling assumption should be wider in this case than under perfect recall. Under weak sampling, in contrast, it is the diversity of the sample and not its size that has an effect on generalisation; thus, a reduction in sample size without a change in diversity would mean that generalisation was unaffected. More generally, as the level of retrieval

[2] In the case that the data x varies over a continuous dimension $\mathcal{H}_c$ will represent a continuum of hypotheses and the sum is replaced with an integral.

failure increases, the Bayesian model predicts generalisation increasingly in line with the prior distribution.

Encoding. If the likelihood is calculated upon encoding, then the strength of evidence that it represents would have to be stored in some way. In this case, the precise effect of later retrieval failure might vary depending on *how* evidence is encoded. For example, if evidence is stored and retrieved with each exemplar individually then failure to retrieve a given exemplar would mean that subsequent generalisation operates over a smaller dataset, as in the retrieval account (although, unlike the retrieval account, using the sampling assumption that was in play at the time of encoding). If instead, evidence were stored and retrieved in aggregate form (via the hypotheses, for example) then failure to recall any particular exemplar need not imply that the associated evidence was lost. In this way, generalisation might still proceed with all the available evidence (presuming the same hypotheses were accessed). The details of representation notwithstanding, if the likelihood is calculated and stored during encoding, and not at retrieval, then generalisation would be shaped by the sampling assumptions available during learning, even if those assumptions are changed at retrieval.

4.2 Method

Our experiment involved a single-category generalisation task modelled on previous work demonstrating that sample size and sampling cover story affect people's willingness to extend category membership to novel examples (Hendrickson et al., 2019; Ransom et al., 2018). Although we employed stimuli identical to those used in that experiment, we modified the method of presentation so that each stimulus was removed from screen after a (typically brief) period of self-paced study. Using a consistent experimental framework allows us to directly compare our results with the previous findings, and thus to determine if the effect of sampling assumptions on generalisation changes as the memory of training examples decays.

One of our manipulations involved the nature of the cover story people received. Either they were told that the data was given by a Helpful teacher (which corresponds to a strong sampling assumption and implies that generalisations should be tighter) or they were given a cover story implying that it was chosen at Random (which corresponds to a weak sampling assumption and implies that generalisations should be looser). Critically, we manipulated whether people were given the sampling story Before or After they saw the training stimuli. If sampling assumptions affect how the data are encoded then people should generalise differently depending on when they received the story.

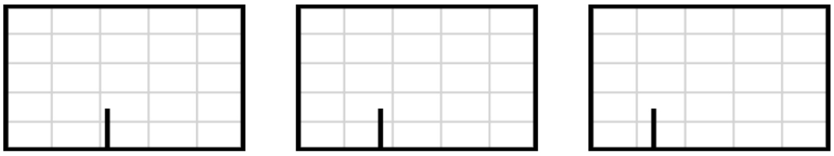


Figure 4.1: **Example stimuli**. Items varied only in the position of the short black vertical line along the bottom edge of the rectangle.

Participants

We recruited 999 people via Amazon Mechanical Turk who were each paid $1.70USD for 5-10 minutes participation. 56% were female, with age varying between 18 and 75 (median: 37 years), drawn predominately from the U.S. population (99%). All participants passed a screening for English language competency prior to participation.

Stimuli

Stimuli were black rectangles containing a vertical black line inside, attached to the bottom edge (see Figure 6.3). They varied along a single dimension (the *stimulus value*): the horizontal position of the line within the rectangle. Participants were told that this was the way in which stimuli varied. Evenly spaced light grey "guide lines" were drawn within each rectangle in order to improve discriminability. There were 12 training stimuli in total, whose stimulus values ranged from 21% to 43% in increments of 2%. They were divided into two sets corresponding to the two training phases, as described below.

Design and procedure

As shown in Figure 4.2, our experiment employed a $2 \times 2 \times 2$ mixed factorial design. Two factors (Sampling Explanation and Presentation Sequence) were manipulated between-subjects while another (Sample Size) varied within-subject. People were thus allocated at random to one of four experimental groups.

Across all groups, the experiment involved presenting people with a number of examples of a novel 1D category and then observing whether they generalised category membership to new items based on the examples they had been shown and what they had been told about those examples.

Sample Size

To facilitate a baseline against which the effect of additional exemplars could be compared, the experiment involved two rounds of testing. The first (Size 4) occurred after a training phase involving four training examples, and the second (Size 12) after seeing eight more.

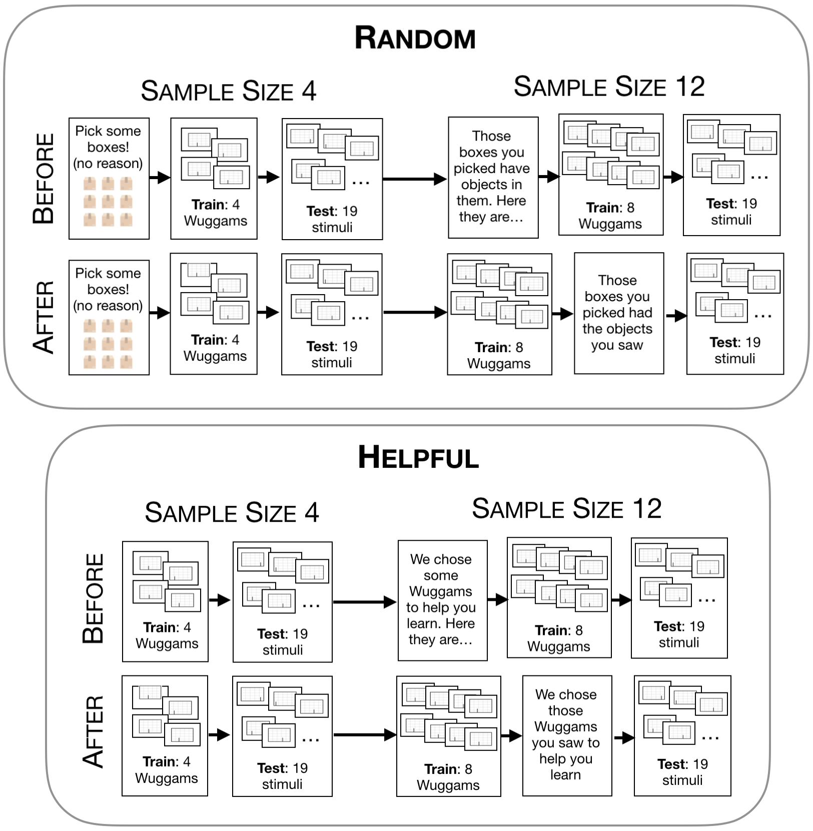


Figure 4.2: **Experiment design**. Our 2x2x2 design varied Sample Size within-subject and Sampling Explanation and Presentation Sequence between-subjects. All participants began by seeing four individually-presented exemplars followed by a generalisation task to novel stimuli. Those in the RANDOM condition were then given a cover story in which the subsequent eight items were chosen at random from boxes that they themselves had previously selected. Those in the HELPFUL condition were told that the items were selected by a helpful teacher. In the BEFORE condition, the cover story was given before seeing the eight new items; in the AFTER, it came after. In all conditions the experiment ended with a repeat of the generalisation test.

Stimuli for the first training phase consisted of the two extreme examples (with values of 21% and 43%) and two others selected at random from the ten whose values lay between the extremes. The eight remaining stimuli formed the second training set and were presented in random order.

Presentation Sequence

This between-subjects manipulation varied when the sampling cover story was presented in relation to the second training set. People in the BEFORE condition were told the cover story (RANDOM or HELPFUL, described below) *before* viewing the second set of training items, while people in the AFTER condition were offered the explanation only after all training items had been presented.

Sampling Explanation

The other between-subjects manipulation varied the details of the cover story explaining how the data in the second training phase were generated. The initial training phase, however, was identical for all participants. No explanation was given for how the exemplars were chosen. People were told only that the purpose of the experiment was to see how people judged whether or not unfamiliar objects were in the same category as known examples. In the second training phase people were given one of two different cover stories explaining how the items were selected.

Helpful. People in the HELPFUL condition were told:

> We have a bunch of boxes containing examples of the full variety of «`Wuggams`». We have chosen 8 of these boxes especially to help you learn the «`Wuggam`» category, bearing in mind the four training examples we showed you originally.

at which point an array of eight icons resembling open packing boxes were displayed in an adjacent panel. Participants in the BEFORE condition then viewed the eight stimuli one by one. Those in the AFTER condition saw the identical explanation (with verb tenses adjusted) only after all eight stimuli in the second training phase had been shown.

Random. The RANDOM condition was designed to encourage people to believe that each training item was selected at random and that it was at least theoretically possible to see examples not in the target category. To achieve this, people in the RANDOM condition were presented with an additional phase preliminary to the first training round. In this phase, a 6×5 arrangement of packing boxes was displayed on screen, and people were asked to select boxes in any order (but not told why this was necessary). After selecting 11 boxes, people were told that the contents would be revealed later in the experiment. Following this, the first training phase commenced, which was identical for all participants.

During the second training phase, participants in the AFTER condition were immediately shown the eight remaining training items without explanation. Those in the BEFORE condition were told that we had many boxes containing examples from our catalogue, and that these examples included but were not limited to Wuggams. After this, the original array of (closed) boxes was displayed, indicating the ones that the participant had previously selected. People were then told:

> At the start of the experiment we asked you to choose some of these boxes at random. These are the boxes that you selected. We're going to open them now and show you whatever kind of item we find inside.

In order to reinforce the notion that it might have been possible to see items from categories other than Wuggams, the display was updated at this point to reveal eight open boxes and three closed ones. People were told that some of the boxes they had chosen were stuck but that we would show them the contents of the boxes that did open. Participants in the AFTER condition received exactly this cover story (with verb tenses adjusted) only after seeing all eight training examples.

Generalisation test

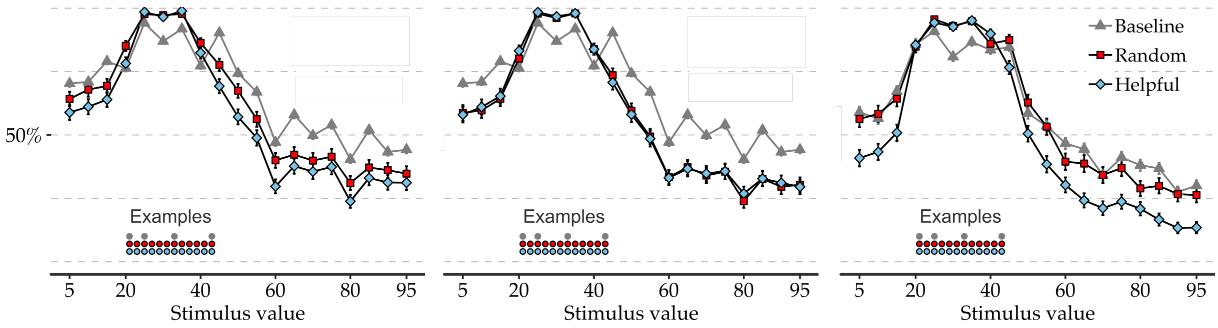
Immediately after both the first and second training phase, participants in all conditions performed the same generalisation test. In it, they were shown 19 stimuli one at a time in random order; this sequence was repeated four times. The stimuli consisted of 19 items with stimulus values ranging from 5% to 95% in increments of 5%. The test query was a yes or no question: "Do you think this object is in the «Wuggam» category?" Neither training stimuli nor the sampling explanation remained on-screen during testing, requiring people to rely on their memory when making judgements.

4.3 Results

Our work is focused on understanding how memory and sampling assumptions interact to affect generalisation. Do we replicate previous findings showing that differences in sampling assumptions lead to differences in generalisation? Does this difference in people's patterns of generalisation change if the sampling manipulation occurs before or after stimulus encoding? We address each question in turn below.

First: do we replicate previous results? Our RANDOM BEFORE and HELPFUL BEFORE conditions are very similar to that of a previous study (Ransom et al., 2018), but are different in one key way. In our version, the training stimuli were removed from the screen after initial presentation; in Ransom et al. (2018) and much of this literature the training stimuli stay visible for the entire experiment. We therefore investigate whether these previously observed effects of sampling manipulation are replicated even when people must rely on their memory of the training stimuli.

To investigate this we first analysed the responses of all participants having seen only the first four exemplars, for which no sampling explanation was given. Against this baseline we separately compared the responses of people in the RANDOM BEFORE and HELPFUL BEFORE conditions. The resulting generalisation curves shown in Figure 4.3(a) reveal that the HELPFUL sampling manipulation led to tighter generalisation than the RANDOM manipulation. This replicates a key finding of Ransom et al. (2018), shown in Figure 4.3(c). To examine the strength of evidence for this finding we analysed generalisation curves for the second test phase (Size 12), calculating the generalisation



(a) BEFORE.

(b) AFTER.

(c) Ransom et al. (2018).

Figure 4.3: Performance on a one category generalisation task as a function of presentation sequence, sampling procedure (manipulated between-subjects) and sample size (manipulated within-subject). The graphs show the proportion of positive responses to the question: “Do you think this object is in the «Wuggam» category?” for each of the test stimuli. People’s performance after seeing four examples of the target category with no sampling explanation given (grey line) is contrasted with their performance after seeing all 12 examples and being given an explanation of how the additional examples were selected (black lines). (a) When the sampling explanation was given prior to the presentation of the final 8 examples (BEFORE condition), people tightened their generalisations as more data was observed, but the extent of tightening was affected by the sampling manipulation; those people who actively sampled the additional examples at random (red squares) tightened their generalisation less than those that were told that the items had been selected by a helpful teacher (blue diamonds). (b) In contrast, when the sampling explanation was given only after all training stimuli were presented (AFTER condition), the sampling manipulation had no effect, with people tightening their generalisation equally in both cases. (c) Using the same experimental framework and stimuli, but keeping the training stimuli on-screen during the testing phase, Ransom et al. (2018) demonstrated the effect of sampling manipulation seen only in the BEFORE condition. But when people must rely on their memory of observed examples, their generalisation is wider overall.

probability for each person and stimulus separately. A Bayesian ANOVA revealed that a model of generalisation probability including stimulus value and sampling manipulation as predictors is strongly preferred to a model containing stimulus value only ($\text{BF}_{10} > 10^6$).

Although we replicated the qualitative difference between sampling conditions, it is evident on visual comparison of Figure 4.3(a) and (c) that people appeared to generalise further when they had to rely on their memory of the training stimuli. To determine the overall effect that this had on generalisation we calculated the marginal probability of extending category membership to novel items as a function of test phase (4 or 12 items) and sampling manipulation (RANDOM or HELPFUL). We then compared this probability between our experiment (the BEFORE conditions) and Ransom et al. (2018).

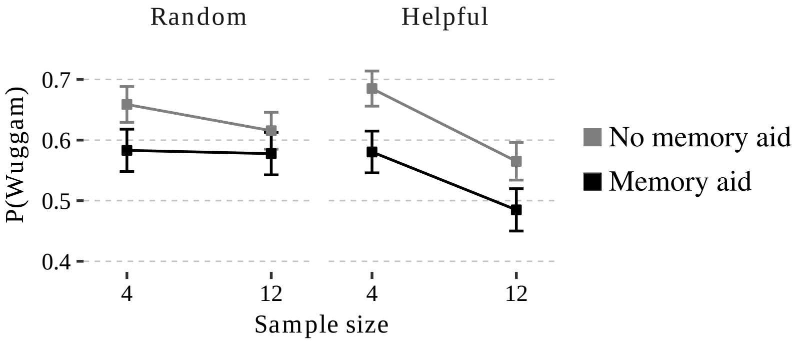


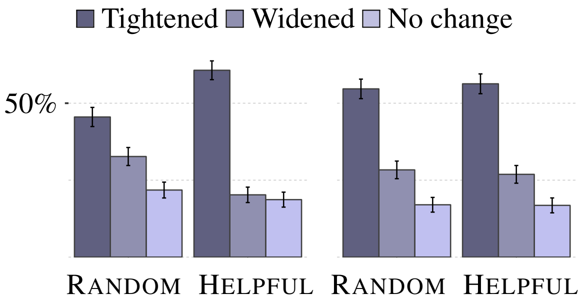
Figure 4.4: The mean effect of additional exemplars on the marginal probability of generalising the learned category to novel stimuli, as a function of sampling assumption and the presence of a memory aid. When training exemplars remained on-screen throughout the testing phase participants were less willing overall to generalise the target category to novel items than when no memory aid was present. In magnitude, the effect of the memory aid on generalisation was comparable to the effect of observing the eight additional exemplars.

The results, shown in Figure 4.4, demonstrate that the absence of a memory aid had a uniform but significant effect on generalisation overall ($\mathrm{BF}_{10} > 10^{100}$).[3] After seeing 12 exemplars, participants in our study (who had no memory aid) showed a willingness to generalise to novel items comparable to participants in Ransom et al. (2018) after seeing only four items that remained on screen throughout. Thus, overall, we find that the *difference* in generalisation according to sampling assumption did replicate, but generalisation was consistently higher when people had to rely on their memory more.

Our second question was whether the effect of sampling manipulation changes when the sampling cover story is given after the training stimuli rather than before. We therefore repeated our analysis for people in the RANDOM AFTER and HELPFUL AFTER conditions, and found that it does: there is no longer a difference in generalisation based on sampling assumption. As Figure 4.3(b) shows, people tighten their generalisations to a remarkably similar degree across the two conditions, despite the fact that they had opposing sampling cover stories (Bayesian ANOVA now favours the model with stimulus value as the only predictor: $\mathrm{BF}_{01} = 42$).

To further assess the effect of our sampling manipulation on the qualitative patterns of responding, we compared each individual's responses between the two test phases, after seeing 4 and 12 exemplars. Figure 4.5 shows the proportion of people who either tightened, widened or showed no net change in their generalisation (marginalised across test items). Consistent with the patterns at the aggregate level, it is evident that the explanation given to participants regarding the source of the additional exemplars does

[3] Based on a Bayesian logistic regression comparing a model of yes/no responses that included stimulus value, sampling manipulation and memory aid as predictors to one without memory aid.



(a) BEFORE (b) AFTER

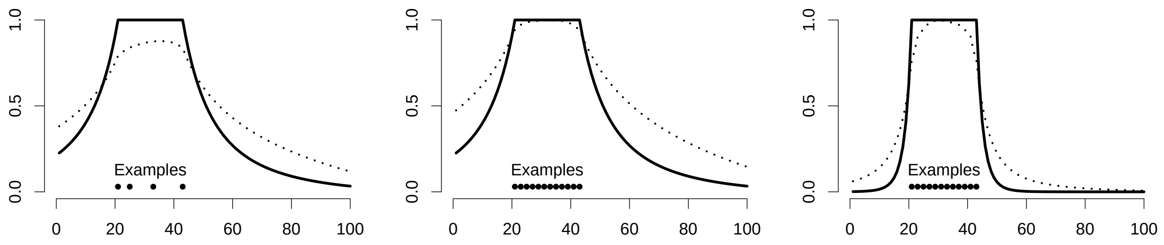
Figure 4.5: The proportion of people who either tightened ($\Delta_p < 0$), widened ($\Delta_p > 0$) or showed no change ($\Delta_p = 0$) in their region of generalisation, after seeing additional examples (where Δ_p reflects an individual's change in rates of responding in favour of the learned category). People are grouped according to the explanation they received about the sampling of extra items, and whether it was given before or after the examples themselves. Error bars show standard error of proportion. (a) In the BEFORE condition, where the sampling explanation was given prior to the presentation of the additional examples, the sampling manipulation had an effect. The majority of people who were told that the items had been selected by a helpful teacher tightened their region of generalisation, while the (slight) majority of people in the RANDOM condition, who actively sampled their own additional examples, widened their region of generalisation or showed no change. (b) In contrast, when the sampling explanation was provided after the additional stimuli had been presented (as in the AFTER condition), the majority of people tightened their generalisations regardless of the explanation given.

affect the trajectory of generalisation as more examples are observed. But this explanation only has an effect if it is given before the exemplars are observed ($\text{BF}_{10} = 300$) and not after ($\text{BF}_{01} = 2.8$).[4]

4.4 DISCUSSION

To our knowledge, our work here is the first to explore *when* sampling assumptions affect generalisation, and by extension when the likelihood is calculated. Our results demonstrate that the sampling cover story only had an effect when it was made explicit prior to the presentation of the data. When it was presented at retrieval, then whatever likelihood was the default at the time of encoding (which, in this case, appeared to have been strong sampling) was the likelihood that shaped generalisation – even though the cover story at retrieval should have contradicted it. While we cannot altogether rule out the influence of sampling assumptions at the point of retrieval, our experiment provides evidence in favour of an encoding account. Under this account, the evidence for different

[4] Bayes' factors are based on a multinomial logistic regression comparing a model of qualitative effect (tighten, widen, no net change) with sampling manipulation as a predictor against an intercept only model.



(a) Baseline. (b) Random. (c) Helpful.

Figure 4.6: Simulated performance on a one category generalisation task as a function of exemplar recall, sampling assumption and sample size. The graphs plot the probability of generalising the learned category as a function of stimulus value. Solid lines represent generalisation performance on the assumption that all exemplars are perfectly recalled at decision time – the default assumption of the Bayesian generalisation model. Dashed lines represent generalisation performance on the basis of imperfect recall. For illustration purposes, the simulation uses an independent probability of recall for each exemplar ($p = 0.5$). Failing to recall exemplars leads to wider generalisation overall. (a) Simulated performance in the Baseline condition (4 exemplars), assuming the default (strong) sampling. When the sample size is small, the effect of forgetting on generalisation reflects a balance of two forces: the reduction in diversity may reduce generalisation within the range spanned by the exemplars, while the reduced sample size leads to wider generalisation outside the range. (b) Simulated performance in the Random condition (12 exemplars), assuming the Baseline performance as a prior and that the 8 additional exemplars are weakly sampled. In the case of imperfect recall, the simulation predicts that the 8 additional items, although imperfectly recalled, lead to wider generalisation as a result of increased diversity. (c) Simulated performance in the Helpful condition (12 exemplars), assuming the Baseline performance as a prior and that the 8 additional exemplars are strongly sampled. Under strong sampling, generalisation tightens quickly around the sampled range with each extra exemplar, thus the predicted effect of forgetting is less in this scenario.

hypotheses is assessed according to the sampling assumption that prevailed at the time that the data were originally presented.

This finding has a variety of interesting implications. First, it suggests that there is no such thing as a "theoryless" learner: at no point do people simply encode the raw data in a veridical fashion. Rather, from the start they are actively engaged in making sense of it for future generalisation even though there is no current need to generalise. The question remains as to how automatic this is: would people be able to inhibit the likelihood calculation if requested to remember each specific data point as precisely as possible, or if they didn't think that a generalisation task would be forthcoming?

This has implications for effective pedagogy as well. It is known that learners benefit from assuming that their teacher is selecting the most informative examples possible given the learner's current beliefs. Such reciprocal assumptions can lead to a highly leveraged form of generalisation in which concepts can quickly be acquired from minimal input (Shafto et al., 2014). Under the idealised account of pedagogical learning, people's inferences should not depend on when the sampling process becomes apparent. However, our results suggest that it is important for the teacher to make the sampling process clear as early as possible.

In a similar way our finding has implications for how people process misinformation and corrections to misinformation. Ransom et al. (2017) found, for example, that people can use truthful but limited data in their efforts to mislead others by attempting to manipulate their counterpart's sampling assumption. Our work suggests that subsequently learning that an information source was biased may not be sufficient to correct the bias. It therefore offers another explanation for the well-established finding that retracting misinformation does not eliminate its influence (Ecker, Lewandowsky, Swire, & Chang, 2011; Johnson & Seifert, 1994). If people are encoding data in such a way that it cannot be disentangled from their theory at the time, interpreting that data under a new theory may be extremely difficult.

Another interesting aspect of this work regards the role of memory. By adopting the experimental procedure of Ransom et al. (2018) but requiring participants to view the simuli one-by-one, we were able to assess how memory decay would interact with sampling assumptions in shaping generalisation. We found that people tightened their generalisations less when they had to rely on their memory more. A simulation of the generalisation task used in our experiment verified our intuition that this should be the case (see Figure 4.6). Our finding is consistent with previous work using complex linguistic and non-linguistic data rather than a simple one-dimensional category (Perfors et al., 2014), which suggests that the result is reasonably robust.

Our memory manipulation (albeit across two experiments) also provides some basis to distinguish between two possible encoding accounts. One possibility is that evidence is stored and retrieved with each exemplar individually and any failure to retrieve an exemplar would mean that computation occurs over a smaller dataset. A second possibility is that evidence is stored in aggregate (across all data points) and retrieved via the hypotheses. In this case, the contribution of each exemplar would be accounted for at the point of encoding, and so the computation should proceed as if the full dataset were retrieved. The two possibilities suggest contrasting predictions. In the first case, we would expect generalisation in the present experiment to be wider than in the previous (Ransom et al., 2018, where perfect recall was supported). In the latter case, we should expect the results of the two experiments to be broadly in line with each other. As already noted, we found that manipulating how easy it was to remember exemplars did affect generalisation in a manner consistent with some degree of recall failure. We interpret this as weak evidence favouring the "exemplar encoding" account over the "hypothesis encoding" account: the data is stored in such a way that the strength of evidence is in some way integral to the encoding of the exemplar, at least to the extent that failure to later retrieve the exemplar equates to a failure to incorporate the associated evidence. Our evidence is only weak, however, because it is not entirely clear what "forgetting" in the context of the hypothesis encoding account would amount to. Fleshing out these distinctions more and testing them more systematically is a goal for future work.

While the present experiment should be taken in the spirit of a “proof of concept”, our research nonetheless suggests that memory, sampling, and generalisation are intertwined in ways that are still not fully understood. By manipulating when different information is available as well as the cognitive load during learning, it is possible to further illuminate this complex relationship.

4.5 Acknowledgements

Thanks to Simon Dennis for helpful comments regarding the initial concept behind this study. Thanks also to the anonymous reviewers for their helpful comments. This work was supported by an Australian Government Research Training Program Scholarship (KR) and ARC Discovery Grant DP180103600

Study II

WHEN THE BASIS FOR INDUCTION IS UNCLEAR

STATEMENT OF AUTHORSHIP

TITLE OF PAPER	Leaping to conclusions: Why premise relevance affects argument strength
PUBLICATION STATUS	Published
PUBLICATION DETAILS	**K Ransom**, A Perfors and DJ Navarro (2016). Leaping to conclusions: Why premise relevance affects argument strength. *Cognitive Science*, *40(7)*, 1775-1796

Principal author

NAME OF PRINCIPLE AUTHOR (CANDIDATE)	Keith Ransom
CONTRIBUTION TO THE PAPER	Designed and ran experiments, performed data analysis, implemented computational model, wrote manuscript and acted as corresponding author.
OVERALL PERCENTAGE (%)	80%
CERTIFICATION	This paper reports on original research I conducted during the period of my Higher Degree by Research candidature and is not subject to any obligations or contractual agreements with a third party that would constrain its inclusion in this thesis. I am the primary author of this paper.
SIGNATURE	
DATE	18/08/19

Co-author contributions

By signing the Statement of Authorship, each author certifies that:

1. the candidate's stated contribution to the publication is accurate (as detailed above);
2. permission is granted for the candidate in include the publication in the thesis; and
3. the sum of all co-author contributions is equal to 100% less the candidate's stated contribution.

NAME OF CO-AUTHOR	Amy Perfors
CONTRIBUTION TO THE PAPER	Supervised development of work, helped with experimental design, model development and editing of the manuscript.
SIGNATURE	
DATE	18/08/19

NAME OF CO-AUTHOR	Danielle Navarro
CONTRIBUTION TO THE PAPER	Supervised development of work, helped with experimental design, model development and editing of the manuscript.
SIGNATURE	
DATE	18/08/19

5 LEAPING TO CONCLUSIONS

Everyday reasoning requires more evidence than raw data alone can provide. We explore the idea that people can go beyond this data by reasoning about how the data was sampled. This idea is investigated through an examination of *premise non-monotonicity*, in which adding premises to a category-based argument weakens rather than strengthens it. Relevance theories explain this phenomenon in terms of people's sensitivity to the relationships amongst premise items. We show that a Bayesian model of category-based induction taking premise sampling assumptions and category similarity into account complements such theories and yields two important predictions: first, that sensitivity to premise relationships can be violated by inducing a weak sampling assumption; and second, that premise monotonicity should be restored as a result. We test these predictions with an experiment that manipulates people's assumptions in this regard, showing that people draw qualitatively different conclusions in each case.

Whereas formal deductive reasoning provides a solid bridge from premise to conclusion, everyday reasoning requires an inferential leap. But what assumptions support such a leap when raw data alone cannot? This question is relevant to the understanding of category-based induction, an important and representative form of inductive reasoning. In a typical category-based induction task, people are presented with a conclusion supported by one or more premise statements and asked to rate the strength of the inductive argument as a whole. Similarity-based models, which assume that argument strength is assessed on the basis of similarity between premise and conclusion categories, have successfully accounted for many aspects of people's performance in such tasks (Osherson et al., 1990; Sloman, 1993). Yet there are other characteristics of people's reasoning in this regard that are not adequately predicted on the basis of similarity. These characteristics have been explained as emerging from people's sensitivity to the relevance of different premises and the relationships amongst them (Medin et al., 2003).

In this paper we explain why and when premise relevance should matter. We argue that people's reasoning is sensitive to premise relationships because they consider the generative process behind the data they observe. If people made no such considerations, and instead assumed that all data consistent with the truth were equally likely to be observed (a so-called *weak sampling* assumption), then a perceived relationship amongst premise items should have no effect on argument strength. We demonstrate this by manipulating people's assumptions about premise selection, and observing that people draw qualitatively different conclusions as a result. Thus, we reproduce an effect of premise relevance on argument strength demonstrated by Medin et al. (2003) in a scenario where relevance should matter, and fail to observe the effect where it should not. Furthermore, we argue that the notion of *cognitive effect*, central to relevance theory explanations of induction, is neatly captured by a Bayesian theory of category-based induction that naturally incorporates different assumptions about premise sampling, along with the role of category similarity. Our results offer important corroborating evidence for the relevance theory of induction.

We first describe the category-based induction task, with a focus on arguments in which additional premises lead to weaker rather than stronger conclusions (known as *premise non-monotonicity*). We then describe a Bayesian analysis of this task which predicts that whether or not people exhibit premise non-monotonicity depends critically on how they assume the premises were generated in the first place. Finally, we present an experiment in which we manipulate these assumptions. As predicted by our model, people's reasoning differs qualitatively as a function of how they think the premises were sampled.

5.1 PREMISE MONOTONICITY AND NON-MONOTONICITY

In a typical category-based induction task participants are asked to rate the strength of inductive arguments like the following:

premise EAGLES have more than one fovea per eye.

conclusion HAWKS have more than one fovea per eye.

Here we use the notation EAGLES → HAWKS to indicate that this problem asks people to generalize a property from EAGLES to HAWKS.[1] Given that EAGLES and HAWKS are similar, participants might rate this as a moderately strong argument. Adding premises to an argument typically strengthens it, an effect referred to as *premise monotonicity* (Osherson et al., 1990). For instance, the argument {EAGLES, FALCON} → HAWKS appears stronger than EAGLES → HAWKS. The additional premise provides evidence that the property of *multiple foveae* should be extended to all birds of prey, and is not a property of EAGLES alone.

However, systematic violations of premise monotonicity have been observed. For example, Medin et al. (2003) found that people were less willing to endorse the generalization {GRIZZLY BEARS, BROWN BEARS, POLAR BEARS} → BUFFALO than GRIZZLY BEARS → BUFFALO, despite the former having more premises. This *non-monotonicity* effect appears to arise because the multiple premise argument provides

[1] More precisely, we might denote this EAGLES $\xrightarrow{\textit{mult. foveas}}$ HAWKS in order to emphasize the property being extended in the argument. For the most part this detail is not needed for our paper.

strong evidence that the property should be extended to bears only, and so weakens the plausibility that buffaloes share the property. This insight is captured in the relevance theory of induction, which suggests that adding premise categories should weaken an argument if the added categories reinforce a property shared by all of the premise categories but not the conclusion (Medin et al., 2003).

This seems sensible, but *why* is it so? If nothing can be assumed about the way premises are sampled, then there is no reason to expect a more relevant premise to be advanced in argument over a less relevant one; the notion that a perceived relationship between premise items represents the appropriate basis for induction, gains no special credence simply by virtue of being put forward. But in the real world arguments are rarely (if ever) constructed from randomly sampled facts. It makes sense for people to assume that arguments are constructed by sampling relevant facts to support conclusions and achieve communication goals. Wilson and Sperber's account of *relevance theory* (Wilson & Sperber, 2004) and Grice's *co-operative principle* (Grice, 1989), upon which their theory is based, each offer explanations for why utterances raise an expectation of relevance on the part of the listener. For Grice, the raised expectation comes about because people, for the most part, follow communicative conventions that encourage relevance. But such a heightened expectation should serve only to sharpen the ability to discriminate inputs on the basis of relevance. A reasonable variation in relevance should exist in the first place.

Wilson and Sperber go further than Grice, arguing that neither a communicative convention nor a communicative context are strictly necessary for an enhanced perception of relevance. A tendency to maximize relevance, they contend, is a fundamental feature of our cognitive systems, arising from the need to make the most efficient use of processing resources. To give an example, there are a number of theoretical results showing that positive evidence has stronger evidentiary value than negative evidence under plausible assumptions[2] about the environment (e.g. Klayman & Ha, 1987; Navarro & Perfors, 2011). Given this, maximizing relevance should lead people to prefer to give and to receive positive evidence, and will therefore treat positive premises (of the form "item *x* has property *p*") as more relevant than negative ones.

If people assume premises are sampled based on relevance then any property shared by the premises will gain plausibility as the correct basis for induction and stronger inferences to that effect should result. For example, if I want to convey the range of animals that share a particular property, and I want to be as relevant as possible, then I

[2] The critical assumption is that we live in a world in which most items do not have most properties. This seems intuitive (e.g., FOXES are *f*urry, but FISH, FEARS and FOOTPRINTS are not), but some care is needed in substantiating the point. From a logical standpoint any "sparse" property (possessed by a minority of entities) is mirrored by a "non-sparse" complement. However, they need not be equally salient nor equally useful when describing the world: people are more likely to think of *f*urry (sparse) as a meaningful property than *n*on-furry (non-sparse). Indeed, what Navarro and Perfors (2011) show is that in any world where entities are not completely homogeneous, the categories and properties that intelligent agents attend to should display this sparsity bias.

should select additional examples that best capture the appropriate range. Returning to the bears example, had I wanted to convey the message that many species had some property, not just bears, you might reasonably have expected me to mention a different kind of animal. So my choosing further examples of bears when extending my argument provides evidence that only bears have the property. Qualitatively, this reasoning explains why people exhibit premise non-monotonicity in this situation. This intuition can be reinforced quantitatively by the mathematics of Bayesian probability theory, as we explain in the next section.

5.2 A model for reasoning in category-based induction

Consider a standard Bayesian approach to category-based induction tasks (Heit, 1998; Sanjana & Tenenbaum, 2003). Suppose the learner is given a one premise argument of the form $x \xrightarrow{p} y$. Let h denote one possible hypothesis about how far property p should be extended, and $P(h)$ denote the reasoner's prior bias to think that h describes the true extension of property p. Having observed that item x possesses property p, the posterior degree of belief in h is given by Bayes' rule:

$$P(h\,|\,x) = \frac{P(x\,|\,h)P(h)}{\sum_{h'} P(x\,|\,h')P(h')}. \tag{5.1}$$

Here, $P(x\,|\,h)$ is the likelihood, which specifies the probability that the argument would have used x as a premise if h were the true extension of property p. The sum in the denominator is taken over all hypotheses that the reasoner might consider regarding the extension of property p. When an argument contains multiple premise items $x_1, \ldots, x_k$, the likelihood is given by the product of each of the individual probabilities, $\prod_{i=1}^{k} P(x_i\,|\,h)$. In order to evaluate the claim that item y also possesses property p, a Bayesian reasoner sums the posterior probabilities of all hypotheses that are consistent with the claim. Thus, the argument strength is given by:

$$P(y\,|\,x) = \sum_{h:y \in h} P(h\,|\,x). \tag{5.2}$$

This model has two components, the prior $P(h)$ and the likelihood $P(x|h)$. In our application of the model, the prior reflects the similarity amongst premise categories. As described in Appendix A, we use empirical similarity data to set $P(h)$ and simulations to check that the qualitatively important effects are not overly sensitive to the particular data collected.

The likelihood is critical to an understanding of when and why premise relevance matters: it naturally captures different assumptions people may make about how the premises were generated. For instance, a naive reasoner might assume that the premise items for an argument are selected at random from the set of true facts about the property

p. This is called *weak sampling*. Since the item x is chosen randomly, weak sampling allows premises to present negative evidence (i.e.,"item x does not have property p"). For a premise presenting evidence that item x has property p, the weak sampling likelihood function is:

$$P(x|h) \propto \begin{cases} 1 & \text{if } x \in h \\ 0 & \text{otherwise} \end{cases} \tag{5.3}$$

In essence, when presented with item x, a learner assuming weak sampling falsifies all hypotheses inconsistent with the premise but does not alter their beliefs in any other respect – the fact that x was chosen over other items has no additional relevance to the reasoning problem. As a result, such a learner will be less likely to demonstrate premise non-monotonicity. If premises are generated randomly, seeing BLACK BEARS in addition to GRIZZLY BEARS does not act as a "hint" that only bears have the property in question. Rather, because there are almost no hypotheses that could be falsified by the additional BLACK BEARS premise that were not already falsified by the GRIZZLY BEARS premise, the additional information is largely irrelevant.

The simplicity of the weak sampling model and its connection to falsification is appealing. However, as we have seen, it provides a poor description of how inductive arguments are constructed in everyday reasoning. If a learner expects an argument to be constructed using positive examples, then a weak sampling assumption is no longer tenable. A simple alternative is *strong sampling* (Sanjana & Tenenbaum, 2003; Tenenbaum & Griffiths, 2001a), in which a premise item is selected *o*nly from those exemplars that possess property p. As noted earlier, this restriction makes sense if people expect to receive relevant evidence. This gives the likelihood function

$$P(x|h) = \begin{cases} \frac{1}{|h|} & \text{if } x \in h \\ 0 & \text{otherwise} \end{cases} \tag{5.4}$$

where $|h|$ denotes the *size* of hypothesis h. In this context, the size is calculated by counting the number of items that possess property p assuming hypothesis h is true.

Under strong sampling, the item presented has relevance beyond falsification. That is, a premise provides more evidence for a small hypothesis than it does for a larger one (Tenenbaum & Griffiths, 2001a). A learner who sees multiple premise items consistent with one small hypothesis will come to prefer that hypothesis over other, broader hypotheses, even when the broader hypothesis happened to be originally preferred. As noted in previous research (Fernbach, 2006; Kemp & Tenenbaum, 2009; Voorspoels, Van Meel, & Storms, 2013), this phenomenon provides a potential explanation for why people sometimes exhibit premise monotonicity and at other times non-monotonicity.

Compare the one premise argument CHIMPANZEES → GORILLAS to the two premise argument $\{$CHIMPANZEES, ORANGUTAN$\}$ → GORILLAS. Both premises are consistent with a small hypothesis (i.e., that all primates have that property). Because gorillas are also primates the additional evidence provided by the ORANGUTAN premise acts to strengthen the argument: premise monotonicity is satisfied. In contrast, compare the one premise argument GRIZZLY BEARS → LION to the two premise argument {GRIZZLY BEARS, BLACK BEARS} → LION. In the one premise variant, the reasoner might reasonably believe that the property extends to all mammals or all predators, and so there is at least some chance that LIONS possess the property. However, when BLACK BEARS is added to the list of premise items, the reasoner has strong evidence in favor of a small hypothesis, namely that the property is common only to bears. This produces a non-monotonicity effect, since an additional positive observation acts to weaken the conclusion.

Importantly, this explanation relies on the assumption of strong sampling. It is this assumption that gives a premise item relevance over and above its use for falsification. In the bears example above, the effect occurs because a second bear premise provides strong evidence for the (smaller) "bears" hypothesis relative to the (larger) "all predators" and "all mammals" hypotheses, even though all three are consistent with both premises. Under a weak sampling assumption, premise items have no relevance beyond falsification, and this shift does not occur.

If strong sampling represents an assumption that premise selection is biased towards relevant items[3] and weak sampling represents the assumption that is is not, then it is reasonable to consider that the bias to expect relevant premises might vary not just in kind, but also in degree. A *m*ixed sampling model can be used to capture this situation in a straightforward way (Navarro et al., 2012). Under mixed sampling, the likelihood function becomes:

$$P(x|h) = \begin{cases} \theta\frac{1}{|h|} + (1-\theta)\frac{1}{|\mathcal{X}|} & \text{if } x \in h \\ 0 & \text{otherwise} \end{cases} \tag{5.5}$$

where $|\mathcal{X}|$ represents the number of possible premise items, and θ represents the probability that the premise item x was strongly sampled. When $\theta = 0$ the model is equivalent to weak sampling and has no bias towards positive evidence. In contrast, when $\theta = 1$ the bias is so extreme that the learner believes it is impossible to receive negative evidence, and the mixed sampling model becomes equivalent to strong sampling.

The notion that people are sensitive to how the premises were generated represents an intriguing and testable prediction. If reasoners have an expectation of premise relevance and thus expect premises to be biased towards positive evidence, they should show premise non-monotonicity for the bears example. If, on the other hand, they assume that premise items have been selected at random (i.e., weakly sampled), then premise monotonicity should be exhibited. Note that this prediction stands in contrast to the

[3] The strong sampling model is not intended to capture all the complexity of selecting items for relevance. For instance, richer pragmatic assumptions can be captured using pedagogical sampling models (Shafto et al., 2014). This complication is not necessary in the current context but some implications are addressed in more detail in the discussion.

predictions of similarity based models (Osherson et al., 1990; Sloman, 1993) neither of which incorporate any sensitivity to the mechanism by which the premises are generated. To that end, we present experimental evidence that premise monotonicity can be systematically manipulated by changing the assumptions people make about the origins of the data. Not only do we see qualitative reversals from monotonic to non-monotonic reasoning consistent with a change from weak to strong sampling, we also find that the transition occurs in a graded fashion consistent with the smoothly varying bias parameter in the mixed sampling model.

5.3 Experiment

Participants

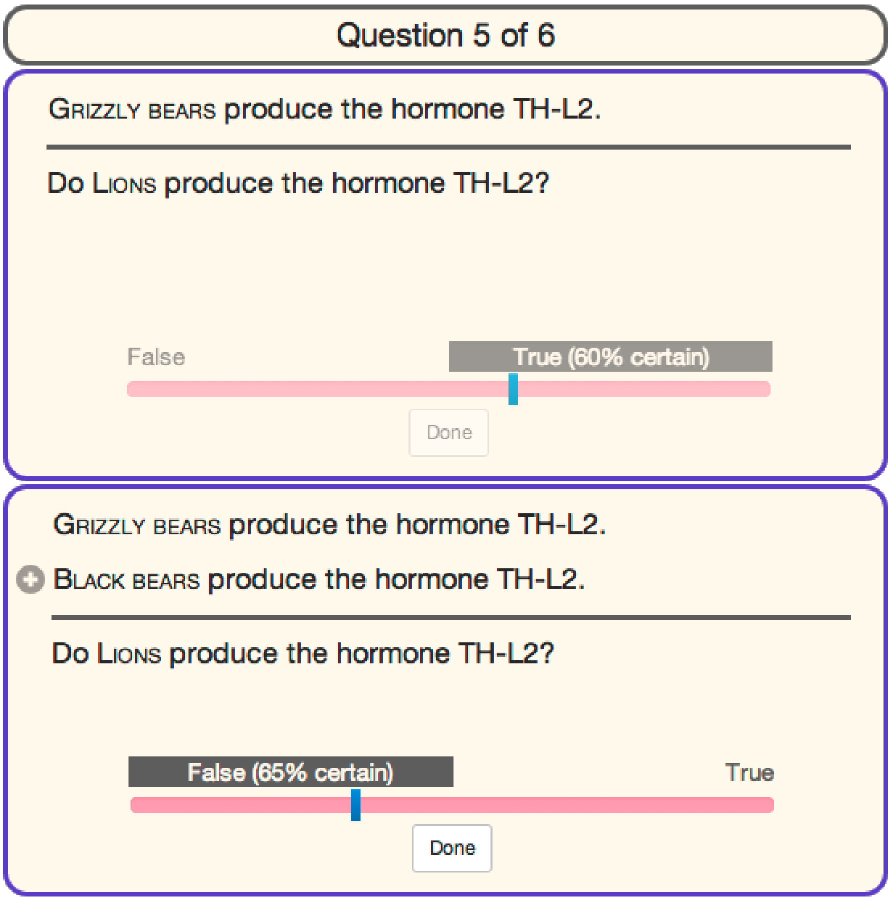
590 adults were recruited via Amazon Mechanical Turk, and were each paid $0.50 (USD) for the 5–10 minutes participation. 52 were excluded due to browser incompatibility, and the remaining 538 were aged 18 to 69 years (median age 28, 65% male). 500 participants were in the United States, with 38 located elsewhere.

Procedure

A cover story informed people that they would be making judgments concerning well established facts about the properties of animals. Each trial began by presenting a fact about one animal and then asking about a second. For example, they might first be told that grizzly bears produce the hormone TH-L2, and then asked whether lions also produce TH-L2. Responses were collected using a slider bar that allowed people to produce answers ranging from “100% false” to “100% true”, as shown in Figure 5.1. They were then told about a second animal, and allowed to revise their original judgment by moving a different slider. The dependent measure for each trial is the difference between these two judgments. If the endorsement of the conclusion is stronger on the second occasion, premise monotonicity is satisfied. If the difference is negative, non-monotonicity is observed.

Conditions

Participants were randomly assigned to one of the four conditions, each involving a different combination of cover story and filler trials. The cover story informed people about how the second fact in each trial was generated, while the filler trials were designed to be consistent with either a strong or weak sampling assumption. In the Both Relevant condition, participants were told that the extra facts were provided by



Question 5 of 6

Grizzly bears produce the hormone TH-L2.

Do Lions produce the hormone TH-L2?

True (60% certain)

Grizzly bears produce the hormone TH-L2.

Black bears produce the hormone TH-L2.

Do Lions produce the hormone TH-L2?

False (65% certain)

Figure 5.1: An illustration of the on-screen presentation of a trial shown at the point where the second premise has been revealed. The one premise argument is displayed on the upper portion of the display, while the two premise form is on the lower portion. The rectangular "slider" (disabled on the upper portion, enabled in the lower) allows participants to respond "True" or "False" and indicate the level of certainty in their response.

past players of the game who were trying to select a helpful example of an animal with the property in question. The story and the filler trials were designed to promote the idea that facts were chosen on the basis of relevance, similar to strong sampling. In the BOTH RANDOM condition people were told that they would select a card from a deck displayed face down on-screen. This card would disclose whether or not a particular animal had the property in question. In contrast to the BOTH RELEVANT condition, the story and filler items were designed to support the assumption of weak sampling by encouraging the belief that facts were being sampled at random. To allow us to investigate whether the premises alone had an effect on sampling assumptions we ran two further experimental conditions. The RELEVANT FILLERS condition employed a neutral cover story giving no information about how the premises were selected, and used the same filler items as the BOTH RELEVANT condition. Likewise, the RANDOM FILLERS condition employed a neutral cover story, but used the same filler items as the BOTH RANDOM condition.

STIMULI

All participants were presented with six trials in a fixed order,[4] as shown in Table 5.1. Three of these were especially key and appeared in all conditions. There were two target arguments structured so that they should elicit non-monotonic responding under a strong sampling assumption (**T**arget 1: {TIGERS, LIONS} → FERRETS; **T**arget 2: {GRIZZLY BEARS, BLACK BEARS} → LIONS). There was also a **C**ontrol argument designed to elicit monotonic reasoning under any mixture of weak or strong sampling ({CHIMPANZEE, ORANGUTAN} → GORILLA). Finally, each person saw three **F**iller trials, designed to reinforce a particular sampling assumption. Consistent with strong sampling, the filler trials in the BOTH RELEVANT and RELEVANT FILLERS conditions consisted solely of positive examples. In contrast, the filler trials in the BOTH RANDOM and RANDOM FILLERS conditions included negative examples as well, and appeared much more random.

5.4 RESULTS

MODEL PREDICTIONS

The Bayesian model of strong and weak sampling described in Equations 5.1–6.4 was used to quantitatively predict how a reasoner holding either assumption would reason about the two **T**arget arguments and the **C**ontrol argument. In order to extract these

[4] Randomization of trial order would not have made sense in this context. Because the filler items were an important part of the experimental manipulation, it was critical that at least some of these precede the target items; because we did not want the design to be too obvious, we also wanted to include at least one filler in between the two targets.

			Additional example	
Trial	Property to be generalized	First generalization	RELEVANT	RANDOM
Filler 1	have more than one fovea per eye	EAGLES → DOVES	+HAWKS	−TORTOISES
Filler 2	have mammary glands	ELEPHANTS → DEERS	+COWS	+ANTEATERS
Target 1	have a bite force greater than 500 BFU	TIGERS → FERRETS	+LIONS	+LIONS
Filler 3	give birth to underdeveloped young	KANGAROOS → WOMBATS	+KOALAS	−FLAMINGOS
Target 2	produce the hormone TH-L2	GRIZZLY BEARS → LIONS	+BLACK BEARS	+BLACK BEARS
Control	require cystocholamine for brain function	ORANGUTANS → GORILLAS	+CHIMPANZEES	+CHIMPANZEES

Table 5.1: The property to be generalized, the first generalization, and additional example used in the BOTH RELEVANT/RELEVANT FILLERS conditions, and in the BOTH RANDOM/RANDOM FILLERS condition. Trials are shown are shown in the order presented in the experiment. All conditions have the same arguments in the key trials (**T**arget 1, **T**arget 2, and **C**ontrol), differing only in cover story and supporting filler trials. The second generalization that people were required to make is formed by combining the first generalization with the additional example. For example, the second generalization for **T**arget 1 becomes {TIGERS, +LIONS} → FERRETS. The "−" symbol is used to indicate that the statement should be negated: e.g., "TORTOISES *don't have* more than one fovea per eye."

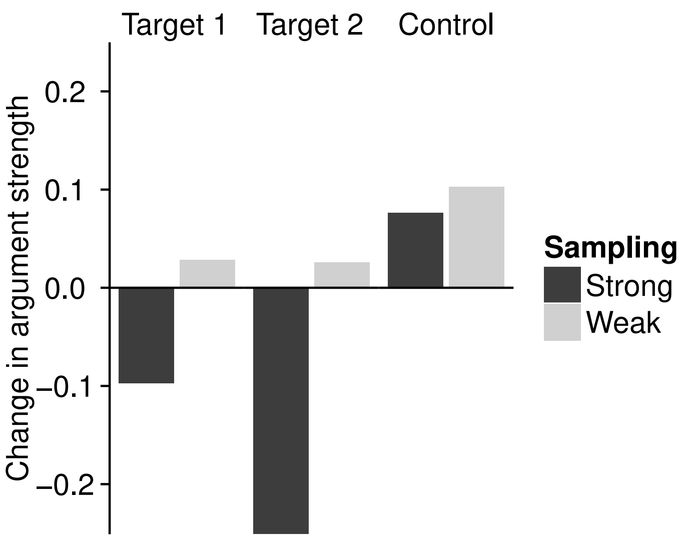


Figure 5.2: Model predictions for the change in argument strength when an additional premise is introduced (i.e., $P(y|x_1,x_2) - P(y|x_1)$). A positive change indicates premise monotonicity, a negative change, non-monotonicity. In the **C**ontrol argument, monotonicity is predicted regardless of sampling assumption. For both **T**arget arguments, a reversal is predicted: premise non-monotonicity is expected only under an assumption of strong sampling. (The difference in the magnitude of the predictions between the two **T**arget conditions emerges due to the structure of people's real-world knowledge about the domain as reflected in the prior, and is incidental to our main point.)

predictions, it was necessary to specify a hypothesis space $\mathcal{H}$ and a prior distribution $P(\mathcal{H})$. The hypothesis space simply consisted of all possible sets of the 14 animals common to the two experimental conditions. In order to estimate the prior, we collected similarity ratings for all pairs of the 14 animals. The estimation procedure was an adaptation of the additive clustering technique (Shepard and Arabie 1979; see also M. D. Lee 2002, Navarro and Griffiths 2008) and is discussed in more detail in Appendix A.

Figure 5.2 shows the resulting model behavior. As predicted previously, when weak sampling is assumed the model indicates premise monotonicity for both **T**arget and **C**ontrol trials. Conversely, under strong sampling it predicts non-monotonicity for **T**arget trials and monotonicity for **C**ontrol trials. Importantly, while the precise numerical prediction shown in Figure 5.2 depends on the way in which the prior was derived, the qualitative effect of sampling assumptions is robust with regard to change in details: as discussed in Appendix A, the Bayesian model predicts a shift towards non-monotonicity under strong sampling provided that the prior distribution reflects the conceptual structure of the animal domain.

		Argument strength					
		Original		Revised		Change	
Condition	N	Mean	SE	Mean	SE	Mean	SE
Target 1							
Both Relevant	135	.283	.021	.210	.021	-.073	.013
Relevant Fillers	134	.313	.023	.259	.022	-.054	.015
Random Fillers	138	.301	.020	.275	.020	-.026	.014
Both Random	131	.277	.022	.307	.026	.031	.021
Target 2							
Both Relevant	135	.523	.015	.444	.023	-.079	.023
Relevant Fillers	134	.538	.017	.484	.022	-.054	.015
Random Fillers	138	.534	.017	.521	.020	-.012	.014
Both Random	131	.578	.018	.616	.021	.038	.013
Control							
Both Relevant	135	.773	.015	.863	.014	.090	.013
Relevant Fillers	134	.765	.015	.860	.013	.096	.009
Random Fillers	138	.759	.013	.853	.013	.093	.011
Both Random	131	.790	.016	.902	.010	.111	.013

Table 5.2: Mean argument strength ratings (linearly scaled to the range 0 to 1) for the original judgment (after seeing the first premise only), the revised judgment (after seeing the second premise), and mean change in argument strength (the revised rating minus the original rating, linearly scaled to the range -1 to 1), summarised by condition and trial type.

Experimental results

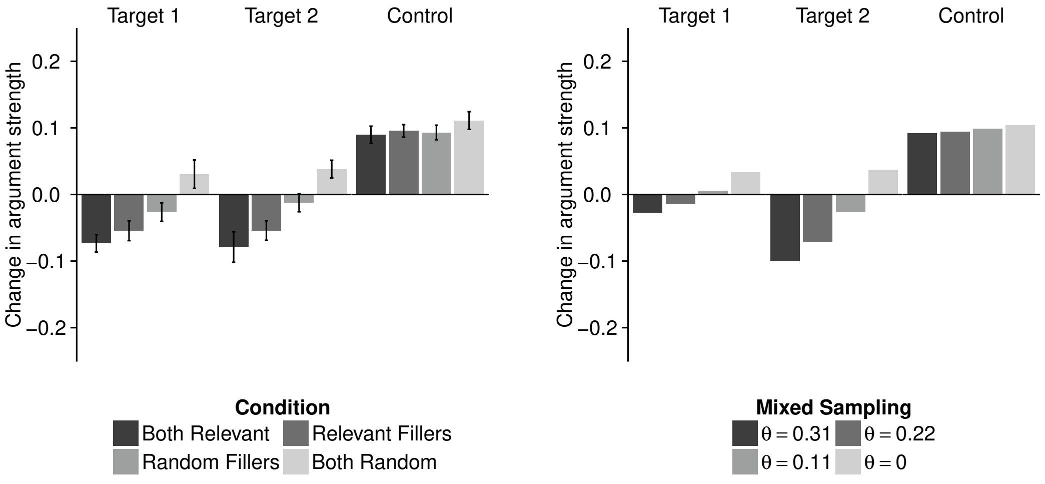
For each trial, participants rated the strength of an argument in a one- and two-premise form. The main question of interest was whether sampling assumptions had an impact upon the way people assessed the evidentiary value of the additional premise. The dependent measure was therefore the response change between the two judgments: a positive response change reflects premise monotonicity, while a negative one reflects non-monotonicity. Table 5.2 presents mean argument strength ratings based on the one- and two-premise forms, as well as the mean change between judgments, by trial type and condition.

	Bayes Factor		
Condition	Target 1	Target 2	Control
Both Relevant	$>$ **1,000** : 1 ($\mu < 0$)	**40** : 1 ($\mu < 0$)	$>$ **1,000** : 1 ($\mu > 0$)
Relevant Fillers	**98** : 1 ($\mu < 0$)	**91** : 1 ($\mu < 0$)	$>$ **1,000** : 1 ($\mu > 0$)
Random Fillers	1 : 1 ($\mu < 0$)	1 : **5.7** ($\mu < 0$)	$>$ **1,000** : 1 ($\mu > 0$)
Both Random	1 : **1.8** ($\mu > 0$)	**13** : 1 ($\mu > 0$)	$>$ **1,000** : 1 ($\mu > 0$)

Table 5.3: Bayes factors indicating the relative likelihood of a one-sided model of mean change in argument strength against the null model ($\mu = 0$), by condition and trial type. The one-sided test performed in each case (given in parentheses) was chosen on the basis of the mean change in argument strength observed. $\mu < 0$, and $\mu > 0$ correspond to the hypotheses that the true mean change in argument strength represents non-monotonic and monotonic responding, respectively. Bold type indicates the preferred model in each case. As predicted, a cover story consistent with a strong sampling assumption lead to non-monotonic responding in the **T**arget trials, but not the **C**ontrol trial, while a cover story consistent with a weak sampling assumption induced monotonic responding across all conditions and trials. Bayes factors are shown to two significant figures.

Figure 5.3(a) shows, as predicted, that people exhibited different response patterns depending on their sampling assumptions. For both **T**arget trials, participants in the Both Random condition exhibited premise monotonicity, while those in the Both Relevant condition showed non-monotonicity. To quantify the amount of evidence for these assertions, for every condition we ran Bayesian analysis comparing three hypotheses: that responding was monotonic (positive change: $\mu > 0$), non-monotonic (negative change: $\mu < 0$) or that the additional premise had no influence (null effect: $\mu = 0$). Analyses were conducted using the BayesFactor package in R (Morey & Rouder, 2014), applying the method outlined by Morey and Wagenmakers (2014) to test one-sided hypotheses. The results of these analyses are summarized in Table 5.3, which reports the Bayes factor between the two best hypotheses in each case. As the table makes clear, there is strong evidence for monotonic reasoning on the control trials regardless of condition, but there is evidence for a shift from monotonic to non-monotonic reasoning in the target conditions.

For the two conditions employing a neutral cover story, our intuition was that a mixed sampling assumption should be induced. Consequently, we expected mean response change in the Relevant Fillers and Random Fillers conditions to be within the bounds of that for the Both Relevant and Both Random conditions. To investigate this intuition, we determined the mix of strong and weak sampling assumptions (captured by θ, as per Equation 5.5) that best fit the mean response change observed for each condition. The fitting process involved finding a value for θ (in the range 0 to 1) that



(a) Human performance

(b) Model fits

Figure 5.3: (a) Average change in people's argument strength ratings for all four conditions, calculated by subtracting their original judgment (after seeing the first premise only) from their revised judgment (after seeing the second premise), then linearly scaled to the range -1 to 1. In keeping with the predictions, people exhibit premise non-monotonicity in the Both Relevant and Relevant Fillers conditions and only for the **T**arget arguments. The results demonstrate that when a relationship amongst premise categories not shared by the conclusion is highlighted, a strong reason is needed in order for such relevance to be ignored and for non-monotonic reasoning to be inhibited. Bars show one standard error. (b) Best fitting value of θ under a mixed sampling assumption. $\theta = 0$ corresponds to a weak sampling assumption, whereas $\theta = 1$ would correspond to an assumption of pure strong sampling. Intermediate values reflect more graded assumptions. The fitted values confirm that when a cover story establishes a high or low expectation of premise relevance consistent with the premises observed, people exhibit an increased bias towards strong or weak sampling, respectively.

		Bayes Factor (: NO EFFECT)		
Model	Order restrictions	Target 1	Target 2	Control
NO EFFECT	$\mu_1 = \mu_2 = \mu_3 = \mu_4$	-	-	-
FILLERS ONLY	$\mu_1 = \mu_2 < \mu_3 = \mu_4$	740 : 1	12,000 : 1	< 1 : 1
STORY ONLY	$\mu_1 < \mu_2 = \mu_3 < \mu_4$	4,100 : 1	17,000 : 1	< 1 : 1
BOTH	$\mu_1 < \mu_2 < \mu_3 < \mu_4$	2,900 : 1	30,000 : 1	< 1 : 1
RANDOM EFFECT	$\mu_1 \neq \mu_2 \neq \mu_3 \neq \mu_4$	520 : 1	4,600 : 1	< 1 : 1

Table 5.4: Bayes factors representing the relative likelihood of the observed changes in argument strength under each model compared with the NO EFFECT model. A higher Bayes factor indicates greater evidence in favour of a particular model. Each model is described in terms of the order restrictions amongst the values μ_1, μ_2, μ_3 and μ_4, which represent the true means of the BOTH RELEVANT, RELEVANT FILLERS, RANDOM FILLERS, and BOTH RANDOM conditions, respectively. Bayes factors are shown to two significant figures.

minimised the squared difference between predicted response change and mean observed response change summed across **C**ontrol and **T**arget trials.

As Figure 5.3(b) shows, the change in relative mixture across conditions follows the expected pattern. The correlation between fitted model and data is 0.94, indicating a good fit overall. Further analysis showed that order restricted models suggesting either an effect of cover story only or both cover story and filler items were both well supported by the data, with the latter having strongest support overall (Bayes factors are shown in Table 5.4). Bayes factors were calculated using a custom JAGS model, employing the product space method of model comparison (Lodewyckx et al. 2011; see Appendix B for details).

Overall, the effect of sampling assumption on premise monotonicity in our experiment was strong enough to cause a genuine reversal in whether people were prepared to endorse the conclusion in one case. For the second target trial, 78% of participants in the BOTH RANDOM condition endorsed the conclusion that lions produce the hormone TH-L2 (by using the on-screen slider to indicate “True”), compared to 37% in the BOTH RELEVANT condition. With respect to the first target trial the effect was less pronounced due to low overall endorsement of the conclusion; 24% endorsement in the BOTH RANDOM condition compared to 11% in the BOTH RELEVANT condition.

5.5 Discussion

Arguments, when presented in everyday life, are intended to bring about a change in the audience. Whether to engage, to teach or persuade, premises are typically selected with a relevant goal in mind. This paper investigates why premise relevance should matter when people evaluate arguments. We demonstrate that people's reasoning in a category-based induction task is dependent on their assumptions about how the premises were sampled. If they think the premises were provided by a helpful confederate choosing positive examples from the categories in question, they show the premise non-monotonicity effect found previously (Medin et al., 2003). However, if they believe that the premises were generated randomly, this effect reverses. These results can be explained by a Bayesian theory of category-based induction that naturally incorporates different assumptions about premise sampling.

Our results support two qualitatively different conclusions. First, our work shows that the perceived strength of an inductive argument is influenced not just by the direct generalizability of premises to conclusion, but also by expectations of premise relevance. By inducing a weak sampling assumption we showed that sensitivity to premise relationships can be violated. Second, this influence is pronounced enough to lead to a reversal of an effect (premise non-monotonicity) that normally obtains for certain kinds of argument structures. Reasoners who hold different sampling assumptions may endorse opposite conclusions as a result.

A previous attempt by Fernbach (2006) to demonstrate premise non-monotonicity by inducing a weak sampling assumption was not entirely successful. Although Fernbach (2006) found a difference in argument strength depending on sampling assumptions, participants in that study did not show a qualitative shift from monotonic to non-monotonic reasoning. Instead, the additional premises raised argument strength in all cases. It is possible that the relevance of the additional premises was not clear enough in that manipulation, which did not vary filler items. In our experiment we used filler items to substantiate the cover story in the Both Relevant and Both Random conditions. For example, our Both Random condition contained negative examples as filler items, without which a weak sampling assumption is difficult to sustain. Previous work involving category learning has also found that people rely on data, not just cover stories, to determine which sampling assumptions are appropriate. For instance, Navarro et al. (2012) found that the data people were shown affected their generalizations, but that sampling assumptions implicit in the cover story did not. A replication of that study which made the sampling assumptions in the cover story more explicit did find a reliable effect of cover story (Vong et al., 2013). Our results showed a reliable effect of both cover story and filler items, with participants in the Relevant Fillers and Random Fillers conditions exhibiting a similar, albeit attenuated, pattern of responding to those in the Both Relevant and Both Random conditions, respectively. This lends further support to the intuitive notion that in many cases people's sampling assumptions reflect some weighted mixture of strong and weak sampling. And while the questions remains open as to whether and how sampling assumptions are updated as new data arrives, it is clear that people do pay attention to the nature of the data when determining how that data was generated.

Prominent models of inductive argument strength, such as the similarity coverage model of Osherson et al. (1990), and the featural similarity model of Sloman (1993) suggest that argument strength is based on the similarity between premise and conclusion, as first observed by Rips (1975). However, these models offer no explicit mechanism to capture sampling assumptions. Each model "hard-wires" a particular assumption instead. In contrast, as we have shown, a Bayesian model along the lines we have illustrated can accommodate the roles of both premise-conclusion similarity and sampling assumptions.

How might the relevance framework for inductive reasoning (Wilson & Sperber, 2004) accommodate our finding that premise sampling assumptions affect argument strength? Relevance theory claims that an input is worth picking out from the mass of competing stimuli when it is *more* relevant, and that an input is more relevant if it produces a larger cognitive effect or requires less effort to process. The addition of a premise that highlights a shared property should raise the relevance of that property when determining the appropriate basis for induction, by decreasing the effort required to call the property to mind. But that should be so in each of our experimental conditions, because identical premises were used in the trials of interest. So that leaves us to posit a difference in cognitive effect to explain a difference in relevance between conditions.

This is where the Bayesian theory of category-based induction comes in. The theory describes how beliefs are revised in response to evidence in terms of the redistribution of probability mass. Such redistribution, we argue, is an excellent candidate measure for cognitive effect. Under this view, the mathematics of Bayes' rule predicts that a strong sampling assumption will always lead to a greater cognitive effect than would a weak sampling assumption because it leads to belief revision due to differences in the likelihood of observing certain data, and not simply due to falsification alone.[5] Relevance theory holds that comparing stimuli on the basis of relevance is a crucial part of human reasoning. The Bayesian theory of category-based induction provides a computational basis for making such comparisons in a way that takes two critical factors – premise sampling assumptions and category similarity – into account. As such, the theory represents an important component that can be integrated into the relevance framework. Likewise, relevance theory complements Bayesian theory insofar as it can make qualitative predictions regarding processing effort. Any algorithmic account of category-based induction should take these predictions into account, as well as relevant empirical findings (e.g. Coley & Vasilyeva, 2010; Feeney, Coley, & Crisp, 2010; Feeney & Heit, 2011).

[5] An important implication of this assumption is that equating cognitive effect directly with change in argument strength is potentially flawed, since the two forms of belief revision can have opposing effects.

In general, we found that people in our experiment quite naturally assumed that premises were selected sensibly or drawn from the category – the difficulty came in trying to persuade them that they were truly random, as in the BOTH RANDOM condition. This observation, in combination with the fact that the premise non-monotonicity found in the BOTH RELEVANT condition corresponds to the standard effect (Medin et al., 2003), suggests that people have an automatic bias to believe that premises are selected sensibly: if not by a helpful teacher, at least in a way consistent with strong sampling (i.e., selected from the category). A biased presumption of relevance is an outcome in keeping with a central claim of relevance theory that people act to maximise relevance when selecting inputs to process (Wilson & Sperber, 2004). This is sensible in the context of category-based induction given that this is how arguments are constructed and used in the real world, but it does mean that we cannot, as researchers, assume that people reason as if we are generating examples randomly (even when we are).

It should be noted that our model incorporates strong sampling, which in the context of category-based induction implies that a category exhibiting the property in question is as likely as any other to be chosen. Seeking to persuade or dissuade another is typically a matter of picking a relevant example of a concept, not a random one. Yet, when a property defines a small or coherent category such as "species of bear" or "black and white striped animals" then there is likely to be little variation in relevance across the category members, and a strong sampling assumption may be appropriate. A pedagogical assumption, in contrast, which gives greater weight to examples that better characterise a property, may be more appropriate for larger, less coherent categories, where there is greater variation in relevance across category members.[6] Shafto et al. (2014) found evidence to suggest that pedagogical sampling compared to strong sampling lead to tighter generalizations on the part of the learner, albeit with simple perceptual stimuli. It is plausible that our BOTH RELEVANT cover story acted to tighten generalizations over and above the predictions of strong sampling. Such a tightening may have acted to increase levels of premise non-monotonicity in the BOTH RELEVANT condition. Further work is needed to determine whether premise non-monotonicity can be observed with a cover story suggestive of a strong sampling assumption alone, in line with our model simulations. Regardless, the likelihood function in the Bayesian model may be adapted to capture either strong or pedagogical sampling (Shafto et al., 2014).

There is substantial evidence to suggest that when attempting to learn, generalize and draw conclusions from data, people are sensitive to the process by which data is generated. This sensitivity to sampling has been previously shown in simple generalization problems (Navarro et al., 2012; Tenenbaum & Griffiths, 2001a), in early word learning (Xu & Tenenbaum, 2007a), and even in infants (Gweon et al., 2010). Other work has demonstrated that people are sensitive to more complicated sampling schemes (Shafto et

[6] Pedagogical sampling (Shafto et al., 2014) may be viewed as a partial instantiation of the *communicative principle of relevance* (Wilson & Sperber, 2004), insofar as it can make predictions about belief revision in an explicitly communicative context.

al., 2014). Our work extends this sensitivity to category-based induction tasks, adding an important clarification to relevance theoretic accounts of a phenomena attributed to relationships amongst premise items. In a world of exclusively weak sampling assumptions, where evidence supports falsification only, the inferential leap receives no boost from premise relevance: the relevant becomes irrelevant.

5.6 Appendix A

In order to generate model predictions (using Equations 5.1–6.4 described in the main paper) it is necessary to specify an hypothesis space $\mathcal{H}$ and a prior distribution, $P(\mathcal{H})$. To do so, we restrict the category labels under consideration to the fourteen experimental stimuli used in both experimental conditions. This is not to say that the experimental participants were aware in advance of the nature and extent of the stimuli used, nor restricted their considerations in this manner. We made this restriction to render analysis tractable, with the view that the predictions remain valid in a qualitative sense, despite this truncation. The fact that our experimental results match our predictions in qualitative terms lends support to this view. Given the fourteen category labels, our hypothesis space $\mathcal{H}$ consists of 2^{14} hypotheses, each corresponding to the proposition that a unique cluster of categories share a given property.

Having established our hypothesis space $\mathcal{H}$, we need to separately derive a plausible prior distribution, $P(h)$, defined over all $h \in \mathcal{H}$. We seek a prior that is independent of any particular property or this specific task, to avoid fitting our predictions too tightly to the properties used in our experimental trials. That is, $P(h)$ represents the probability that a blank (unseen) property is shared by those items that belong to a particular category h. In keeping with prominent models of category-based induction (Osherson et al., 1990; Sloman, 1993), we assume that generalizing a property from one item to another involves an assessment of their similarity. Intuitively, since hypotheses in our model correspond to clusters of items, we seek to establish a weighting for each cluster that reflects its coherence. Prior probabilities will be derived from these clusters, with higher prior probabilities assigned to more coherent clusters.

To establish clusters and associated weights we apply the *additive clustering* (ADCLUS) model (M. D. Lee, 2002; Navarro & Griffiths, 2008; Shepard & Arabie, 1979) to similarity data gathered from a separate experiment, described in more detail below. On the basis of observed similarity data, ADCLUS identifies structure in the domain free from the undesirable restriction that such structure should take a strictly hierarchical form. The model defines the similarity of any two objects as the sum of the weights across all clusters containing both objects. It attempts to find a set of clusters and weights maximising the fit between empirical similarity data and the theoretically reconstructed measures. Finding an optimal fit is an under-constrained and computationally expensive exercise, hence the model implementation seeks to find a good and parsimonious fit.

Starting with an initial configuration of clusters and weights, a gradient descent algorithm is employed to find a suitable local optimum. On each iteration of the gradient descent process, clusters with non-appreciable weights may be discarded.

In order to provide empirical interstimulus similarities as input to the ADCLUS model, a separate experiment was conducted to gather similarity ratings via a triad task for the fourteen animal stimuli common to all conditions of our experiment. 63 adults were recruited via Amazon Mechanical Turk, and were each paid \$0.60 (USD) for the 5–10 minutes participation. 5 were excluded due to browser incompatibility, and the remaining 58 were aged 19 to 75 years (median age 31, 41% female). 50 participants were in the United States, with 8 located elsewhere. For each triad presented, people were asked to pick which animal was *least* similar to the others. Each person rated 60 randomly selected triads. Since there were a total of 364 possible triads, this meant that each triad was rated by 9–10 participants on average. The pairwise interstimulus similarity for two stimuli a and b was calculated as the proportion of all triad ratings for a, b, and some other stimulus c, where c was rated as being the least similar.

The final stage in our model implementation involves the assignment of prior probabilities based on the clusters and weights identified by ADCLUS. Let $\mathcal{H}_C$ denote those hypotheses (clusters) identified by the ADCLUS process, and w_h denote the weight associated with hypothesis $h \in \mathcal{H}_C$. We form an initial estimate of the prior distribution directly from these outputs:

$$P_w(h) \propto \begin{cases} w_h & \text{if } h \in \mathcal{H}_C \\ 0 & \text{otherwise} \end{cases} \tag{5.6}$$

This initial estimate is not quite right, however. The ADCLUS model does not deal meaningfully with clusters corresponding to a single category. Yet intuitively, in the context of our experiment, properties that pertain to a single category (TIGER, for example) are quite plausible. Therefore we need to combine the prior derived from the cluster weights with one that assigns non-zero probability to the singleton hypotheses (the set of which we denote $\mathcal{H}_S$). For the latter, we use a size-based prior:

$$P_s(h) \propto \begin{cases} \frac{1}{|h|} & \text{if } h \in \mathcal{H}_C \cup \mathcal{H}_S \\ 0 & \text{otherwise} \end{cases} \tag{5.7}$$

Lastly, we combine these two prior distributions to form the prior used to generate our model predictions in such a way that the probabilities for singleton hypotheses calculated in Equation 5.7 are preserved:

$$P(h) = \begin{cases} P_s(h) & \text{if } h \in \mathcal{H}_S \\ P_w(h) \sum_{h' \in \mathcal{H}_C} P_s(h') & \text{if } h \in \mathcal{H}_C \\ 0 & \text{otherwise} \end{cases} \tag{5.8}$$

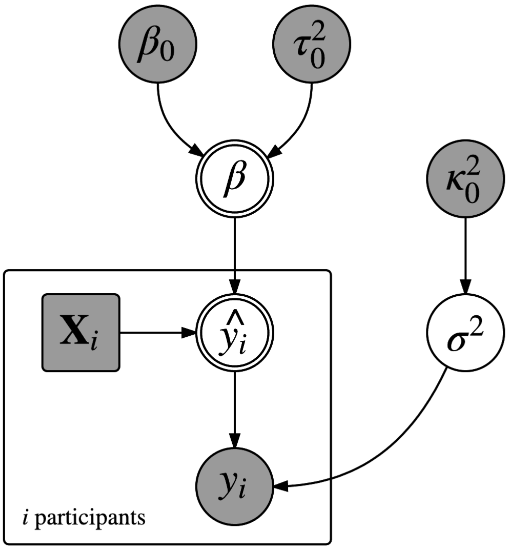
As the reader will note, our method for defining the hypothesis space and for deriving prior probabilities affords a certain latitude. Using the ADCLUS model, the precise clusters and associated weights identified depend on the values chosen to seed the optimization process. Whilst we retain the seeding heuristic of Shepard and Arabie (1979), we also experimented with other heuristics. We found that although such alternatives lead to different numerical predictions, the important qualitative effect was robust: greater levels of premise non-monotonicty were predicted under a strong sampling assumption than under a weak sampling assumption for the **T**arget (but not the **C**ontrol) arguments. Similarly, alternative methods may be employed for assigning probabilities to singleton hypothesis, but once again, the qualitative predictions appear robust in the face of such changes.

5.7 APPENDIX B

As discussed in the main text, differences in mean change in argument strength across conditions indicated that our experimental manipulation had some effect. To investigate the factors driving the effect we compared a number of plausible models to determine which might best account for our experimental results. The models considered were based on the change in argument strength predicted by our Bayesian model of category-based induction, derived from empirical similarity ratings. Under a strong sampling assumption, our model predicts non-monotonic responding for both **T**arget trials; under a weak sampling assumption, monotonic responding is predicted.

Furthermore, the fitted values of θ derived from the mixed sampling model suggest an ordering in terms of mean response change across conditions. Thus, consistent with the suggested orderings, three plausible models concerning the nature of the effect were compared, namely: that the effect was driven by the filler items only (FILLERS ONLY), that it was driven by the cover story only (STORY ONLY), or that it was driven by both of these factors (BOTH). The order restrictions for each model are shown in Table 5.5. A fourth unrestricted model was also considered, namely that results were driven by a random effect (RANDOM EFFECT).

For each of the four models, we calculated the Bayes factor representing the relative likelihood of the observed changes in argument strength under the model against the "no effect" model (NO EFFECT). To do so, we employed a Markov chain Monte Carlo (MCMC) procedure known as the *product space method* (Lodewyckx et al., 2011). The technique supports the comparison of two models (M_o and M_1, for example) by building a hierarchical "supermodel" combining the models via a random variable (M, say) that acts as a model index. The Bayes factor for the relative likelihood of M_1 against M_0 becomes the posterior odds ratio ($M_1 : M_0$) for the two models, divided by the prior odds ratio. Theoretically, the prior model probabilities may be chosen with freedom, although



$$\begin{aligned}
y_i &\sim \text{Normal}\left(\hat{y}_i, \sigma^2\right)\\
\hat{y}_i &= \beta_1 + \beta_2 X_{1i} + \beta_3 X_{2i} + \beta_4 X_{3i}\\
\sigma^2 &\sim \text{Scaled Inverse } \chi^2\left(1, \kappa_0^2\right)\\
\beta_1 &\sim \text{Normal}\left(\beta_0, \tau_0^2\right)\\
\beta_2 &= \delta_1\\
\beta_3 &= \delta_1 + \delta_2\\
\beta_4 &= \delta_1 + \delta_2 + \delta_3
\end{aligned}$$

Figure 5.4: A graphical model supporting comparison of condition means. For each of the five models considered, β_1 represents the condition mean of the reference condition BOTH RELEVANT. β_2, β_3, and β_4, represent the difference between the mean of the reference condition and the mean of the RELEVANT FILLERS, RANDOM FILLERS, and BOTH RANDOM conditions, respectively. The models differ only in the definition of δ_1, δ_2, and δ_3.

technical considerations require careful selection if reliable MCMC estimates are to be obtained. Finally, the prior probabilities for each model may be estimated as follows:

$$\hat{P}(M_k \mid \text{Data}) = \frac{\text{Number of posterior samples where } M = k}{\text{Total number of posterior samples}} \tag{5.9}$$

from which the Bayes factor easily follows.

Figure 5.4 shows the graphical model capturing the common elements for each of the models tested. The vector quantity

$$\mathbf{X}_i = (X_{1i}, X_{2i}, X_{3i})$$

represents a dummy coding of condition for each participant. The vector quantity

$$\boldsymbol{\beta} = (\beta_1, \beta_2, \beta_3, \beta_4)$$

captures the relationship between the means μ_1, μ_2, μ_3, and μ_4 of the BOTH RELEVANT, RELEVANT FILLERS, RANDOM FILLERS, and BOTH RANDOM conditions, respectively; that is,

$$\beta_1 = \mu_1,\ \beta_2 = \mu_2 - \mu_1,\ \beta_3 = \mu_3 - \mu_1,\ \text{and } \beta_4 = \mu_4 - \mu_1.$$

The δ_i parameters represent the difference between adjacent condition means, and are each sampled from a normal distribution with mean 0 and variance τ_0^2. The range restrictions

		Parameter range		
Model	Order restrictions	δ_1	δ_2	δ_3
NO EFFECT	$\mu_1 = \mu_2 = \mu_3 = \mu_4$	0	0	0
FILLERS ONLY	$\mu_1 = \mu_2 < \mu_3 = \mu_4$	0	$(0, \infty)$	0
STORY ONLY	$\mu_1 < \mu_2 = \mu_3 < \mu_4$	$(0, \infty)$	0	$(0, \infty)$
BOTH	$\mu_1 < \mu_2 < \mu_3 < \mu_4$	$(0, \infty)$	$(0, \infty)$	$(0, \infty)$
RANDOM EFFECT	$\mu_1 \neq \mu_2 \neq \mu_3 \neq \mu_4$	$(-\infty, \infty)$	$(-\infty, \infty)$	$(-\infty, \infty)$

Table 5.5: The range restriction imposed on the Normal$(0, \tau_0^2)$ distribution from which the δ_i parameters are sampled for each model. A value of 0 indicates that the respective parameter is always 0.

on the values sampled differ across the five models, as shown in Table 5.5. The mean of the reference condition has a normal prior distribution with mean β_0, and variance τ_0^2. The prior for the error variance (σ^2) is a scaled inverse χ^2 distribution, with 1 degree of freedom and scaling parameter κ_0^2. To ensure that these prior distributions do not favour any one particular model, and that the posterior is effectively independent of the prior, the values for β_0, τ_0^2, and κ_0^2 were derived from the data using the procedure outlined in Klugkist, Laudy, and Hoijtink (2005, p. 482).

Study III

WHEN THE BASIS FOR SAMPLING IS UNCLEAR

STATEMENT OF AUTHORSHIP

TITLE OF PAPER	Where the truth lies: how sampling implications drive deception without lying
PUBLICATION STATUS	Unpublished and unsubmitted work written in manuscript style
PUBLICATION DETAILS	Not yet submitted for publication.

Principal author

NAME OF PRINCIPLE AUTHOR (CANDIDATE)	Keith Ransom
CONTRIBUTION TO THE PAPER	Designed and ran experiments, performed data analysis, implemented computational models, and wrote manuscript.
OVERALL PERCENTAGE (%)	80%
CERTIFICATION	This paper reports on original research I conducted during the period of my Higher Degree by Research candidature and is not subject to any obligations or contractual agreements with a third party that would constrain its inclusion in this thesis. I am the primary author of this p
SIGNATURE	
DATE	18/08/19

Co-author contributions

By signing the Statement of Authorship, each author certifies that:

1. the candidate's stated contribution to the publication is accurate (as detailed above);
2. permission is granted for the candidate in include the publication in the thesis; and
3. the sum of all co-author contributions is equal to 100% less the candidate's stated contribution.

NAME OF CO-AUTHOR	Wouter Voorspoels
CONTRIBUTION TO THE PAPER	Supervised development of work, helped with experimental design, computational modelling and editing of the manuscript.
SIGNATURE	
DATE	18/08/19

NAME OF CO-AUTHOR	Danielle Navarro
CONTRIBUTION TO THE PAPER	Supervised development of work, helped with experimental design, computational modelling and editing of the manuscript.
SIGNATURE	
DATE	18/08/19

NAME OF CO-AUTHOR	Amy Perfors
CONTRIBUTION TO THE PAPER	Supervised development of work, helped with experimental design, computational modelling and editing of the manuscript.
SIGNATURE	
DATE	18/08/19

6 WHERE THE TRUTH LIES

Efficient communication leaves gaps between message and meaning. Interlocutors, by reasoning about how each other reasons, can help to fill these gaps. To the extent that such *meta-inference* is not calibrated, communication is impaired, raising the possibility of manipulation for deceptive ends. We examined how people reason in such a situation by having people act as the perpetrator or target of deception across two related experiments. Importantly, the nature of the task meant that outright lying was impossible. As a result, deception required either concealing information or supplying technically correct but misleading information. We find evidence for two distinct patterns of behaviour. One group of people appear to make assumptions about communicative intent based on context and message content. Senders in this group were more likely to mislead, and receivers were more effectively misled. A second group of people appeared to adopt a more defensive stance displaying the same cautious approach in all situations. We explain this behaviour using a computational level account of the kinds of inferences required by both receiver and sender. These distinct patterns arise from different assumptions about the generative process behind communication.

6.1 INTRODUCTION

Inference on the basis of real-world communication is a complex and under-constrained problem. Messages (not unlike flat-packed furniture) rarely come complete with everything necessary to assemble what was intended. Over and above decoding the message on syntactic and semantic grounds, the receiver must also fill in gaps based on her existing knowledge and her inductive biases. In so doing, she may make assumptions about the way that the sender chooses what to say on the basis of what he means to convey (Grice, 1989). Likewise, a sender who seeks to convey a given meaning may make his own assumptions about the receiver and how she will decode his meaning from the contents of his message. Critically, both sender and receiver may recognise that for each assumption they themselves make, their interlocutor may assume that they make it. This pattern of reciprocal and potentially recursive "meta-inference" may be leveraged by both parties. This can enhance their ability to communicate accurately yet efficiently and result in stronger conclusions and more decisive action.

However, meta-inferential reasoning presents challenges of its own. Meta-inference, like inference in general, is under-determined: as a result, successful inference relies on making assumptions that are appropriate to the problem space. Without adequate mechanisms to ensure that the assumptions of senders and receivers are reciprocally calibrated, communication can be significantly impaired. Communicative conventions which rest on the presumption that communication is generally truthful, cooperative and goal-directed offer a basis for calibrated meta-inference.

At the same time, such conventions raise the possibility of manipulation for deceptive ends. Once we accept that cooperative principles may not always hold, communicative meta-inference becomes a vital prerequisite. Without it, people who try to lie will fail to do so successfully and people who are lied to will be consistently taken in. The spectre of deception thus both necessitates and complicates meta-inference. The inferential problem is especially difficult because skilful deceivers may strive to avoid detection by never lying outright. This kind of deception, known as *paltering* (Schauer & Zeckhauser, 2009), occurs when the communicator takes advantage of what they know about the listener's mental state to provide information that is not strictly false but will cause the listener to draw the wrong conclusion.

How *do* people reason in contexts where the goals of sender and receiver are not necessarily aligned? In the present study, we investigate this issue in a setting where deception is warranted but outright lying does not occur. Using behavioural data from two related experiments, one sender focused and one receiver focused, we examine how meta-inference (and ultimately inference) is affected when cooperative norms may no longer apply. We then rely on a model-based analysis to address a number of further questions. First, intuitively, we expect receivers to reason from evidence (messages) differently based on the perceived intent of the sender: why should this be the case when the veracity of the evidence is beyond question? Second, do receivers use cues from the message content itself in order to gauge the sender's intent? And finally, how do these factors affect the sender when deciding whether to mislead or conceal information?

Our approach here complements descriptive accounts of pragmatic reasoning (Grice, 1989) and verbal deception (e.g., Dynel, 2011), which yield valuable insight into the measures and counter-measures that senders and receivers employ. Our contribution lies in providing a computational account of how different strategies may be weighed in the balance and how precisely such strategies give rise to different behaviours and inferences. By casting people's beliefs about the conventions that govern communication as sampling assumptions, and formalising message production as the computational inverse of comprehension, we can examine the trade-offs involved with greater precision. We demonstrate that a form of meta-inferential signalling affects the interplay between meta-inference and inference, and our results highlight the added complexity of meta-inference when the possibility of deception arises.

Before presenting our experimental work, we characterise the meta-inferential challenge that deception without lying represents and provide an overview of the theoretical basis for our analysis.

THE LIAR'S TOOLBOX: A META-INFERENTIAL CHALLENGE

Imagine the following scenario:

> You are a graduate student, attending an academic conference for the first time. Nervous about your presentation the next morning, you have some wine at the conference dinner to help you relax. One thing leads to another and after a night of heavy drinking, you oversleep and miss your talk. Travelling home from the conference, you meet a colleague at the airport. She asks you how your talk went. The colleague is a potential future employer, so you are keen not to look foolish.

Assuming that you'd prefer not to reveal what really happened, how can you conceal the truth from her? There are three main strategies you might consider, each corresponding to different violations of the Gricean maxims:

OUTRIGHT LYING: One possibility is to proffer a blatant falsehood: *"My talk went really well! I was touched by the standing ovation."* By communicating facts which the sender knows false, outright lying represents a violation of the fundamental norm of communication. But as long this violation goes undetected, the receiver may leverage the assumption of cooperation implicit in the context and draw the desired incorrect conclusion. That said, lying is often fraught with difficulty. The liar may be uncertain about what the receiver already knows and it may be easier for the receiver to detect if new facts come to light. Outright lying is thus not necessarily the safest option, even for a completely amoral and self-interested communicator.

BEING UNINFORMATIVE: To avoid outright lying, it may be preferable to say something irrelevant or otherwise uninformative: *"The conference dinner was fun."* Where no new information is disclosed the receiver's inference is seemingly restricted to her prior beliefs. But overtly flouting maxims of relevance and quantity in this way is likely to raise suspicion. Indeed, the blatant violation of Gricean norms is often deliberately used as a communicative strategy of its own.

MISLEADING: A third option is paltering: providing truthful but information with a misleading implication in mind: *"I was nervous beforehand, but the session was over before I knew it and there weren't any questions I couldn't handle."* There are considerable advantages to this strategy. Outright lies often bring harsh consequences when detected. Misleading implication which, by defintion, is not

> part of what is explicitly conveyed, offers a sense of plausible deniability (J. J. Lee & Pinker, 2010; Pinker, Nowak, & Lee, 2008) and diminished repercussions (Schauer & Zeckhauser, 2009). Perhaps because of this, people may be less likely to view this form of deception as equivalent to lying (e.g., Coleman & Kay, 1981; Hardin, 2010), although this perception may change when one is on the receiving end (Rogers, Zeckhauser, Gino, Norton, & Schweitzer, 2017). Importantly, because misleading involves being genuinely informative, norms of relevance and quantity are not overtly violated. This acts to limit suspicion and reduce the risk of detection while at the same time leveraging the receiver's presumption of cooperation. Thus, by selectively sampling facts in the right way, the sender may lead the receiver to a false conclusion as a result. Of course, this strategy carries risks of its own. For one thing, it requires the sender to accurately judge the conclusions that the receiver will draw. These inferences are likely to be determined in part by the receiver's assumptions and level of prior suspicion. Allowing for individual variation on the part of the receiver makes matters more complex. Misleading thus becomes a delicate balancing act: enough information must be disclosed to avoid or reduce suspicion, but not enough that the chance of inferring the truth increases.

What kind of option do people tend to choose in this kind of situation? In a preliminary study, we asked 96 first year psychology students (87 women) at the University of Leuven to imagine seven different scenarios like the one above. Participants selected a response from seven options consisting of two lies, two uninformative statements, two misleading statements, and the truth. Figure 6.1 presents their preferences, collapsed across scenarios and equivalent response options.

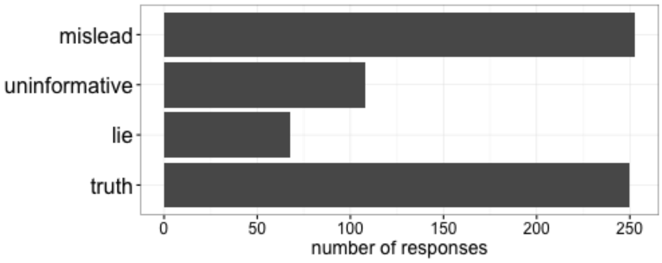


Figure 6.1: When choosing how to communicate in a variety of different scenarios with a clear motivation to deceive, people showed a strong preference to mislead rather than be uninformative. Telling an outright lie was the least preferred option.

Two important conclusions emerge from Figure 6.1. Firstly, people were uncomfortable with deception: 37% of responses involved telling the full truth and only 10% were outright lies: a surprising number perhaps given that each scenario provided a clear motivation to deceive. Secondly, among those who chose not to tell the truth, people showed a clear preference for misleading over lying or being uninformative (37%, 10%

and 15% respectively). This finding is consistent with other work on the topic (Montague, Navarro, Perfors, Warner, & Shafto, 2011; Rogers et al., 2017).

Why do people seem to prefer to actively mislead rather than be entirely uninformative? At first glance, it seems rational to be as uninformative as possible, because you are providing no information that the receiver can use to revise her beliefs at all. Effective misleading, on the other hand, involves salting your statements with a grain of truth. It thus runs a greater risk of the receiver inferring the real truth.

An important motivation for choosing a misleading utterance over a strictly uninformative one is because the latter is suspicious. Consider the likely response of choosing to be uninformative in our earlier scenario:

Colleague: How did your talk go?
You: The conference dinner was fun.
Colleague: Talk didn't go so well?
You: The main conference room comfortably seats 400 people.
Colleague: That bad, huh? What happened?

Sperber et al. (2010) propose that people have a toolbox of cognitive mechanisms for *epistemic vigilance* that reduces the risk of being deceived. The ability to track cooperation in others forms an integral part of such a defence. Whether through dedicated cognitive mechanisms or domain general capacities, obvious departures from communicative norms can be reliably detected by children as young as 3–6 years old (e.g., Eskritt, Whalen, & Lee, 2008; Okanda, Asada, Moriguchi, & Itakura, 2015; Skarakis-Doyle, Izaryk, Campbell, & Terry, 2014). Responding in an uninformative way violates the principle of cooperation so blatantly that the deception is revealed.

A deceiver, sensitive to the epistemic vigilance of his counterpart may prefer instead to provide truthful but misleading utterances, a technique which may reduce or bypass such scrutiny altogether (Reboul, 2017). However, in so doing, he faces a delicate trade-off. Chosen well, such utterances may not only allay the receiver's suspicion, but by virtue of the inferential boost accorded to cooperative speakers, the receiver may be led to a false conclusion, terminating the search for further information (Bonawitz et al., 2011; Montague et al., 2011). Yet if suspicion is already raised, the receiver is unlikely to fall for the false implicature and may use the information to get closer to the truth (Dynel, 2011).

This analysis points to two opposite forces, balanced in the selection of one strategy over another. On one hand, the knowledge that the receiver may engage in inference about the helpfulness of the statement may lead the sender to opt for a misleading yet informative statement. On the other hand, if the sender considers that the receiver will be suspicious a priori, he may resort to being uninformative. In the following section we present a computational account of meta-inference which has the potential to capture this sort of reasoning.

Sampling assumptions as meta-inference

Consider a communication scenario where one person (the *receiver*) seeks to update her beliefs on the basis of information disclosed by another (the *sender*). The sender, for his part, selects information designed (according to his intention) to help or hinder the receiver in her efforts. We may characterise the reasoning of two such communicating parties as a form of Bayesian inference (following, for example, N. D. Goodman & Frank, 2016; Shafto et al., 2014).

Turning first to the problem faced by the receiver: how should she update her beliefs based on the evidence provided by the sender? Let h denote one possible hypothesis that the receiver is currently considering, and $P(h)$ denote her belief in the hypothesis prior to receiving information from the sender. Then, having observed new information x (revealed by the sender) the receiver updates her belief according to:

$$P_{\textsc{Receiver}}(h|x) \propto P_{\textsc{Sender}}(x|h)P(h), \tag{6.1}$$

where $P_{\textsc{Sender}}(x|h)$ represents the assumption the receiver makes about the sender's sampling strategy (the way he chooses information to convey). The sender, in turn, is assumed to select information according to a sampling strategy targeted to the receiver:

$$P_{\textsc{Sender}}(x|h) \propto (P_{\textsc{Receiver}}(h|x))^{\alpha} \tag{6.2}$$

where $P_{\textsc{Receiver}}(h|x)$ represents an assumption the sender makes about the belief update rule adopted by the receiver.

The goals of the sender are captured by the parameter α. A positive value for α corresponds to a sender who wishes to reveal the truth (that is, to increase the receiver's posterior belief in the correct hypothesis h); a negative value for α implies that the sender wishes to conceal the truth (by reducing the receiver's posterior belief). The magnitude of α indicates the degree to which the sender selects optimally: the larger the magnitude, the closer to optimal his selection becomes, the smaller the magnitude the more his choice resembles random selection. There are other ways to capture conflicting goals, like assigning separate utility functions for the sender and receiver with regard to truth-predicated action, but we chose this for its relative simplicity.

Using the model to describe a particular communication scenario requires Equations 6.1 and 6.2 to be considered simultaneously. Describing how the receiver updates her beliefs amounts to specifying the sampling strategy for the hypothetical sender that she thinks she is facing. Likewise, describing the sender's sampling strategy requires stating an update rule for the hypothetical receiver that he considers. This may be a deeply recursive process, depending on the level of *"he thinks that she thinks that he thinks..."* reasoning that occurs. However deep the reasoning, we end up with a series of sender and receiver models nested under one another, starting at the top level with the model of the sender and receiver whose behaviour we wish to capture, progressing to the model the sender has of the receiver, and that the receiver has of the sender, and so on. Past empirical work (e.g., Colman, 2003; Franke & Degen, 2016; Stiller et al., 2015; Vogel et al., 2013) has suggested that recursive reasoning in this fashion may be limited in depth. We can avoid infinite nesting by specifying a sender with $\alpha = 0$. In this case there is no need for further nesting, since such a sender selects information at random without regard for the receiver. Likewise, any alternative update rule for the receiver that is effectively independent of the sender will also suffice as a ground term.

In any communication scenario there is a potential mismatch in the meta-inferential assumptions of the two parties. In pedagogical situations, where both parties have incentive to improve the effectiveness and efficiency of communication, such asymmetries may be of little consequence; similar qualitative patterns of inference emerge whether all assumptions are reciprocated or not. But when the goals of sender and receiver are at odds, qualitatively different patterns of reasoning may emerge depending on who is aware of the mismatch, who is aware of who is aware, and so on. By structuring a model as a series of nested sub-models, we can capture differing degrees of reciprocal awareness between sender and receiver regarding the sender's intent.

In this paper we use this computational framework to explain people's behaviour a pair of related experiments involving a simple "deception game" (see Figure 6.2). In the first (Experiment 1), participants took the role of the receiver and were asked to infer the truth on the basis of potentially deceptive evidence. In the second (Experiment 2), people acted as the sender and were asked to provide evidence relevant to the hypotheses in question while at the same time preventing the truth from being discovered. We find that people's inferences and choices in this task are sensitive to the level of suspicion of the receiver. Moreover, qualitative individual differences in how people reason in the deception game correspond to different assumptions about intent and the data sampling process, as described by our computational framework.

6.2 Experiment 1: Reasoning from deceptive communications (receiver)

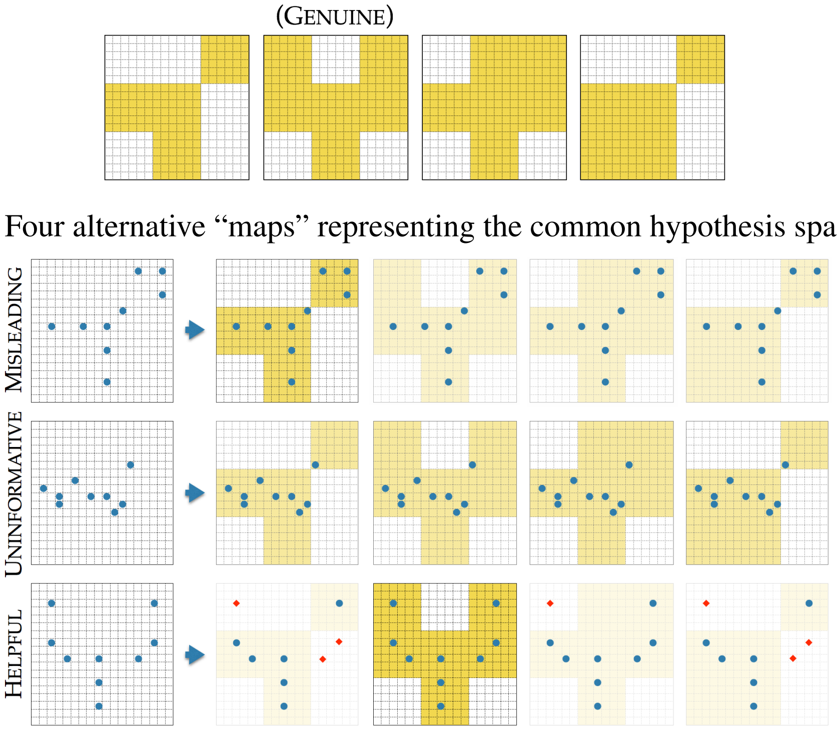
Method

Participants

We recruited 99 adults via Amazon Mechanical Turk, who were each paid $2.00USD for 5-10 minutes participation. One participant was excluded for browser incompatibility. The remaining 98 participants were 59% male and aged 19-64 (median age 29).

Procedure

A cover story informed people that they were taking part in an experiment simulating an online game based on data provided by past players of varying skill levels. People



(a) Four alternative “maps” representing the common hypothesis space.

(b) Different patterns of evidence and the inferences they may licence.

Figure 6.2: The deception game. (a) People taking the role of the receiver (Experiment 1) and the sender (Experiment 2) see the same four “maps” in corresponding trials; the shaded area marks the region where treasure is buried. Only one of the four maps is genuine. (b) As sender, people seek to conceal the identity of the genuine map, but are nonetheless required to reveal some locations where treasure is actually buried (blue dots). They are given three options to choose from: representing *Misleading*, *Uninformative*, or *Helpful* evidence. As receiver, people attempt to infer the identity of the genuine map on the basis of the evidence provided, which varies in its potential to drive inference. The brightness of the shaded areas has been varied here to illustrate how plausible a trusting receiver might consider each map after viewing the evidence (brighter maps represent more plausible hypotheses and red dots indicate disconfirmatory evidence).

were told that they would take the role of an “explorer” (the receiver in our terminology) who must decide on a turn by turn basis which of four treasure maps is genuine based on evidence provided by a past player taking the turn of a “pirate” (the sender). The evidence consisted of points corresponding to a subset of locations drawn from the genuine map. Each point corresponded to a location where treasure was actually buried, but a sender could provide misleading or uninformative evidence through a strategic selection of points.

People’s beliefs about the **sender’s intent** in supplying the evidence was the basis of a within subjects manipulation. In the Teammate condition, participants (as receivers)

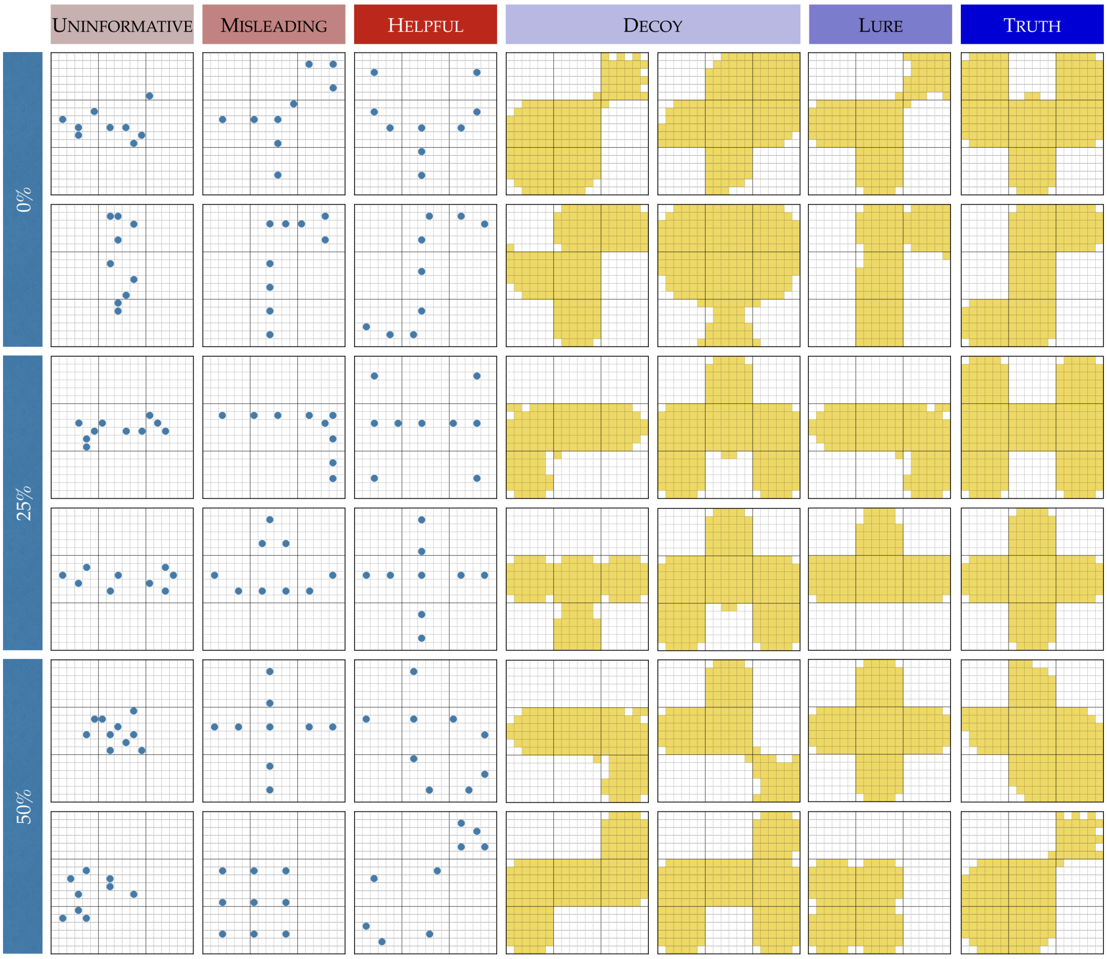


Figure 6.3: The experimental stimuli. Each trial in both experiments involved one of six sets of stimuli (rows), comprised of four maps (yellow regions) and three sets of evidence (blue dots). The task for the sender was to select one of the three sets of evidence to give to the receiver, while the task for the receiver was to select one of the four maps on the basis of the evidence they were given. The *Uninformative* evidence is consistent with all four corresponding maps. The *Helpful* evidence is consistent with only one map (the *Truth*). The *Misleading* evidence is designed to encourage a false conclusion (that the *Lure* map is genuine), but is also consistent with the *Truth*. The informativeness of the *Misleading* evidence was manipulated by controlling the number of *Decoy* maps with which it was consistent (row labels indicate the percentage of hypotheses (maps) ruled out by the *Misleading* evidence).

were told that the sender's goal had been to help a teammate identify the genuine map. In the OPPONENT condition the receivers were told that the goal of the speaker had been to keep its identity concealed.[1] Regardless of condition, participants knew that the sender could not provide false information. Thus, evidence could be relied upon to rule out a given map if any of the locations indicated did not overlap the shaded region shown on the map.

After the training session, people were shown a block of trials for the TEAMMATE condition and a block of trials for the OPPONENT condition. Within each block, participants saw each of the six map sets on three separate occasions: once in conjunction with the *Uninformative* evidence, once with *Misleading*, and once *Helpful* (see Figure 6.3). Thus each block consisted of 18 trials in all. The on-screen order of maps displayed in each trial, the trial order within each block, and the block order itself were all randomised. On each trial, people were required to consider the four maps and the evidence provided, and, taking into account whether the sender was a TEAMMATE or an OPPONENT, indicate which of the four maps they believed was most likely to be genuine.

Materials

The full set of experimental stimuli is shown in Figure 6.3. Each set consisted of four maps and three pieces of associated evidence. The **quality** of the evidence was systematically varied from trial to trial. *Helpful* evidence constituted a pattern of locations that bore a close resemblance to the genuine map and ruled out three of the four alternatives. *Uninformative* evidence, in contrast, bore little similarity to any of the four maps, and could not be used to rule any out. *Misleading* evidence was designed to bear a strong resemblance to one of the three false maps. In addition, the **informativeness** of the *Misleading* evidence varied across the six sets of stimuli, ruling out either none, one or two of the *Decoy* maps, but never the genuine map (the *Truth*) or the map that it was designed to resemble (the *Lure*).

BASIC RESULTS

Our first question of interest is whether people take the intention of the sender into account when interpreting the evidence offered. Figure 6.4, which plots the responses of participants based on what they were told about the sender, suggests that they do indeed. To examine the strength of evidence for this finding we conducted a Bayesian multinomial logistic regression, comparing two mixed-effects models. In the EVIDENCE ONLY model, responses were predicted on the basis of the type of evidence presented in each trial. In the EVIDENCE + SUSPICION model, predictions also included an indicator of suspicion

[1] To rule out consideration for player reputation, and to allow us to present each set of maps more than once, the instructions made it clear that people faced a different player on each trial. To reinforce this, the name and colour of the icon representing the pirate player was different for each trial.

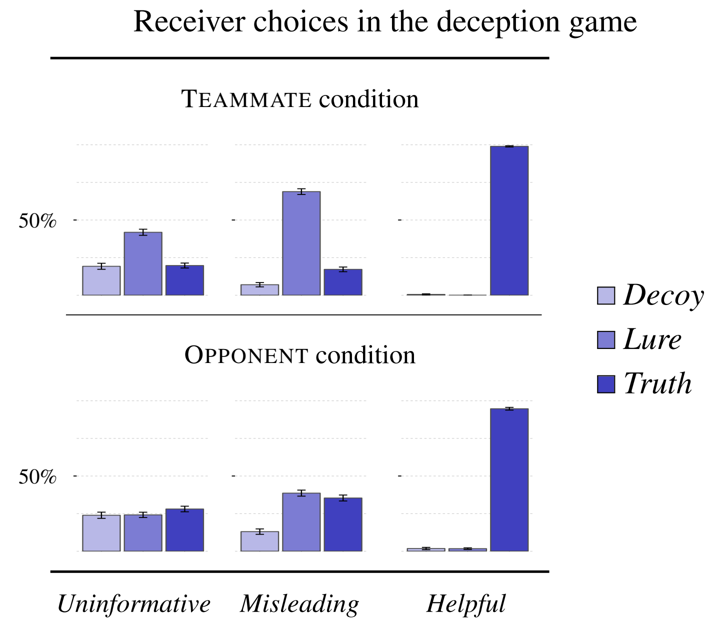


Figure 6.4: Proportion of participants selecting each response in Experiment 1. People playing the role of receiver were asked to identify which of four maps they believed to be genuine on the basis of the evidence provided. On each trial, people chose between two incorrect *Decoy* items, one *Lure* (a subset of the genuine map) and the *Truth*. In the TEAMMATE condition, people were told that the evidence had been provided by a helpful teammate; in the OPPONENT condition they were told that it had come from an opponent trying to conceal the truth. People correctly recognised that the *Helpful* evidence was consistent only with the *Truth*, and responded accordingly. When the evidence was *Misleading* (consistent with both the *Lure* and the *Truth*, but closest in size to the *Lure*) people were far more likely to choose the *Lure* in the TEAMMATE condition where there was reason to trust the sender. Likewise, when faced with *Uninformative* evidence (consistent with all four choices, but closest in size to the *Lure* in three out of six cases) people also displayed a preference for the *Lure* in the TEAMMATE condition. Error bars show standard error, and the proportion of responses favouring the *Decoy* items, is averaged over the two options.

(based on condition). The analyses revealed strong evidence ($BF_{10} > 10^6$) in favour of the EVIDENCE + SUSPICION model over the EVIDENCE ONLY model, consistent with the notion that people reason differently depending on the context of communication.[2] When people thought the sender was trying to help them they reasoned beyond the immediate evidence, drawing strong (but mistaken) conclusions as a consequence. But when they thought the sender was trying to conceal the truth, people adopted a more conservative approach, appearing to select an option at random from amongst those not directly ruled out by the evidence.

As one might expect of meta-inferential reasoners, people interpreted the *Misleading* evidence differently depending on what they had been told about the sender. In the TEAMMATE condition, people interpreted it as strong evidence in favour of the deceptive

[2] Both models included a random intercept for each individual, and were fit using the **brms** package (version 2.5.0) in R (version 3.4.3). Trials involving *Helpful* evidence were excluded because responses in favour of the *Truth* were at ceiling in both conditions.

Lure, while in the OPPONENT condition they were much more cautious. The effect of suspicion in the face of *Misleading* evidence represents a five-fold reduction in relative rates of choosing the *Lure* over the *Truth* (95% CI: 3.6 to 6.7).[3] Yet even in the TEAMMATE condition more than a quarter of responses to *Misleading* evidence were in favour of items other than the *Lure*, raising the possibility that some people took different views of the evidence than others. We return to this issue in our model-based analyses.

A further curiosity is that people also interpreted *Uninformative* evidence differently, depending on what they believed about the intention of the sender — this, despite the fact that the data had no explicit evidentiary value. When *Uninformative* evidence was provided, the *Lure* map was chosen with greater frequency in the TEAMMATE condition, where cooperation was expected. While the impact of suspicion was reduced in this instance (when compared with the case for *Misleading* evidence), the effect nonetheless represents a three-fold reduction in relative rates of choosing the *Lure* over the *Truth* (95% CI: 2.1 to 4.1). Given that the deceptive *Lure* was the smallest hypothesis in size compatible with the *Uninformative* evidence in three out of six cases, and the smallest in size compatible with the *Misleading* evidence in all cases (see Figure 6.3), a possible role for hypothesis size in the decision process suggests itself. We consider the nature of the meta-inferential assumptions that might account for this finding in our subsequent model-based analyses.

6.3 EXPERIMENT 2: SENDING DECEPTIVE INFORMATION

Experiment 1 demonstrated that people do take the likely intent of the sender into account when seeking to leverage the evidentiary value of information provided. Given that this is the case, do people account for this tendency in others in their own meta-inferential reasoning when they are acting as a sender? That is, do they seek to exploit the receiver's trust when they have it, and alter their strategy accordingly? In order to investigate these issues we invited people to play the deception game as a sender who was motivated to conceal the truth from their counterpart.

METHOD

Participants

We recruited 100 adults via Amazon Mechanical Turk, and paid them $1.25USD for 10-15 minutes minutes participation. Two of these participants were excluded for browser incompatibility. Data from a further 22 participants who failed to demonstrate

[3] This effect was quantified using our regression model extended to included an interaction between the type of evidence presented and level of suspicion.

a sufficient understanding of the experiment were excluded from subsequent analyses.[4] The remaining 76 participants were 46% female and aged 20-63 (median age 28.5).

Procedure

As in the first experiment, the cover story for the sender version informed people that they were taking part in an experiment based on an online game and that they would take the role of a pirate (the sender). On each trial, people were shown the genuine treasure map and three false maps, and were asked to select evidence to reveal to the explorer (the receiver).[5]

People's **sampling strategy** (deciding what evidence to disclose) was manipulated within subjects. In the CONTROL condition, the goal was to provide evidence that would help the receiver to correctly identify the genuine map. In both the TEAMMATE and OPPONENT conditions, the goal was to prevent the receiver from guessing correctly. Participants were told that the receiver was expecting evidence from a teammate (in the CONTROL and TEAMMATE conditions) or an opponent (in the OPPONENT condition). Participants were restricted in their choice of evidence to one of three options, namely: *Helpful*, *Misleading* or *Uninformative* evidence.

Experimental trials employed the same stimuli used in Experiment 1 (see Figure 6.3). However, an additional four filler trials involved new stimuli, with evidence designed to reduce tactical responding; that is, whilst a seemingly random pattern of dots was characteristic of *Uninformative* evidence in the experimental trials, similar random patterns in the filler trials could be used to rule out one or more maps. Additionally, the least informative evidence in each of the filler trials was not a random pattern, but a pattern bearing a resemblance to one of the four maps. These filler trials were not analysed.

Participants undertook a training exercise similar to that used in the first experiment. In the test phase, people saw each of the ten map sets (six experimental and four fillers) three times (once per condition), making 30 trials in all. The on-screen order of maps, as well as the order of evidence items was randomised on a trial by trial basis. Trial order was also randomised, with trials from each of the three conditions randomly interleaved. The participant's goal in each trial (corresponding to the three conditions) was clearly stated via on-screen instructions. On each trial people were required to choose evidence from amongst the available options that best achieved the stated goal of the trial, taking into account the four maps shown and the identity of the genuine map.

[4] Participants were excluded if they failed to select the *Helpful* evidence on at least 40% of the CONTROL trials (where the goal was to help the other player), or if they chose the *Helpful* evidence in 40% or more LOW SUSPICION trials (where the goal was to hinder, and double bluffing was unreasonable).

[5] Once again, to avoid reputational concerns, people were told to assume that they were facing a different explorer on each trial, one who had not played the game before, and was unaware of the pirate's identity.

Basic results

Experiment 1 established support for an intuitively reasonable notion: people take the sender's likely intent into account when determining the evidentiary value of information provided. The aim of the present experiment was to investigate whether people embody this intuition when motivated to deceive. Figure 6.5 plots the proportion of people choosing to provide the different types of evidence to a hypothetical receiver, as a function of experimental condition. The figure suggests that, as expected, people select content to convey according to their goal and the context in which the content will be interpreted. A Bayesian multinomial logistic regression, comparing a model with condition as a predictor to an intercept only model, revealed strong evidence in support of this finding ($BF_{10} > 10^6$).[6] In the Low Suspicion condition, where the aim was to conceal the truth from a player expecting help from a teammate, participants made liberal use of the *Misleading* option. In contrast, when people faced an opponent in the High Suspicion condition, they overwhelmingly preferred to reveal as little as possible, selecting the *Uninformative* option in almost every case. Overall, the effect of suspicion on sender participants represents a three-fold reduction in relative rates of actively misleading rather than simply limiting disclosure (95% CI: 2.4 to 4.6).Unsurprisingly, in the Control condition where the goal was to help, people were able to identify the evidence that the receiver would find most helpful, and almost always selected it.

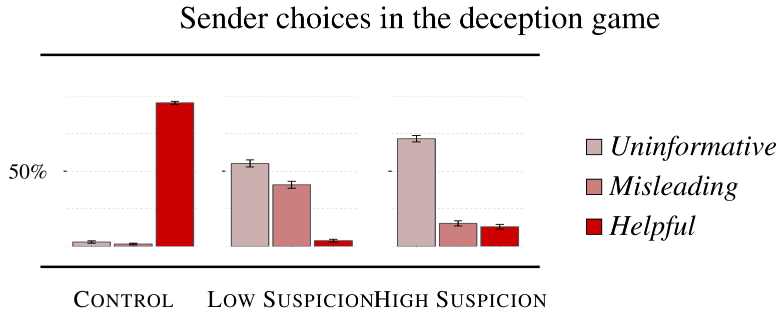


Figure 6.5: Proportion of participants selecting each response in Experiment 2. When acting as senders, people are only willing to provide the *Misleading* evidence when they believe the receiver does not suspect deception (left panel); when suspicion is high (middle panel) people overwhelmingly prefer the *Uninformative* alternative. In the control condition, people were asked to choose the option that would most benefit the receiver, and they did so consistently (right panel).

The pattern of behaviour across conditions — specifically, the change in the willingness to mislead — is largely what one would expect if people were reasoning about the inference of the receiver in a context sensitive manner. Despite this, the majority of

[6] Both models included a random intercept for each individual. Control trials were excluded because responses in favour of the *Helpful* evidence were at ceiling.

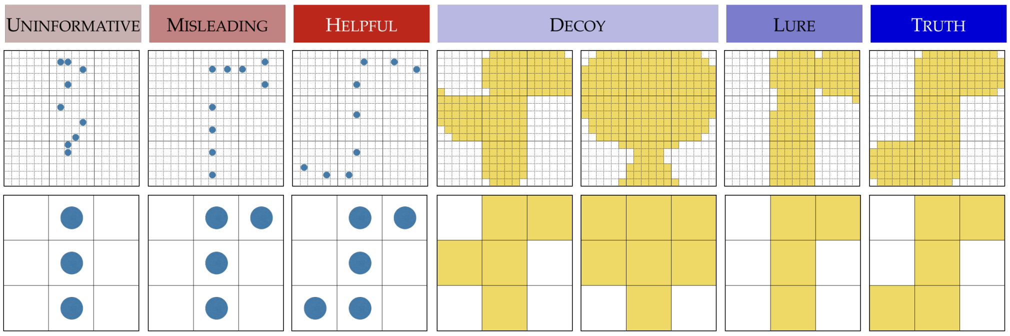


Figure 6.6: Model Implementation. We use a 3×3 grid approximation of the original stimuli to represent maps (hypothesis) and patterns of marked locations (evidence). A cell in the coarse grid representation (bottom row) is "on" if cells in the corresponding area of the original stimuli (top row) are "on".

senders in the LOW SUSPICION condition preferred to be uninformative. This seems contrary to the intuition that a meta-inferential reasoner expecting help should be reliably misled by misleading evidence, and the truth better concealed as a result. To better understand the assumptions that might drive a quantitative shift in sender preference, but not a qualitative reversal, we turn now to our computational analysis of the deception game.

6.4 MODELLING META-INFERENCE IN THE DECEPTION GAME

Our two experiments were designed to investigate how communicating reasoners might take account of the inferences of their interlocutor in situations where a variety of assumptions might reasonably hold. Taken together, the pattern of responses across both experiments appear largely consistent with an intuitively reasonable approach to meta-inferential reasoning. When receivers think that the sender can be trusted, they leverage that assumption to reason beyond the data; if the sender cannot be trusted no such leverage occurs. Senders, seemingly aware of this, are thus more willing to mislead trusting receivers than they are suspicious receivers. But if the veracity of data is not in question (because lying is not an option), then precisely what is it that mediates its strength as evidence? What are receivers (and consequently senders) sensitive to? Using the computational framework outlined at the start of the paper we can model various assumptions that might underpin this sensitivity, and ask which of these best captures the patterns of behaviour observed.

		Model			
Condition	Schema	Weak	Strong	One Step	Reciprocal
(Receiver)					
Teammate	$\langle Rec_T \rangle$	*Weak*	*Strong*	*Help* (*Weak*)	*Help* (. . . (*Weak*))
Opponent	$\langle Rec_O \rangle$	*Weak*	*Strong*	*Hinder* (*Weak*)	*Hinder* (. . . (*Weak*))
(Sender)					
Control	*Help* ($\langle Rec_T \rangle$)	*Help* (*Weak*)	*Help* (*Strong*)	*Help* (*Help* (*Weak*))	*Help* (. . . (*Weak*))
Low Susp.	*Hinder* ($\langle Rec_T \rangle$)	*Hinder* (*Weak*)	*Hinder* (*Strong*)	*Hinder* (*Help* (*Weak*))	*Hinder* (*Help* (. . . (*Weak*)))
High Susp.	*Hinder* ($\langle Rec_O \rangle$)	*Hinder* (*Weak*)	*Hinder* (*Strong*)	*Hinder* (*Hinder* (*Weak*))	*Hinder* (. . . (*Weak*))

Table 6.1: **Four alternative models of sender and receiver behaviour in the deception game**. Receivers are assumed to reason according to Equation 6.1 on the basis of the sampling assumption defined, and to respond in proportion to their strength of belief in each hypothesis. Senders are assumed to select evidence from amongst the options provided with probabilities defined according to Equation 6.2. "*Help* ()" and "*Hinder* ()" denote opposite forms of intentional sampling where the selection of data is biased according to the sender's goal. "(. . .)" denotes a recursive and reciprocal assumption. A "*Weak*" sampling assumption means that evidence is used solely to disconfirm hypotheses, while a "*Strong*" assumption implies that data constitutes stronger evidence for smaller hypotheses, in accordance with the size principle. The schema column illustrates the common relationship amongst the sender and receiver assumptions within each model. See main text for further details.

Model implementation

To model the deception game in a tractable way we use a simplified 3×3 grid to represent the experimental stimuli, as illustrated in Figure 6.6. Thus both the hypothesis space $\mathcal{H}$ and the space $\mathcal{X}$ from which evidence is drawn, consist of the $2^9 = 512$ possible patterns of on/off grid cells. For any given trial, the hypothesis space is further restricted to one of the four maps in question by means of a trial-specific prior that rules out the remaining possibilities. No such restriction is placed on the evidence X, since people playing the role of the receiver were not aware of any restrictions regarding the selection of evidence, save for the fact that it was constrained to be truthful.

In the analyses that follows we consider the four models of deceptive communication summarised in Table 6.1. Our goal is use the models to investigate which set of assumptions best captures the behaviour observed across both experiments. Each model is actually a family of nested sub-models corresponding to the five experimental conditions (two receiver: Teammate and Opponent, and three sender: Control, Low Suspicion and High Suspicion). The assumptions of the trusting and suspicious receivers lie at the core of each model, so we turn to these first.

The first model we consider, the Weak model, captures the notion that the receiver (whether trusting or suspicious) makes no assumption about the relative likelihood of an observation under each of the hypotheses in question. In terms of the computational framework, we capture this with a *weak sampling* assumption which defines the probability of an observation x in the event that hypothesis h holds, as

$$P(x|h) = \begin{cases} P(x) & \text{if } x \in h \\ 0 & \text{otherwise} \end{cases} \tag{6.3}$$

In other words, observations are used to rule out hypotheses that do not fit the evidence (i.e. where $x \notin h$), but the evidence is otherwise *uninformative* about the remaining hypotheses. The fact that a receiver adopts such an assumption, however, does not necessarily imply an absence of meta-inferential reasoning. A weak sampling assumption may capture the responses of a cautious receiver who is simply unwilling to impute any *particular* assumption on the part of the sender, perhaps in response to perceived variability in the reasoning style of others. Instead she may choose to rely only on the fact that the data itself was not false (consistent with the instructions given).

Yet in the context of the deception game, the receiver might reasonably justify a stronger assumption that leverages a perceived dependency between the evidence observed and the truth of the matter in question. The fact that only positive (and reliable) evidence may be provided constrains the sender in his choice. And importantly, the less that a given hypothesis entails (i.e. the fewer the observations compatible with it), the more the sender is constrained. According to the STRONG model, the receiver takes account of this by making a *strong sampling* assumption, where

$$P(x|h) = \begin{cases} \frac{1}{|h|} & \text{if } x \in h \\ 0 & \text{otherwise} \end{cases} \tag{6.4}$$

Once again, such a sampling assumption need not indicate a lack of meta-inference on the part of the receiver. Rather, as long as the receiver is unwilling to assume that evidence selection is biased one way or another, then this size principle (Tenenbaum & Griffiths, 2001a), which gives greater weight to smaller hypotheses, seems justified.

The next logical step in the progression of meta-inferential assumptions is for receivers to assume that senders also engage in meta-inference. A receiver who reasons in this way will expect the sender to bias selection in favour of evidence that is more informative (in the TEAMMATE condition) or less informative (in the OPPONENT condition). A single level of *"the receiver thinks that the sender thinks..."* reasoning may be modelled straightforwardly in the computational framework by a sequential instantiation of Equations 6.1 and 6.2. The receiver's assumption about how the sender intentionally biases the selection of evidence is captured by the α parameter: $\alpha = 1$ implies a bias in favour of more informative evidence, while $\alpha = -1$ implies the converse. For the receiver, such an intentional sampling assumption either increases or decreases the evidentiary weight that data would otherwise have under a more basic assumption, depending on the perceived intent of the sender.[7] This essentially asymmetric reasoning style, where the receiver attempts to reason one step further than the sender, forms the basis of the ONE STEP model.

The RECIPROCAL model, in contrast, describes the case where each reasoner credits the other with taking meta-inferential reasoning to its logical extreme. That is, for any assumption that the receiver makes about the sender, the sender reciprocates that assumption by assuming that she makes it, and vice versa – here there is no imbalance with regard to depth of reasoning. Notwithstanding the way that recursive and reciprocal reasoning proceeds, computationally speaking, it can only be satisfied by finding a meta-inferential equilibrium - a fixed point beyond which further recursive reasoning does not change the outcome (Shafto et al., 2014).[8]

Turning to the models of sender behaviour (see Table 6.1), each model represents an instance of Equation 6.2 that defines the probability with which the sender selects evidence from amongst the options provided. The value of the parameter α is matched to the stated aim of each condition: in the CONTROL condition, where the goal was to help the receiver uncover the truth, we set $\alpha = 1$, while in the LOW SUSPICION and HIGH SUSPICION conditions, where the goal was to hinder, we set $\alpha = -1$. In addition to capturing the sender's intent, a sender model must define the sender's beliefs about the receiver's sampling assumption. For each of the models considered, the sender makes the same assumption about the receiver as the model itself does.[9] Each sender model thus proceeds straightforwardly from the corresponding receiver model. In the CONTROL and LOW SUSPICION conditions, the sender is assumed to model the would-be receiver in line with TEAMMATE model, while in the HIGH SUSPICION condition the sender is assumed to target the OPPONENT receiver.

MODEL-BASED ANALYSES

We can now use the models we have defined to examine people's reasoning within the deception game. We are interested in whether people reasoned probabilistically about the generative process underlying communication within the game, and how this changed based on whether cooperation or competition was expected. Each model we have defined represents a different trade-off between the generality of the underlying assumptions and

[7] In the computational framework, the evidentiary weight of data ultimately stems from either its power to disconfirm hypothesis (weak sampling) or from the size principle (strong sampling). We use a weak sampling assumption as the ground term in the ONE STEP model.

[8] A single-state fixed point is not guaranteed — bi-stable equilibria may exist under reasonable assumptions, for example. However, for each of the four models a single-state fixed point exists.

[9] This need not be the case, we might wish to model a disconnect where the sender's assumption about the receiver does not match the model's direct assumption about the receiver. However, we constrain the models to be coherent in this way because it is both a theoretically reasonable and parsimonious starting point.

the degree to which inference is driven by those assumptions. So a comparison between model predictions and behavioural data allows us to assess how sensitive people were to the relative likelihood of evidence under one hypothesis over another.

For the receiver models, we compared predictions with responses (aggregated across all participants) to each of the 18 combinations of stimuli (six map sets and three types of evidence). Model fit, as measured by Root Mean Squared Error (RMSE), was assessed separately for each type of evidence as well as on an overall basis. Model predictions and fits to the choices of our receiver participants are shown in Figure 6.7 for the TEAMMATE condition and in Figure 6.8 for the OPPONENT condition. Because the *Helpful* evidence is incompatible with all but one hypotheses in every case, the receiver models predict that the receiver will identify the truth with complete certainty, fitting our behavioural data almost perfectly. For this reason, we omit those predictions from our plots, but include them in the calculation of overall fit.

In each of the models considered, the receiver makes a progressively stronger assumption about how the sender chooses what to reveal. When the receiver trusts the sender to cooperate, each additional assumption leads to progressively tighter conclusions. This cumulative ratcheting effect can be seen in Figure 6.7. The figure shows that a receiver who adopts a weak sampling assumption is not easily misled. But one who believes that the sender is trying to help and that he reciprocates her assumptions, will leap to the wrong conclusion. Less intuitively, perhaps, this ratcheting effect applies even when the evidence is seemingly uninformative: information that would otherwise be ambiguous can still tighten conclusions, by virtue of the size principle. Indeed, comparing the predictions of the STRONG model to people's choices in response to *Uninformative* evidence suggests that the size principle was in effect.

This becomes important when considered from the perspective of the sender who wishes to conceal the truth. The sender's goal in this case (following directly from Equation 6.2) is to do what he can to reduce the receiver's belief in the true hypothesis (at least in relative terms). Certainly a misleading message has the potential to achieve this. But as we have seen, if the receiver's inference is consistent with the size principle then even a message that reveals no new information may act to reduce her belief in the true hypotheses. Thus, when considering these alternatives, the sender may conclude that the additional information disclosed by misleading yet informative evidence is not sufficiently offset. Figure 6.9 (TEAMMATE condition) reveals that this is the case in the deception game. The figure plots the accuracy with which receivers identify the genuine map given the different types of evidence. It shows that, according to model predictions, the *Uninformative* evidence is always most effective at keeping the truth from the receiver (compromising accurate identification as a result). In the case of the WEAK model, this follows directly from what it means to be uninformative. For the remaining models, it follows from the size principle. As a direct consequence, the sender models for the LOW

Model fits to Receiver choices: TEAMMATE condition

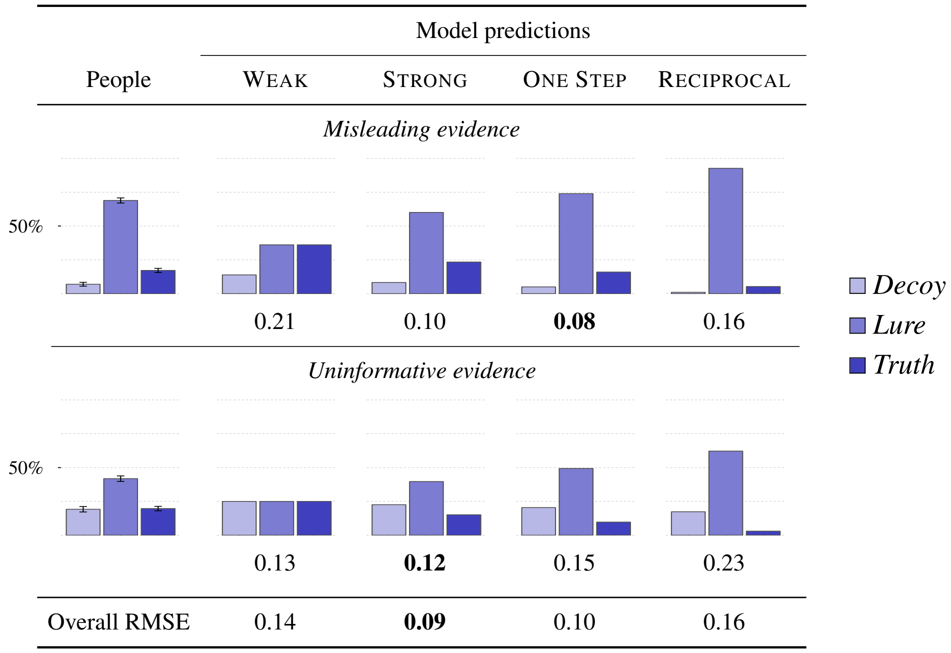


Figure 6.7: Predictions of four models compared with the choices of people playing the role of receiver in the TEAMMATE condition. The models are arranged in order, based on the strength and complexity of the assumptions involved. The WEAK model captures no constraints on the data, and represents a stance where the generative process is effectively ignored by the receiver. The STRONG model assumes only that the data represents positive evidence of the concept in question, and is otherwise unbiased by the sampling process. The ONE STEP model builds upon the STRONG model by assuming that the sender biases selection towards more informative content. The RECIPROCAL model assumes not only that the sender is trying to help in this way, but that both sender and receiver share a mutual awareness of each other's assumptions. The ratcheting effect of progressively layered assumptions can be seen in the top row: the more complex models increasingly favour the *Lure* item reflecting the fact that stronger assumptions licence stronger conclusions. The numbers below each graph show the model fits, as measured by Root Mean Squared Error (RMSE), with lower numbers indicating a better fit. The row at the bottom of each table shows the overall fit for each model in the given condition. While the STRONG model best captures the behaviour of participants in the TEAMMATE condition when evidence is *Uninformative*, when the evidence is *Misleading* it appears as though participants adopted a stronger assumption (although differences between the two are minor).

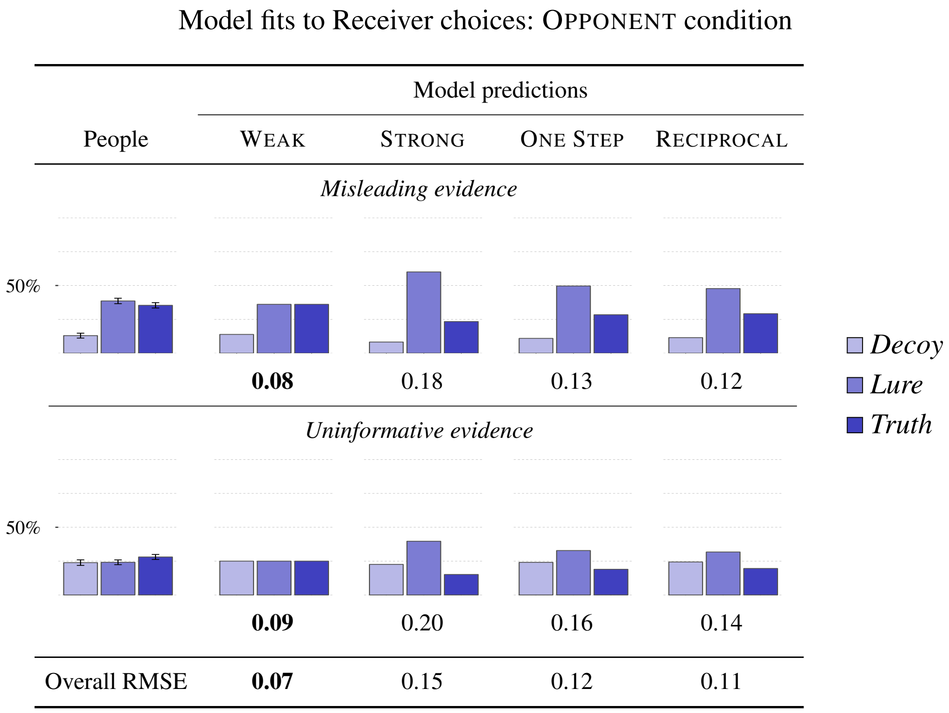
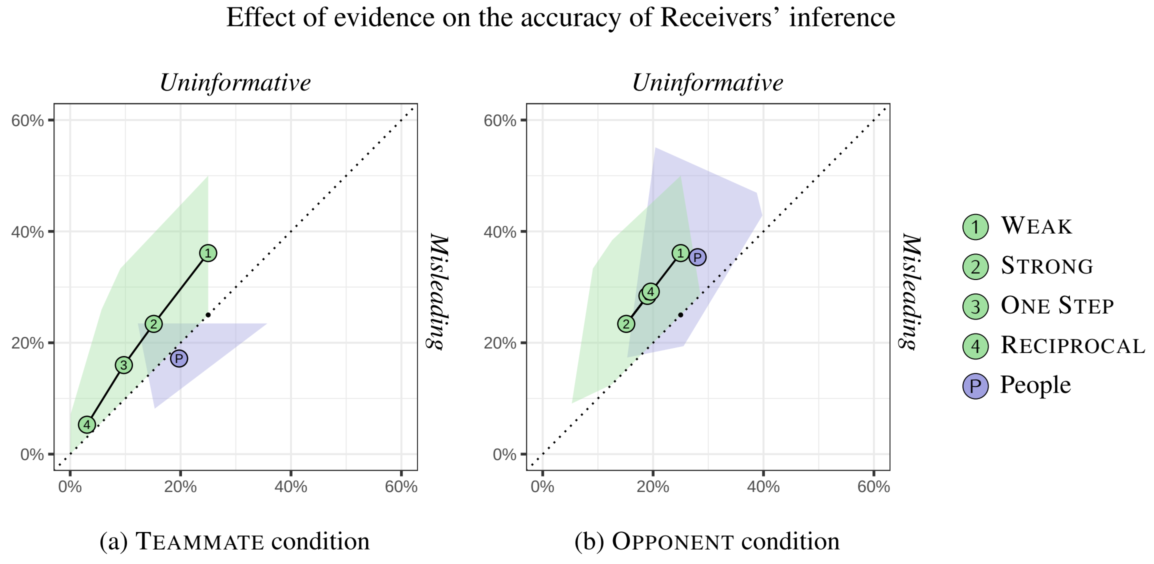


Figure 6.8: Predictions of four models compared with the choices of people playing the role of receiver in the OPPONENT condition. The WEAK and STRONG models, which are not context sensitive, make exactly the same predictions as described for the TEAMMATE condition. The ONE STEP and RECIPROCAL models are context sensitive however. In the OPPONENT condition, these models assume that the sender is trying to conceal the truth rather than reveal it ($\alpha = -1$). Their respective predictions reflect a trade-off between uninformativeness and the size principle, falling “between” the predictions of the WEAK and STRONG models. As described in the previous figure, lower RMSE values represent better model fits. In the OPPONENT condition it is the WEAK model, where the receiver assumes that the sender will be maximally uninformative (effectively disregarding the process by which the data is generated), that best captures people’s behaviour in the OPPONENT condition.

SUSPICION condition predict a preference for choosing the *Uninformative* option, as Figure 6.10 (LOW SUSPICION condition) shows.

If being uninformative is an effective way of concealing the truth from a trusting receiver, consider then what inference a receiver in the OPPONENT condition should draw. As we have seen, a reasonable starting point for this receiver is to assume that the sender prefers to be uninformative. In our model, information becomes more informative with respect to a small hypothesis than to a large one, and hence less likely to be produced by an uncooperative sender. Yet the size principle dictates the reverse — namely smaller hypotheses consistent with the evidence are more likely. Under the assumptions of the model, this leads the receiver to find a balance between two opposing forces. As a



(a) TEAMMATE condition

(b) OPPONENT condition

Figure 6.9: Receiver accuracy based on the type of evidence provided. The plotted points represent model predictions (green circles) and people's performance (blue circles) aggregated across the six sets of stimuli, while the polygons illustrate the spread of predictions – each vertex corrsesponds to a single set. Model predictions in the TEAMMATE condition illustrate the effect of deceptive evidence on a trusting receiver. As people adopt stronger assumptions, their attempts to uncover the truth become increasingly inaccurate, leading to almost complete inaccuracy in the RECIPROCAL case. In contrast, the OPPONENT models predict that stronger assumptions lead to little change in receiver accuracy. The plots illustrate the connection between the sender and receiver models. A sender who wishes to keep the truth from the receiver should choose the type of evidence that leads to the lowest accuracy. Regardless of the strength of the receiver's assumption, and whether or not they trust the sender, the model predictions indicate that the *Uninformative* evidence consistently leads to lower accuracy. This is in contrast to the observed accuracy of our receiver participants, who were least accurate in the TEAMMATE condition when presented with *Misleading* evidence.

consequence, the inferences predicted by the ONE STEP model are less certain than those of the STRONG model but sharper than those of the WEAK model (see Figure 6.8).

Somewhat paradoxically, as the receiver becomes less prepared to reason beyond the data, the sender pays a lower penalty for disclosing information. Thus, in contrast to the TEAMMATE models which predict that stronger assumptions lead to sharper conclusions, the OPPONENT models show no such pattern. Instead, progressively stronger assumptions produce predictions that follow the pattern of dampening oscillation shown in Figure 6.8, and converge on the RECIPROCAL model. The predictions of the RECIPROCAL model, however, which represent an equilibrium where neither sender nor receiver "out thinks" the other, seem unintuitive. A more intuitive way for the receiver to take the sender's reasoning to its "logical" extreme, is to consider that he will display an optimal bias towards being uninformative (Hespanha, Ateskan, & Kizilocak, 2000). As a consequence, the receiver should not attempt to reason beyond what the data falsifies - i.e. she should

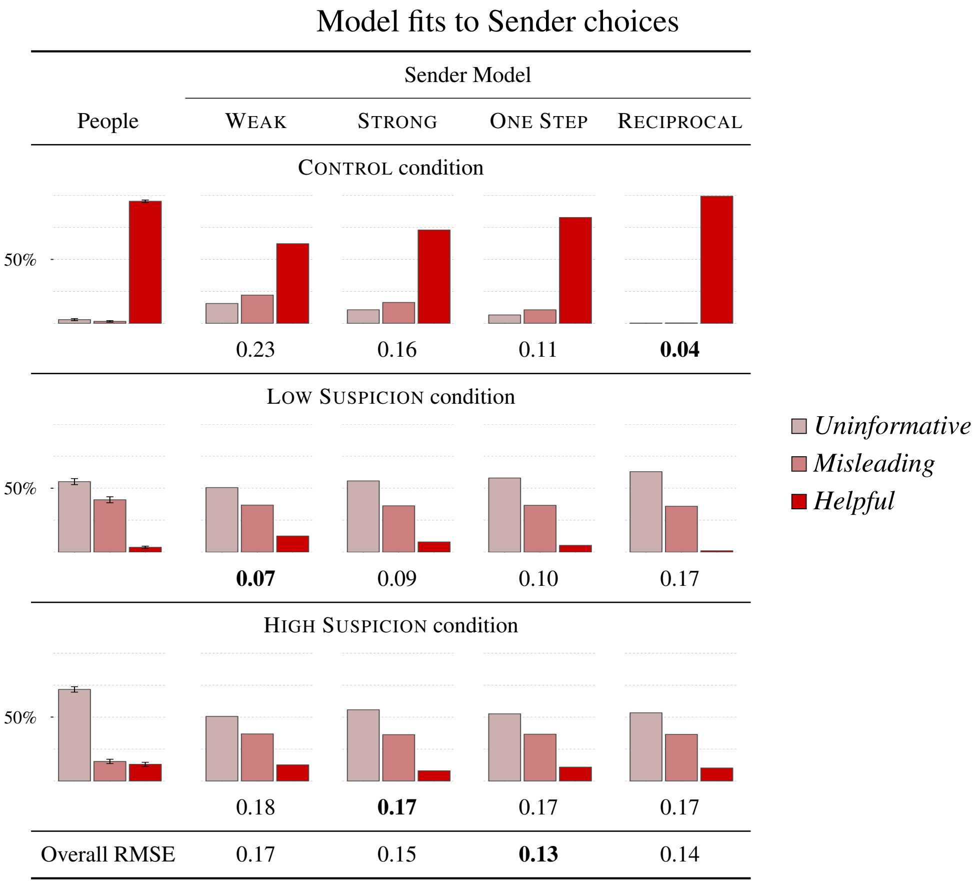


Figure 6.10: Predictions of four models compared with the choices of people playing the role of the Sender. In the CONTROL condition, where the sender's goal is to reveal the truth, all models predict a strong preference for the most informative (*Helpful*) message, as exhibited by participants. In the LOW SUSPICION and HIGH SUSPICION conditions however, the sender has the opposite goal – to hide the truth. The relative homogeneity of the predictions under these conditions reflects the unanimous prediction of the underlying receiver models that *Uninformative* evidence is most effective in this regard. Model fits (RMSE) are shown beneath each plot, and averaged across conditions in the bottom row of the table. Lower numbers represent better fits. While the models capture participants' overall preference for the *Uninformative* option in the LOW SUSPICION and HIGH SUSPICION conditions, the qualitative reduction in the use of the *Misleading* option is not predicted. (as the relatively poor fits in the HIGH SUSPICION condition indicate).

adopt a weak sampling assumption.[10] As Figure 6.8 shows, the WEAK model best captures the behaviour of people in the OPPONENT condition. Whether we choose to model progressively stronger assumptions by an increasing bias towards the uninformative (larger negative α), or through increased depth of meta-inference, the predictions of the STRONG model, which assumes that data selection is unbiased, represent an upper bound

[10] In terms of the computational model, taking the limit as $\alpha \to -\infty$ of Equation 6.2, yields a likelihood function compatible with Equation 6.3 – i.e. weak sampling.

on the strength of inference expected. Thus, once again, *Uninformative* evidence is most effective at concealing the truth from the receiver (see Figure 6.9 – OPPONENT condition), and the sender models predict a preference for using it (see Figure 6.10 – HIGH SUSPICION condition).

DISCUSSION

Our analysis thus far has demonstrated that our framework captures two important properties of the receiver's inference. Firstly, it predicts the (obvious) effect of information content: trusting receivers draw stronger conclusions from more informative evidence. Secondly, and more importantly, it predicts an effect of assumption strength: stronger assumptions lead to stronger conclusions. While our analysis was not intended as a parameter fitting exercise, the good qualitative fits with theoretically motivated sampling assumptions suggests that the behaviour of participants in our receiver experiment is consistent with a sampling assumptions explanation. The strength of the assumption adopted depends on the perceived intent of the sender, as dictated by the setting in which communication takes place. In the TEAMMATE condition, people reasoned beyond the data, giving greater weight to those hypotheses under which the data might make sense ($\alpha > 0$). In the OPPONENT condition, where any attempt to reason beyond the data might be exploited, people behaved in line with a weak sampling assumption, using data only to falsify hypotheses ($\alpha \ll 0$).

When it comes to capturing the behaviour of our sender participants, however, the qualitative fits in Figure 6.10 are less compelling. While the models matched response patterns in the LOW SUSPICION condition reasonably well, and captured people's overall preference for uninformative evidence in the HIGH SUSPICION condition, they failed to predict context sensitive meta-inference. They could not account for the disparity people showed between conditions as a function of suspicion. In the case of the ONE STEP and RECIPROCAL models, which were specified with context-specific assumptions in mind, this represents a challenge to the sampling assumptions account.

It is instructive at this point to recall the intuition behind the sender's decision. The virtue of a misleading utterance is that it appears (to a trusting receiver) to conform closely to communicative norms. Consequently, the intuition goes, the receiver will accord it a stronger inferential boost than they would a less informative utterance, promoting a strong yet misleading conclusion. But the ultimate goal for the sender (as we have framed it) is not to maximise the receiver's belief in one of the false hypotheses, but to minimise her belief in the true hypothesis. Under a weak sampling assumption ($\alpha \ll 0$), the evidence receives no inferential boost and so of course the sender should prefer to use uninformative evidence. Similarly the so-called strong sampling assumption ($\alpha = 0$) is not sufficiently strong as to warrant a change in preference. The intuition behind reciprocal and recursive meta-inference however, is that each layer of additional "he thinks, she thinks...” reasoning acts to increase the inferential boost that informative evidence receives relative to less informative evidence. Thus, a sufficiently strong (recursive) assumption on the part of the receiver should justify a reversal in the sender’s preference so that he prefers to mislead. Yet what our analysis has shown is a limitation of our framework in this regard. As it stands, our framework fails to predict the necessary *interaction* between the information content of evidence and the strength of assumption in determining the strength (or rather weakness) of inference. At least not in the way that matters — the cumulative ratcheting effect of progressively stronger assumptions preserves and never reverses the relative superiority of uninformative evidence in limiting receiver accuracy. Thus we have essentially demonstrated that two key (intuitively reasonable) meta-inferential assumptions — that trusting receivers make stronger assumptions (of a positive information bias) than suspicious ones, and that strong assumptions more strongly benefit informative content — are insufficient to explain sender behaviour in this situation. What additional or alternative assumptions might senders be making when deciding whether to conceal information or to actively mislead?

The receiver model fits shown in Figure 6.7 reveal a potential clue. The figure shows that while the STRONG model provides a better fit to people’s responses to the *Uninformative* evidence, the ONE STEP model provides a better account of the *Misleading* evidence. This suggests that receivers may make stronger assumptions on the basis of more informative evidence. This is an idea with some intuitive appeal; for example, if on hearing your words I believe you have chosen them carefully, I may be more likely to infer what you have implied.

If receivers’ assumptions are sensitive in this way, or the sender believes them to be, it should change the nature of the sender’s evaluation. Instead of comparing what a receiver making a fixed assumption would infer from two alternative messages, the problem becomes one of comparing the alternatives under the different assumptions they would induce. For example, if senders assume that receivers reason beyond the data only when norms of relevance are upheld, then this might increase the incentive for the sender to mislead a trusting receiver. In the following section we extend our computational framework to accommodate these kind of “content-sensitive” sampling assumptions.

6.5 MODELLING CONTENT-SENSITIVE SAMPLING ASSUMPTIONS

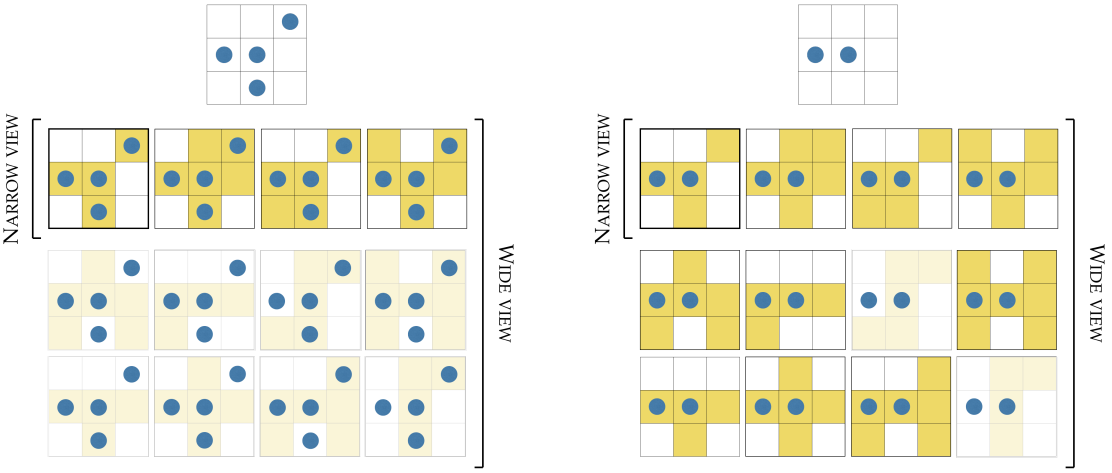
The computational models we have considered are based on a simple premise: namely, that people (as receivers) use information to rule out competing hypotheses and are therefore sensitive (as senders) to its evidentiary value when choosing information to convey. A given sampling assumption reflects a particular estimate about the degree to which evidence selection is biased in favour of the informative ($\alpha > 0$) or the uninformative ($\alpha < 0$). In the models we have examined, this estimate has been pre-determined solely on the basis of whether cooperation or competition is expected: that is, $a = +1$ (cooperation), or $\alpha = -1$ (competition). Although this approach has the virtue of simplicity, it fails to account for the *ostensive* nature of cooperative communication, in which the goal is not merely to produce utterances that are informative, but also ones that are easily recognised as such. In order to clarify how sampling assumptions might account for our experimental results, it makes sense to consider the notion of a receiver whose assumptions are sensitive to message content, and the implications for sender behaviour that this might have.

To see how a receiver might adjust her sampling assumption after observing the evidence provided, imagine that we (as an observer) already know which is the true hypothesis and are aware of the possibilities that the receiver is considering. If the aim of a cooperative sender is to select evidence that reduces the receiver's uncertainty about the matter at hand, then we can estimate his selection bias after seeing the evidence he selects, in the same sense that we might estimate the bias of a coin after seeing only a single toss. Of course, the receiver does not know the true hypothesis, but she may nonetheless form an estimate by considering all possibilities in order to determine the "likely helpfulness" of the information provided. We may model the receiver's assessment of the sender's likely helpfulness via the following straightforward extension of Equation 6.1:

$$P_{\text{Receiver}}(h|x) \propto \sum_{s \in \mathcal{S}} P_{\text{Sender}}(x|h,s)P(h)P(s) \qquad (6.5)$$

where s represents an assumption that the receiver makes about the sender's sampling strategy, and $\mathcal{S}$ denotes the set of alternative strategies considered. As a simplifying assumption which should reasonably hold in the context of our experiments, we assume that the receiver considers the sender's sampling strategy to be independent of the true hypothesis.

If receivers are vigilant for ostensive signs of cooperation, then there are implications for the sender. In practical terms, the cooperative sender might select information to reduce the receiver's uncertainty not only about the hypotheses under consideration but also about the way in which the information was sampled. Because senders and receivers will not in general have perfect mutual information about each other's knowledge state, it makes sense for helpful senders to provide information that would be judged as being helpfully sampled, independent of any particular reciprocal assumption about prior knowledge. Intuitively, for example, the evidence sample shown in Figure 6.11(a) feels more likely to have been provided by a competent and helpful sender than does the evidence sample shown in Figure 6.11(b), despite the fact that each sample is equally ambiguous in the narrow sense. In terms of our model, a sender who wishes not only to be informative but also to be seen to be informative, selects information consistent with a strong selection bias ($\alpha \gg 0$, for example) under an appropriately general prior distribution. The deceptive sender may choose to mimic the helpful sender by making his informative intention clear, all the while selecting evidence intended to misinform. By being uninformative he may leave his sampling method (and his meaning) ambiguous.



(a) Ostensively clear evidence

(b) Ambiguous evidence

Figure 6.11: **Examples of ostensively clear and ambiguous evidemce.** In an example scenario drawn from the deception game, the receiver must use the evidence sample given (top) to distinguish amongst four hypotheses (narrow view) drawn from a larger set of possibilities (wide view). In interpreting the weight of a given piece of evidence, senders and receivers must take a stance regarding what constitues good evidence. Under a narrow view, evidence is informative only to the degree that it distinguishes amongst those hypotheses being directly considered. In this case both evidence samples are equivalent – each is equally ambiguous under a weak sampling assumption (because they rule out none of the four hypotheses in the narrow view), or equally helpful under a strong sampling assumption (identifying the target (dark border) due to the size principle). If instead, the receiver interprets informativity in the broader sense then the picture changes. In this case, evidence sample (a) is helpful in distinguishing the target from a considerably wide range of alternatives (pale yellow), whereas sample (b) is relatively unhelpful even in the wider context. Because the meaning of sample (a) is ostensively clear in this context it licences a stronger sampling assumption than sample (b), which remains relatively ambiguous.

Condition	Schema	Model
(Receiver)		
Teammate	$\langle Rec_T \rangle$	$Ostensive \equiv \frac{1}{3}(Strong) + \frac{2}{3}(Help_{\supset}(\ldots(Weak)))$
Opponent	$\langle Rec_O \rangle$	$Hinder\ (Ostensive)$
(Sender)		
Control	$Help\ (\langle Rec_T \rangle)$	$Help\ (Ostensive)$
Low Suspicion	$Hinder\ (\langle Rec_T \rangle)$	$Hinder\ (Ostensive)$
High Suspicion	$Hinder\ (\langle Rec_O \rangle)$	$Hinder\ (Hinder\ (Ostensive))$

Table 6.2: **The Ostensive model of sender and receiver behaviour**. The sender and receiver models for each condition follow the same model schema as the models introduced previously (see Table 6.1). The core *Ostensive* assumption corresponds to a reasoner who believes (with probability $p = \frac{1}{3}$) that the data is strongly sampled or (with probability $p = \frac{2}{3}$) that the data is helpfully sampled. The prior probabilities for the two assumptions were chosen to match the proportion of uninformative and informative stimuli used in the experiment (1:2). $Help_{\supset}(\ldots(Weak))$ denotes a recursive and reciprocal assumption, based on a general prior distribution (over a superset of the hypotheses currently under consideration). See main text for further details.

To complete our content-sensitive model (which we shall refer to as the Ostensive model), we need to specify the set of strategies $\mathcal{S}$ that the receiver considers in the typical course of cooperative communication. For simplicity we consider two strategies only, but in general we could integrate over any aspect of the model specification, such as the depth of recursion, the value of α and so on. To reflect the possibility that the evidence was selected by a helpful sender we adopt the sampling assumption from the Reciprocal model (see Table 6.1 – Teammate condition). The alternative assumption that the receiver considers is that an indifferent sender selected the information at random ($\alpha = 0$), which is equivalent to a strong sampling assumption.This *Ostensive* sampling assumption is intended to describe the receiver's inference in the Teammate condition. The sampling assumptions that complete the Ostensive model are shown in Table 6.2.

In the next section, we apply this new model (in addition to the previous ones) to the data from Experiments 1 and 2. In order to investigate the presence of individual differences in reasoning, we focus on two sub-groups of participants that we identified in a *post hoc* fashion upon visual inspection of the data, as described below. Although this grouping is *post hoc* and the corresponding analysis should be taken with caution, we find that it (and the associated model fits) is revealing about the different kinds of reasoning that occur in deceptive communication.

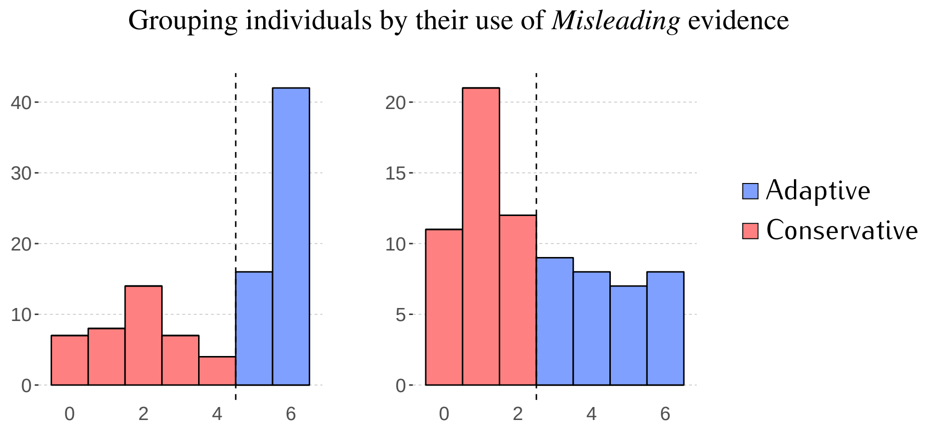
Experimental results: Individual differences in content sensitivity

The results of our two experiments demonstrate that people's communicative inferences take into account the context in which communication takes place and whether cooperative norms can be taken for granted. Moreover, the data so far suggest that for receivers at least, people's reasoning is sensitive to context (suspicion level). However, context sensitive tailoring of deceptive strategy on the part of the sender is less evident. In order to investigate whether people are sensitive to the possibility that message content may signal the sender's intent, we now take a closer look at the response distributions of both receivers and senders.

Turning first to our receiver participants, upon visual examination of the data it appeared that there were two qualitatively distinct patterns of behaviour based on how people responded to the *Misleading* evidence in the TEAMMATE condition. As Figure 6.12(a) reveals, the relevant response distribution is bi-modal. In addition, we used a Bayesian model to infer two independent binomial response rate parameters from the given response distribution. The model favours the same division that we identified by visual inspection. Further a Bayes' factor analysis revealed strong support for a model with two independent response rates over a model assuming only one ($\mathrm{BF}_{10} > 1{,}000$). We therefore defined, in a *post hoc* fashion, two qualitatively distinct groups. The Adaptive group, consisting of all participants who were consistently misled (choosing the *Lure* on five or six out of six relevant trials), appeared to be sensitive to the perceived intent of the sender and to adapt their assumptions accordingly. In contrast, the other participants, which we have labelled Conservative, appear to be largely *insensitive* to the sender's likely goal, displaying comparable conclusions in either condition.

The responses of receiver participants aggregated according to these groups are shown in Figure 6.13. The Adaptive receivers drew stronger conclusions when evidence was *Uninformative* as well as *Misleading* in the TEAMMATE condition but showed a very different pattern in the OPPONENT condition, suggesting they were sensitive to the sender's intent. In contrast, the Conservative receivers responded similarly regardless of the nature of the sender or whether the evidence was *Misleading* or *Uninformative*.

We can apply a similar analysis to the sender data from Experiment 2. For the sender, the essential decision in each trial is whether to attempt to actively mislead the receiver or instead to just be as uninformative as possible. Where the sender stands in this regard should be influenced by their assumptions about the receiver – there is little point in revealing more than is necessary to a receiver who is unlikely to take the bait. We therefore divided people in two groups based on how frequently they chose to provide the *Misleading* evidence in the LOW SUSPICION condition. Although the relevant response distribution, shown in Figure 6.12(b), is not as clearly bi-modal as was the case in our receiver analysis, the division into two groups is supported by the same Bayesian analysis used previously, applied in this case to the relevant sender response data ($\mathrm{BF}_{10} > 1{,}000$). Thus, Adaptive senders are those who chose to mislead on three or more of the six relevant trials. In analogy with the receiver groups, the remaining participants comprise the Conservative group.



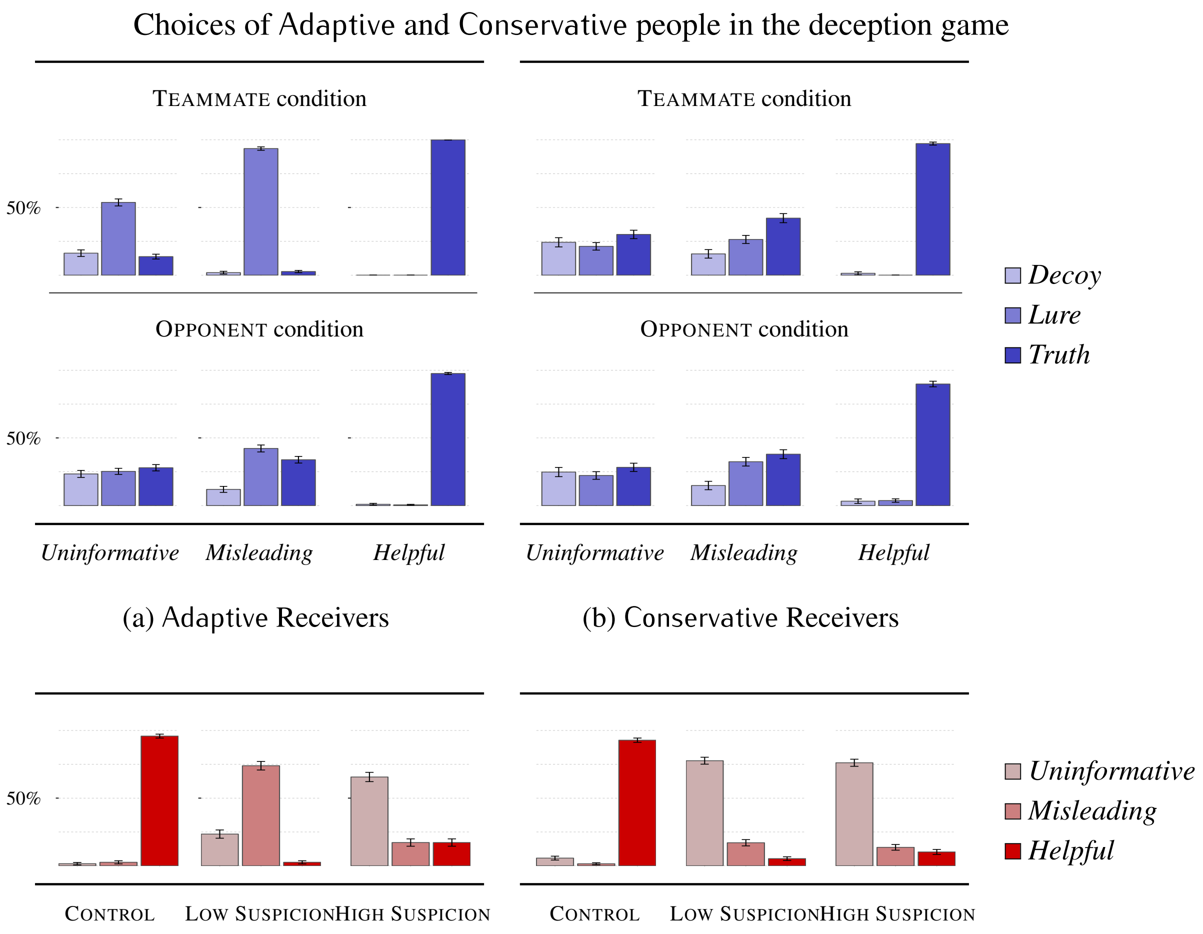
(a) Receivers

(b) Senders

Figure 6.12: **Use of *Misleading* evidence by receivers and senders.** The histograms show the number of trials (out of six) that (a) receivers in the TEAMMATE condition chose the *Lure* item in response to *Misleading* evidence, and (b) senders in the LOW SUSPICION condition chose to provide the *Misleading* evidence. The vertical axis indicates the number of people responding with the frequency given on the horizontal axis. For a post hoc analysis, participants were separated into two groups on the basis of visual inspection of the response distributions. The dashed line separates Conservative participants (red bars) to the left, and Adaptive participants (blue bars) to the right. The receiver distribution is clearly bi-modal, while the sender distribution is less so. Nonetheless, a Bayesian analysis revealed strong evidence in favour of the partitioning illustrated (see main text for detail).

Sender choices for Adaptive and Conservative people are shown in Figure 6.13(c) and (d) respectively. The figure shows that Adaptive senders, defined on the basis of their preference to actively mislead an unsuspecting receiver in the TEAMMATE condition, reverse this preference when the receiver is likely to be alert to the deception in the OPPONENT condition. In contrast, Conservative senders appear insensitive to the presence or absence of trust on the part of the receiver, strongly favouring the *Uninformative* option in both the LOW SUSPICION and HIGH SUSPICION conditions.

Taken together, the above analyses suggest that there may be a meaningful link between Adaptive senders and Adaptive receivers, and between Conservative senders and receivers as well. When Adaptive receivers believe that the sender can be trusted they are readily deceived by *Misleading* evidence. As Figure 6.13 reveals, the proportion of Adaptive receivers who correctly infer the truth is lowest in this case. By favouring the use of *Misleading* evidence when facing a trusting receiver, Adaptive senders appear to target Adaptive receivers. However, Adaptive receivers appear to benefit from their strategy by being able to draw stronger conclusions when their inferences about sender intent are correct. In contrast, the figure reveals that *Uninformative* evidence is most effective at concealing the truth from Conservative receivers, and that this strategy is the one favoured by Conservative senders.



(a) Adaptive Receivers

(b) Conservative Receivers

(c) Adaptive Senders

(d) Conservative Senders

Figure 6.13: **Upper panel:** Receiver choices for two different types of participants: (a) Adaptive and (b) Conservative. Adaptive receivers (N=58) were defined as those more likely to select the *Lure* when faced with *Misleading* evidence in the TEAMMATE condition. Adaptive people also drew stronger conclusions from the seemingly uninformative evidence, but only when the sender's cooperation was expected. The difference in their inferences between the OPPONENT and TEAMMATE condition suggests that they were sensitive to the sender intent when deciding what conclusions to draw. In contrast, Conservative receivers (N=40) responded in the same manner regardless of whether the evidence was *Misleading* or *Uninformative*, as well as irrespective of the sender's intent. **Lower panel:** Sender choices for (c) Adaptive and (d) Conservative participants. Conservative senders (N=44) were defined as those more likely to favour *Uninformative* evidence in the LOW SUSPICION condition. This preference is reversed in favour of *Misleading* evidence for Adaptive senders (N=32). But while Conservative senders prefer to be uninformative without regard to receivers' suspicions, Adaptive senders adapt their strategy accordingly, providing *Misleading* evidence when the receiver is likely to be low in suspicion but *Uninformative* evidence when the receiver suspects them already.

To summarise, we have grouped our participants on the basis of how they reason about the effect of *Misleading* evidence in the TEAMMATE condition. In doing so, we have isolated those participants (the Adaptive ones) whose responses have driven the context sensitive behaviour we observed and modelled in aggregate in the first part of this paper. In what follows, we revisit our computational model to determine whether a sampling assumptions account can explain the behaviour of these two distinct groups.

MODEL-BASED ANALYSES: INDIVIDUAL DIFFERENCES IN CONTENT SENSITIVITY

We now use the extended version of our model developed above in order to address two important questions that arose from the original analysis. Firstly, to what degree does the behaviour of our receiver participants indicate that they are adopting content-sensitive sampling assumptions? Specifically, do people appear to draw stronger conclusions (based on a stronger sampling assumption) when presented with *Misleading* evidence compared to *Uninformative* evidence? Secondly, if the sender assumes that the receiver makes a content-sensitive sampling assumption, how does this impact his choices? Can this type of reasoning account for the pattern of deceptive behaviour observed in our sender experiment?

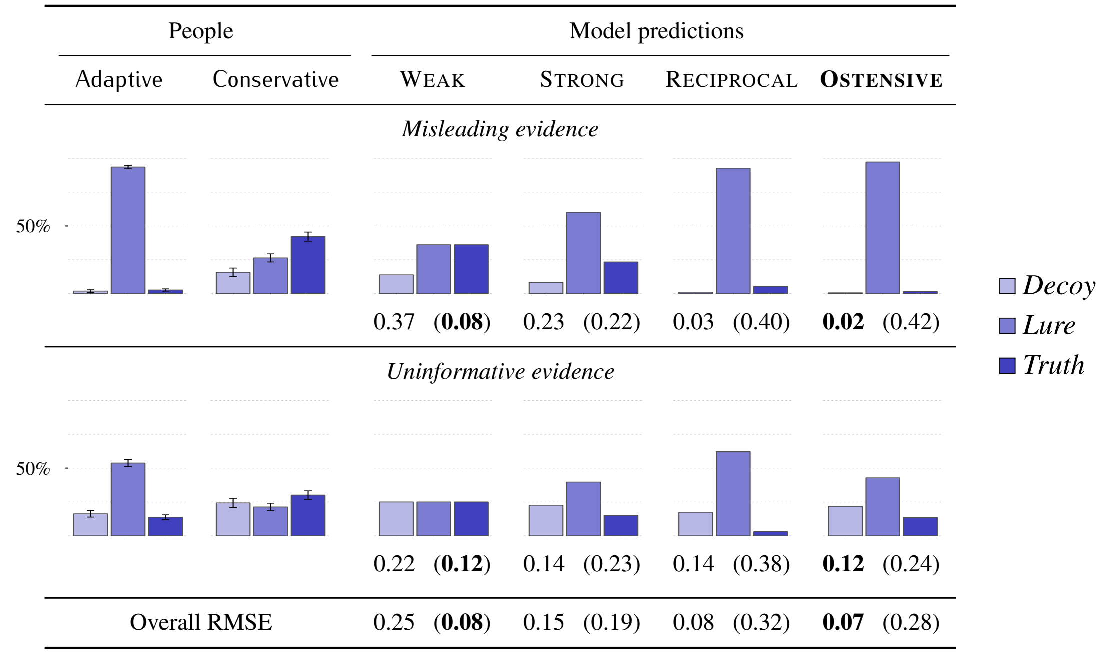
To address these questions, we compared model predictions of the OSTENSIVE model (as well as the four original models) to the choices of Adaptive and Conservative participants separately. Model predictions and associated fits are shown in figs. 6.14, 6.15 and 6.17.[11] Because our group-level analysis indicated that the pattern of context sensitive behaviour is driven primarily by Adaptive participants, we focus our discussion on those participants first, returning subsequently to consider what sampling assumptions best account for Conservative participants.

Adaptive participants

Figure 6.14 illustrates that for the Adaptive receivers, the OSTENSIVE model best captures their behaviour, suggesting that they are indeed drawing inferences about the way that the sender sampled the data based on how helpful the data appears to be. Because the *Misleading* evidence appears to be consistent with what might reasonably be chosen by a helpful sender, the model predicts that a trusting receiver will draw strong conclusions from it, in line with the predictions of the RECIPROCAL model. In contrast, the *Uninformative* evidence is inconsistent with helpful sampling and is therefore more likely to have been sampled at random. In this case, the OSTENSIVE model predicts weaker conclusions, more in line with the STRONG model. We find that this tendency to treat misleading and uninformative evidence in a qualitatively different way has the

[11] Predictions and fits were calculated for all models and conditions, but those not relevant to he present analyses have been dropped from the figures.

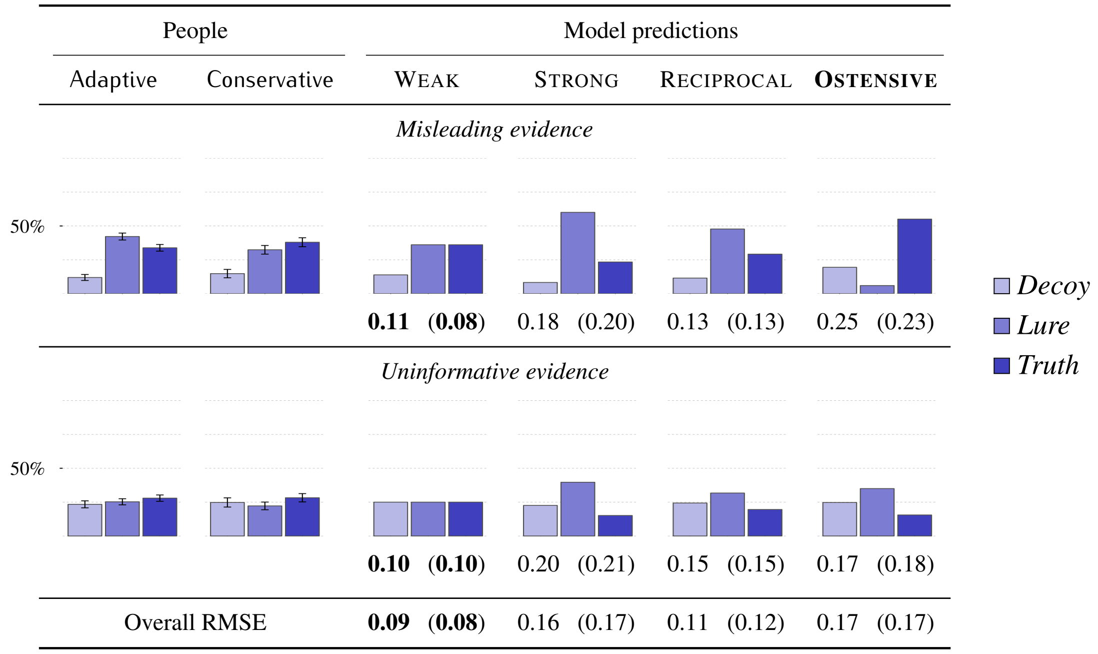
Model fits to choices of Adaptive and Conservative Receivers: TEAMMATE condition



	WEAK	STRONG	RECIPROCAL	OSTENSIVE
Misleading evidence	0.37 (**0.08**)	0.23 (0.22)	0.03 (0.40)	**0.02** (0.42)
Uninformative evidence	0.22 (**0.12**)	0.14 (0.23)	0.14 (0.38)	**0.12** (0.24)
Overall RMSE	0.25 (**0.08**)	0.15 (0.19)	0.08 (0.32)	**0.07** (0.28)

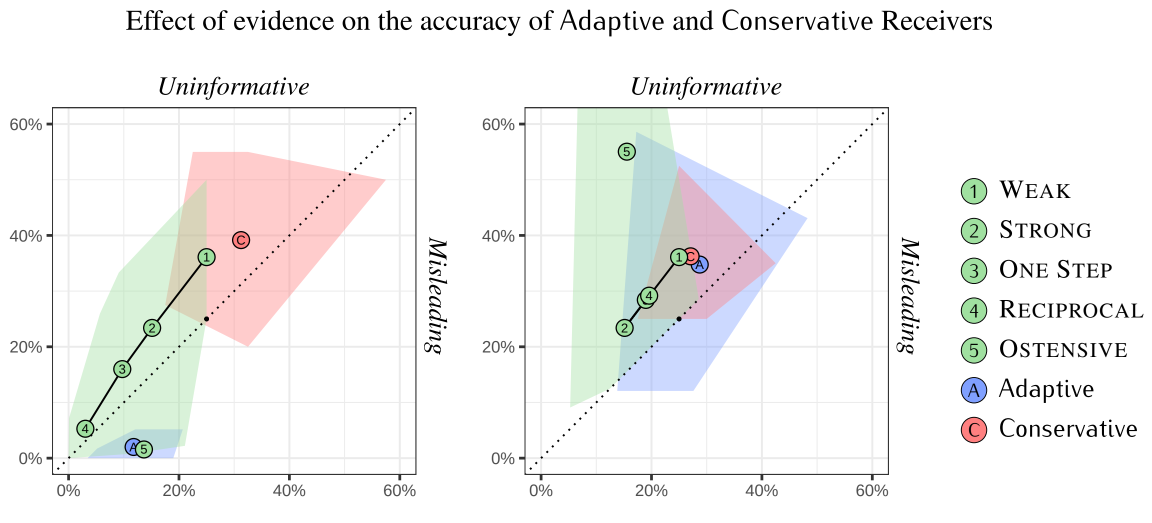
Figure 6.14: Predictions of the OSTENSIVE model compared with three content insensitive models of meta-inference and the choices of Adaptive and Conservative participants in the TEAMMATE condition. In the WEAK, STRONG, and RECIPROCAL models (described earlier), the way that the assumptions are arrived at in the first place, is left undefined. The OSTENSIVE model in contrast, describes the computational problem faced by the receiver as one of joint inference over sampling strategy and the hypotheses in question. Under this form of joint inference, certain scenarios are considered more likely than others: helpful (but misleading) content is more likely to have been helpfully selected, while uninformative content is more likely to have been selected randomly or without care. The closely matching predictions made by the OSTENSIVE and RECIPROCAL models (for *Misleading* content) and by the OSTENSIVE and STRONG models (for *Uninformative* content), follow as a consequence of the *content-sensitive* nature of the OSTENSIVE model. The numbers below each graph show the model fits for Adaptive and (Conservative) participants, as measured by RMSE. Once again, lower RMSE values represent better model fits. Adaptive receivers are best fit by the OSTENSIVE model, appearing to rely on a stronger assumption when given misleading evidence than when faced with something less informative. Conservative receivers in contrast, gain little leverage from their sampling assumptions irrespective of the content, and are best fit by the WEAK model.

Model fits to choices of Adaptive and Conservative Receivers: OPPONENT condition



People		Model predictions			
Adaptive	Conservative	WEAK	STRONG	RECIPROCAL	**OSTENSIVE**
Misleading evidence					
		0.11 (**0.08**)	0.18 (0.20)	0.13 (0.13)	0.25 (0.23)
Uninformative evidence					
		0.10 (**0.10**)	0.20 (0.21)	0.15 (0.15)	0.17 (0.18)
Overall RMSE		**0.09** (**0.08**)	0.16 (0.17)	0.11 (0.12)	0.17 (0.17)

Figure 6.15: Predictions of the OSTENSIVE model compared with three content insensitive models of meta-inference and the choices of Adaptive and Conservative participants in the OPPONENT condition. The OSTENSIVE model is based on the intuition that message content may be informative both in the usual way and with regard to how it was sampled. Content that is informative and easily recognisable as such, the intuition goes, can be particularly misleading. The predictions of the model regarding *Misleading* evidence, show that a receiver alert to this form of deception, rather than be misled, could effectively leverage her suspicion to get closer to the truth. Yet, the plots clearly indicate that this is not what people did. Similarly, neither the STRONG nor the RECIPROCAL model represent a close match for either group of receivers, since both models embody a modest amount of meta-inferential leverage (due to the size principle), despite the receiver's suspicions. Only the WEAK model, which effectively discounts all evidence of a meta-inferential nature, provides a reasonable account of either group of participants. Model fits (RMSE) for Adaptive and (Conservative) participants are shown beneath each plot.



(a) TEAMMATE condition

(b) OPPONENT condition

Figure 6.16: Accuracy of Adaptive and Conservative receivers based on the type of evidence provided. The plotted points represent model predictions (green circles) and people's performance (blue circles) aggregated across the six sets of stimuli, while the polygons illustrate the spread of predictions – each vertex corrsesponds to a single set. Model predictions in the TEAMMATE condition highlight the qualitatively different predictions of the OSTENSIVE model which assumes that the receiver forms their sampling assumption based in part on the information content itself. Consequently, only the OSTENSIVE model predicts that *Misleading* evidence will have a greater negative impact on receiver accuracy than *Uninformative* evidece, and is able to capture the accuracy of Adaptive receivers as a result. In the OPPONENT condition in contrast, the OSTENSIVE model predicts a backfire effect whereby the *Misleading* evidence improves rather than impairs the receiver's accuracy. Notably, this backfire effect did not occur. When faced with a potentially deceptive sender, both Adaptive and Conservative receivers favoured a literal interpretation of the evidence (in keeping with the "no lying" rule), as predicted by the WEAK model.

expected consequence: Adaptive receivers in the TEAMMATE condition were less likely to uncover the truth when given *Misleading* evidence (see Figure 6.16(a)).

It is important to note that the predictions of the OSTENSIVE model were not reflected by our participants in the OPPONENT condition. Under the OSTENSIVE model, a suspicious receiver can discount the ostensive implication of the *Misleading* evidence, that way ruling out the *Lure* hypothesis and improving their chances of uncovering the truth (see the OSTENSIVE model prediction in Figure 6.16(b)). Instead, for all participants, it seems more likely that they adopted a weak sampling assumption across the board. Nonetheless, the qualitative reversal predicted still offers a possible explanation of sender behaviour. If the sender does assume, as the OSTENSIVE model predicts, that misleading pays off when the receiver is trusting and backfires when she is suspicious, then a qualitative reversal of deceptive strategy as a function of receiver suspicion is justified. Indeed, as Figure 6.17 shows, the OSTENSIVE model best fits the behaviour of Adaptive

Model fits to choices of Adaptive and Conservative Senders

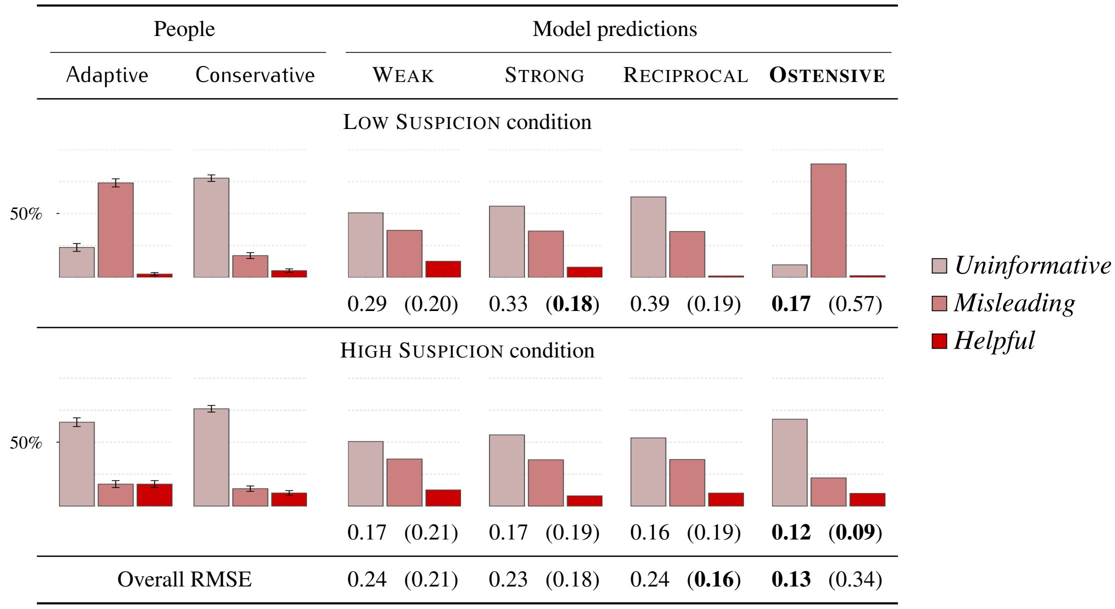


Figure 6.17: Predictions of the OSTENSIVE model compared with three content insensitive models of meta-inference and the choices of Adaptive and Conservative participants playing the role of the Sender. The OSTENSIVE model is based on the idea that data, in addition to being informative about some matter at hand, may also be informative regarding how it was selected in the first place. The remaining models describe meta-inference that is not sensitive to content in this way. Consequently, only the OSTENSIVE model is able to capture the strong preference for *Misleading* content exhibited by Adaptive people in the LOW SUSPICION condition. Furthermore, because the model predicts that attempts to actively mislead will backfire when the receiver is suspicious, it is the only model that captures the qualitative reversal of preference that Adaptive people show between conditions. Model fits (RMSE) are shown beneath each plot, and averaged across conditions in the bottom row of the table. Lower numbers represent better fits. The fits reveal that only the OSTENSIVE model, which heavily penalizes *Misleading* evidence when the receiver is suspicious, comes close to capturing the strength of people's preference for being uninformative in the HIGH SUSPICION condition. The fits also reveal that the behaviour of Conservative people in the LOW SUSPICION condition is not well explained by the models.

senders; it is the only one that predicts a significant change in sender behaviour between the LOW SUSPICION and HIGH SUSPICION conditions. It therefore captures the strong preference to mislead a trusting receiver as well as the equally strong preference to be uninformative when the receiver is likely to be suspicious.

Conservative participants

Our Conservative receivers showed little difference in their behaviour between the TEAMMATE and OPPONENT conditions, preferring to avoid strong conclusions in both situations. Accordingly, we find that the WEAK model captures recipient behaviour consistently for both conditions and for both *Misleading* and *Uninformative* evidence (see figs. 6.14 and 6.15).

What about the senders? Under a weak sampling assumption, the receiver uses evidence solely to disconfirm incompatible hypotheses. It logically follows then that the less information the sender reveals the less chance the receiver has of inferring the truth. But, as Figure 6.16 shows, if the receiver adopts a weak sampling assumption, then the advantage (from the sender perspective) of offering *Uninformative* evidence over *Misleading* evidence is small. If senders choose their strategy according to this small relative difference, then we might expect to see senders exhibit a correspondingly small relative preference for *Uninformative* evidence. Yet, as the poor fits of the WEAK model for the senders indicate (see Figure 6.17), this is not how senders behave. Rather, regardless of condition, Conservative senders show the same strong preference to be uninformative. This inclination to avoid the misleading option suggests that Conservative senders, like Adaptive senders anticipating suspicious receivers, believe that attempts to mislead the receiver will backfire. The predictions of the OSTENSIVE model in the HIGH SUSPICION condition which capture this "backfire" concept provide the best (and only reasonable) fit to the behaviour of conservative participants. Because Conservative senders responded similarly in the LOW SUSPICION and HIGH SUSPICION conditions, we also assessed the degree to which the *Ostensive* assumption from the HIGH SUSPICION condition captured behaviour in the LOW SUSPICION condition. This yielded a better fit than the next best fitting assumption (*Strong*: 0.18).

DISCUSSION

The goal of the present study was to examine how people reasoned about evidence in situations where deception was possible but lying was not an option. People played the deception game in two related experiments: both as "receivers" and "senders" of messages (evidence). When viewed as an homogenous sample, people as receivers were sensitive to the context in which communication took place. They drew strong conclusions from evidence, but only when they thought the sender could be trusted. But evidence for context sensitive behaviour amongst senders was weaker overall. Though people were more willing to mislead when they thought deception was not expected, they favoured uninformative evidence regardless of context. Using a computational framework to predict responses on the basis of meta-inferential sampling assumptions, our original analysis found that while the behaviour of receivers as a whole was consistent with a sampling assumptions account, the behaviour of senders in the aggregate was not so easily accounted for by standard sampling assumption discussed in the literature.

Our more detailed analysis was thus prompted by two issues. Firstly, the surprising result that sender behaviour appeared to be somewhat insensitive to context. And secondly, the fact that our original model which builds on standard sampling assumptions cannot account for even the modest change in sender strategy that was observed. With respect to the first issue, subsequent analyses of participant's choices revealed a plausible explanation. For both senders and receivers there appear to be two qualitatively distinct groups of people: Adaptive participants who tailored their behaviour according to context, and Conservative participants who maintained a single consistent approach. Adaptive receivers reasoned well beyond a literal interpretation of the evidence, but only when the sender's cooperation was implied. Adaptive senders demonstrated a clear reversal in preference for misleading evidence between conditions.

To address the limitations of our original model, we explored the predictions of the OSTENSIVE model. Table 6.3 summarises the findings of our original and revised analyses. The analysis indicates that Adaptive participants, acted consistent with an ostensive-inferential view of communication. When the context dictated that cooperative norms should apply, full cooperation was not taken for granted. Instead, Adaptive receivers leveraged ostensive signs of helpfulness. Adaptive senders overwhelmingly preferred to provide ostensively helpful yet misleading evidence when seeking to mislead.

In contrast, when the context suggests that cooperative norms may not apply, the analysis indicates a disconnect between sender and receiver assumptions. Anticipating that ostensive signals can backfire when the receiver is suspicious, Adaptive senders declined to offer them. Yet absent trust, Adaptive receivers ignored ostensive signals and took a literal view of data. Despite the disconnect, both sets of assumptions represent sensible defensive positions. For receivers, a weak sampling assumption means that they are immune to strategic exploitation. For senders, the extra caution is largely without cost (unless there are a sufficiently large population of receivers who were trusting despite obvious cause for suspicion). The concept of a "defensive" assumption may explain the behaviour of Conservative participants. Our analysis revealed that the best fitting assumption for both senders and receivers in this group matched the assumption of Adaptive participants in an adversarial context. We comment further on the Conservative stance, and other broader issues in the following general discussion.

Despite the limitations of our group-level analyses, due to its post hoc nature and the limited amount of data collected per individual, we find that the revised analysis gives a more compelling account of behaviour overall, and for senders in particular. Importantly, by isolating the group of participants responsible for driving context-sensitive behaviour, and introducing the OSTENSIVE model to capture content-sensitive sampling assumptions, we have better captured the way that qualitative patterns of responding were driven by both content and context.

6.6 GENERAL DISCUSSION

Using a non-verbal communication game where deception was motivated but outright lying was not an option, we investigated how the spectre of deception changes the way

	All People		Adaptive People		Conservative People	
Condition	Assumption	Fit	Assumption	Fit	Assumption	Fit
(Receiver)						
Teammate	*Strong*	0.09	*Ostensive*	0.07	*Weak*	0.08
Opponent	*Weak*	0.07	*Weak*	0.09	*Weak*	0.08
(Sender)						
Low Suspicion	*Weak*	0.07	*Hinder* (*Ostensive*)	0.17	*Hinder* (*Hinder* (*Ostensive*))	0.09
High Suspicion	*Strong*	0.17	*Hinder* (*Hinder* (*Ostensive*))	0.12	*Hinder* (*Hinder* (*Ostensive*))	0.09

Table 6.3: **Best fitting models of sender and receiver behaviour in the deception game**. See tables 6.1 and 6.2 and main text for model descriptions. For model fits (RMSE), lower numbers represent better fits. Overall the fits indicated by our *post hoc* group-level analysis suggest a more nuanced picture than the aggregate analysis revealed. In particular, the revised analyses suggests a role for the kind of content-sensitive sampling assumption captured by the Ostensive model. Further it suggests that Conservative senders and receivers may have adopted a form of "worst case" assumption whether or not suspicion was warranted.

that people reason about evidence. Across two experiments, in both production and comprehension tasks, we found that people's behaviour was guided by their inferences about how others reason from evidence. When selecting evidence to provide, people reasoned about the way that suspicion affects the comprehension process. And when interpreting evidence provided, people considered the ways that evidence may be used deceptively. Support for this conclusion in its most basic sense can be seen in the context sensitive pattern of responses we observed in both experiments.

On the comprehension side, people were sensitive to the (presumed) intent of the sender. They drew strong conclusions from evidence when the context dictated that it made sense to trust the sender, but reached guarded conclusions otherwise. This behaviour is consistent with the findings of comparable studies investigating pragmatic implicature in non-cooperative contexts. For example, using a picture selection task Pryslopska (2013) found that the pragmatic interpretation of "some" as meaning "some but not all" was more likely when the context emphasised cooperation over competition. In a similar vein, Dulcinatti (2018) demonstrated (via a picture selection task, as well as a task involving purely verbal reasoning) that a range of scalar and ad hoc implicatures drove people's conclusions in cooperative but not competitive scenarios.

On the production side, people's selective presentation of evidence was also sensitive to their communicative goal, even without recourse to outright lying. While comparable studies are somewhat rare, people's ability to selectively employ seemingly helpful yet ultimately mileading information has been demonstrated in verbal reasoning tasks in adults (Dulcinatti, 2018, ch. 6) and in concept teaching tasks in children as young as 4 years old (Rhodes, Bonawitz, Shafto, Chen, & Caglar, 2015). Our experiment extends these results by using the same experimental task to examine how people adjust their strategy in the face of low and high suspicion. When people believed their intentions would not be viewed with suspicion many preferred to make misleading implications, but when their motive to deceive was made plain in context, people strongly preferred to give nothing away.

Context-sensitive and content-sensitive sampling assumptions

Our empirical results established the basis for our computational analyses of people's assumptions which reveal useful insights into how people think that others reason from evidence. Our modelling suggests that people (both as senders and receivers) reasoned probabilistically about the generative process underlying communication within the game, and interpreted evidence flexibly in light of those assumptions. The changing nature and strength of people's assumptions brought about by the cooperative or competitive context drove the corresponding change in responding observed. Our findings replicate and extend core findings in the sampling assumptions literature. From our analysis of receiver behaviour, we find evidence of the size principle in operation. Following this principle people tended to generalise from evidence to the smallest compatible hypothesis even when that evidence was otherwise uninformative. This principle, and assumptions which build upon it have been shown to shape inductive reasoning in a variety of tasks including learning abstract concepts (Tenenbaum, 2000), word learning (Xu & Tenenbaum, 2007a), category learning (Hendrickson et al., 2019), property induction (Fernbach, 2006; Sanjana & Tenenbaum, 2003), and similarity judgments (Navarro & Perfors, 2010; Tenenbaum & Griffiths, 2001a). Receiver behaviour in the Opponent condition was well captured by a weak sampling assumption. Frequently, this assumption has been used in the literature to model aleatory uncertainty in the generative process – when observations are sampled at random and independently of the concept of interest, for example (Heit, 1998; Kemp & Tenenbaum, 2009; Shepard, 1987). Our results highlight that it applies in situations of epistemic uncertainty also, even when observations are clearly restricted to the true concept.

Despite the fact that trusting receivers reasoned beyond the evidence even when it was seemingly uninformative, our analysis suggests that people may have adjusted their sampling assumptions based on the data observed. Although our evidence in support of this is modest at best, it is nonetheless consistent with previous findings that people's sampling assumptions may be shaped by the data (Hendrickson et al., 2019; Ransom et al., 2016) and that people perform joint inference over the knowledge and intent of their informant and the truth of the matter at hand (e.g., N. D. Goodman & Frank, 2016; Gweon et al., 2010; Shafto, Eaves, et al., 2012). The data from the Opponent condition is less ambiguous. There are no signs that suspicious receivers were drawn into content-based second guessing of strategy whether the given evidence appeared purposefully or haphazardly sampled. Any joint inference (and the function of epistemic vigilance that it supports) were effectively suspended given the high prior probability that the sender was uncooperative. Our computational model cannot speak directly to the question of whether joint inference regarding sender intent is *actually suspended* in this case, or whether such inferences are drawn and over-ruled. Nonetheless, such questions are an interesting avenue for future investigation. This issue potentially connects with a debate in the pragmatics literature regarding whether implicatures are drawn as the context demands (e.g., Russell, 2006) or are always computed by default but sometimes discarded (e.g., Levinson, 2000).

Our analysis of the problem facing the would-be deceptive sender reveals that what may seem like an obvious heuristic – mislead the trusting, conceal from the suspicious – is not so readily justified. Setting aside the fact that the apparently obvious intuition was shared by only half our participants, our simulations revealed two important disconnects with standard (content-insensitive) sampling assumptions like strong and weak sampling. The first of these is when the receiver is not suspicious. Like the rising tide that lifts all boats, a content-insensitive sampling assumption based on the size principle supports reasoning beyond the data no matter what the data. Under such assumptions, the mathematics of Bayesian inference suggests that however misleading a given piece of evidence may appear, a subset of that same evidence is always a better option.[12] When the receiver instead believes that the sender's goal is opposed to her own, her best bet is to ignore those aspects of the signal that the sender controls (e.g., Hespanha et al., 2000). In the deception game, this means adopting a weak (uninformative) sampling assumption – a tactic overwhelmingly followed by our receiver participants. Thus the second disconnect is that senders did not appear to behave in line with the assumption that the receiver would ignore all unreliable aspects of the evidence. Instead, senders showed a bias against misleading evidence to an extent not justified by the information penalty alone.[13]

Our analysis offers a plausible (albeit speculative) explanation for the senders' bias: that is, senders assumed that receivers would reason further beyond some data than others. Whether people arrived at this assumption through some form of mental simulation or via an intuitive theory derived from experience, the fact that people assumed that receivers would act in this way complements the modest evidence from our receiver experiment

[12] This is certainly the case in the deception game where only positive (and truthful) evidence is allowed, and applies regardless of the strength of the informativity bias or the number of recursive layers of "he thinks, she thinks reasoning". Additionally initial simulations show that this may be a robust result that applies to any discrete likelihood function obeying reasonable constraints related to the size principle: namely, any given observation should be more (or equally) likely under the smaller of any two hypotheses with which it is compatible; and for any two observations compatible with the same hypothesis, the one that is compatible with fewer alternatives should be more (or equally) likely. A formal proof and further investigation of the generality of this property are an area for future work.

[13] Under the (somewhat standard) assumption that $\alpha = 1$, as used in our model. One might alternatively account for the bias observed in the HIGH SUSPICION condition by using a higher value to reflect more optimal choosing on the part of the sender. But doing so consistently would also predict more extreme values in the LOW SUSPICION condition which were not observed.

that this is the case. Alternatively, people might simply be mistaken – in itself this would represent an intersting disconnect between production and comprehension that would be worth pursuing. Regardless, to the best of our knowledge, our finding that such an assumption is operating on the production side is a novel one, and one with interesting implications if it can be replicated.

Ostensive meta-inference

If people do have an (implicit) awareness that comprehension may be affected by content-sensitive sampling assumptions, it is interesting to consider whether and how this effects communication on the production side. For instance, do senders attempt to increase the chances that a particular sampling assumption will be adopted by their counterpart by signalling it in some way? The use of ostensive signals such as eye-gaze, pointing and tone modulation have been shown to play an important role in infant learning. Such signals help the infant to understand that they are being addressed, to make clear the referent when teaching object labels, and even to indicate that that the information being conveyed is of a generalizable nature (Csibra & Gergely, 2009; Topál, Gergely, Miklósi, Erdőhegyi, & Csibra, 2008).

More broadly, a central idea of Relevance Theory (Wilson & Sperber, 2004) is that of *ostensive-inferential* communication, the purpose of which is not only to inform one's interlocutor, but also to inform them of your intention to inform them. The idea of what we might call *ostensive meta-inferential* communication is closely related. A simple example can be found in everyday discourse. Replying "It's after 5." when a colleague asks you the time suggests not that the time is "5:01" as it might under a strongly informative assumption, but more likely that it is some time after 5 o'clock (and presumably before 6 o'clock). The use of the modifier "after" may signal that the recipient should not generalise too narrowly from the data. Using our computational model we analysed one particular form that ostensive meta-inference might take. The Ostensive model captured the notion that although two stimuli might license the same inference in a particular context, the more ostensive one would licence stronger inference in a broader range of contexts. The results of our sender experiment suggests that senders considered such implications when weighing up their options. This was evident in their avoidance of the *Misleading* option in the High Suspicion condition, even when it was technically no more informative than the *Uninformative* option (see Figure 6.11 for an example of such).

In experiments investigating the generation of referential expressions, the production of contextually redundant information (so-called *over-specification*) has been frequently observed, while under-specification is comparatively rare (Pogue, Kurumada, & Tanenhaus, 2016). And while under-specification is consistently rated as unhelpful by receivers, over-specification is not viewed in this way (Engelhardt, Bailey, & Ferreira, 2006). Indeed, by making communication more robust, over-specification can facilitate faster object identification (Arts, Maes, Noordman, & Jansen, 2011). In our experiment, misleading but uninformative stimuli can be considered "over-specified", at least in relation to the purely uninformative stimuli. Thus, these findings lend support to the idea that people in our experiment would consider the ostensive properties of stimuli when reasoning about evidence. If over-specification is common and helpful, then for some senders it will make sense to favour it when the receiver has no reason to be suspicious and to avoid doing so otherwise (for fear of the strategy back-firing).

There is some evidence to suggest that a complementary tactic of ostensive under-specification may too play a role in deceptive communication. In a study of non-verbal deception with parallels to our own, Montague et al. (2011) used a "rectangle game" to investigate the use of deceptive strategies and their impact on learners. Participants played the part of informants who indicated points within or outside of a rectangle, or learners who had to infer the true boundary from the evidence provided. The cover story and instructions provided to learners left the helpfulness of informant testimony in question. Although informants were allowed to lie outright, it was not the preferred strategy in the competitive condition, presumably because learners were allowed to verify information. Instead, informants in that condition favoured points which were relatively uninformative (and had no significant correlation with learner error - a measure of deceptive success in this case). Because informants in the cooperative condition were required to provide more points than the two strictly required to mark the opposite corners of a rectangle, they too provided uninformative points (which also had no significant correlation with learner error). Nonetheless, informants displayed context sensitivity in the choice of uninformative points. Uninformative evidence provided by cooperative informants was mostly positive (within the rectangle), while competitive informants favoured negative evidence (exterior points). In information theoretic terms, whether negative evidence is more or less informative than positive evidence depends upon the structure of the hypothesis space and the size of the hypothesis in question (Navarro & Perfors, 2011). But given the lack of correlation in Montague et al.'s data between learner error and uninformative evidence of either kind, the qualitative reversal of strategy observed between cooperative and competitive informants is intriguing.

A plausible connection with ostensive signalling arises as a consequence of the frequently sparse nature of the hypotheses with which learners are concerned (Navarro & Perfors, 2011). In an environment where hypotheses are sparse the expected information value of negative evidence (in advance of actually determining it) is less than that of positive evidence. Deceptive informants sensitive to the average uninformativeness of negative evidence (rather than its context-specific value) may thus prefer it over positive (yet uninformative) evidence without any further inference required. There is evidence to suggest that this ostensive use of negative evidence may impact people's sampling assumptions. For example, it has been noted that negative evidence or evidence from a second concept can induce a weaker sampling assumption on the part of the learner (Hendrickson et al., 2019; Ransom et al., 2016).

Taken together, our own results and those of Montague et al. (2011) support the idea that deceptive informants are sensitive to the ostensive qualities of data as well as its context-specific information content. An interesting avenue for future research would be to investigate whether any such sensitivity is heightened in deceptive contexts or representative of communication more broadly. An awareness of such differential sensitivity has the potential to benefit verbal deception detection techniques such as *forced choice tests* (Frederick & Speed, 2007) and *model statements* (Vrij, Leal, & Fisher, 2018).

Individual differences in meta-inferential stance

Responses across both experiments were subject to important qualitative differences amongst individuals. On the comprehension side, only Adaptive receivers were sensitive to a difference in the evidentiary value of data in cooperative and competitive contexts. Likewise on the production side, only Adaptive senders were sensitive to suspicion in forming meta-inferential assumptions. Similar patterns of individual differences have been noted elsewhere in the literature. In a related study, Franke and Degen (2016) used a similar Bayesian modelling framework to analyse production and comprehension behaviour in a (cooperative) reference game. They found that while listener behaviour appeared consistent with *Gricean* reasoning (analogous to our One Step model) in the aggregate, closer analysis revealed that the majority of listeners used so-called *exhaustive* reasoning (analogous to our Strong model), with the average being skewed by a smaller number of highly pragmatic participants. On the back of their analysis Franke and Degen (2016) highlight the importance of considering individual differences in computational level analysis, lest averaging effects obscure the different computational strategies being employed. Based on our own analysis we echo these sentiments.

Given our analyses, how should we interpret the differences in assumptions between Adaptive and Conservative participants? One obvious answer relates these differences to differences in the depth of reasoning in which people engaged. Such differences have been observed in experimental studies employing strategic reasoning games (e.g., Hedden & Zhang, 2002; Ohtsubo & Rapoport, 2006; Stahl & Wilson, 1995). Stahl and Wilson (1995) for example, analysed people's responses across twelve 3 × 3 symmetric games. Comparing various models of player behaviour, they found that most people could be grouped into one of four major categories: *level 0* types who choose randomly, *level 1* types who reasoned as if their opponent was a level 0 type, naive Nash types who used an equilibrium strategy (analogous to our Reciprocal model), and *worldly* types (the largest group) who reasoned that their opponent might be any one of the preceding types. Stahl and Wilson's finding that a significant proportion of people were sensitive to individual differences in reasoning styles connects with our own finding regarding Adaptive participants. If people expect a resonable amount of variation between (or within) individuals then the cognitive effort required to infer content-sensitive sampling assumptions may be justified. And given a sufficient population of Adaptive receivers, sensitivity to the meta-inferential implications of content makes sense for senders motivated to deceive.

But what about our Conservative participants – what might explain their behaviour? A simple explanation is that Conservative receivers failed to engage in meta-inferential reasoning at all. But given that the deception game explicitly entails the use of positive evidence only, an assumption that evidence was selected at random should justify a strong sampling assumption, not the weak assumption that Conservative receivers adopted. This does not rule out the *no meta-inference* explanation of course. The tendency of experimental participants to underweight the value of evidence has long been noted (e.g., Edwards, 1968; Phillips & Edwards, 1966), and a variety of explanations have been offered (for a review, see Corner, Harris, & Hahn, 2010). Navarro et al. (2012), found evidence that people adopt conservative sampling assumptions across a range of simple generalisation tasks. By modelling the strength of assumptions drawn (as a linear combination of strong and weak sampling), they found considerable variation amongst individuals. However, the "no meta-inference" explanation cannot account for Conservative senders – such behaviour among senders would have meant choosing information at random, for which there was no evidence.

An alternative explanation for the behaviour of some Conservative receivers at least is not that they didn't (or couldn't) engage in the kind of meta-inference required, rather that they drew strong inferences but rejected them in favour of a more literal/logical interpretation. Feeney, Scrafton, Duckworth, and Handley (2004) found evidence of comparable pragmatic inhibition. Their study looked at how people respond to uses of "some" that are felicitous (e.g. *some cars are red*) or infelicitous (e.g. *some birds have wings*). Reaction time data indicated that people took longer to endorse the literal meaning of infelicitous examples, suggesting extra cognitive effort was required to reject a misleading implication (for example, that *some but not all birds have wings*). The idea that some receivers draw but reject misleading inferences would help to explain the presence of conservative senders who avoid making such implications in the first place. However, given that our experiment was not designed to distinguish between the "no meta-inference" and "rejected meta-inference" explanations, this remains an area for future investigation.

6.7 Conclusion

We presented a computational framework for modelling the production and comprehension of information in a combined experimental and computational study of deception without lying. Our work makes two main contributions. First, we have provided an empirical demonstration that by formalising the production of messages as the computational inverse of comprehension it is possible to capture the behaviour of people seeking to mislead or conceal information from suspicious or naive targets. On the flip side, we have shown that by casting people's beliefs about the contingent nature of message production as probabilistic sampling assumptions, the same model can capture people's inferences when they are knowingly or unknowingly the target of deception. Reflecting on the findings of decades of deception research, Levine and McCornack (2014) argue that the principle drivers of deceptive behaviour are rational and utilitarian. People deceive when they need to, making the best of the information they possess given the contextual constraints. Further, they argue that the practical concerns of deception detection would be better served by an understanding of message content and the context in which it is produced, than by the myriad non-verbal cues which have proved relatively ineffective (see for example, Bond & DePaulo, 2006). By showing that the framework can capture a diversity of behaviour — that is, production and comprehension tasks in both cooperative and non-cooperative scenarios and across contexts where suspicion does and does not naturally arise — we hope to have demonstrated its applicability for further deception research.

Importantly, by using the framework to examine the predictions that particular models *cannot* make, we have been able to test alternative hypotheses concerning the ways that content and context combine to drive inference beyond the data provided. Our second contribution is thus an empirical demonstration and analysis of the context- and content-sensitive nature of meta-inference.

The process of reasoning about the inferences of another, has been studied in a variety of settings, including concept learning and teaching (Shafto et al., 2014), learning from goal directed actions (Baker, Saxe, & Tenenbaum, 2009; Shafto, Goodman, & Frank, 2012; Ullman et al., 2009), intentional selection (Durkin, Caglar, Bonawitz, & Shafto, 2015; Shafto & Bonawitz, 2015), preference learning (Jern, Lucas, & Kemp, 2017), attitude attribution (Hawthorne-Madell & Goodman, 2015; Walker, Smith, & Vul, 2015), and pragmatic language understanding (M. C. Frank & Goodman, 2012; Franke & Degen, 2016; N. D. Goodman & Stuhlmüller, 2013; Harris, Corner, & Hahn, 2013; Hawkins, Stuhlmüller, Degen, & Goodman, 2015). These studies share a common view that people make probabilistic assumptions about the way that others reason and act, and that they take this into account when drawing conclusions and communicating. Our work adds to this growing body of literature demonstrating that people enjoy the benefits of such meta-inference, learning more from less when interlocutors cooperate, while guarding against those seeking to exploit such tendencies in order to mislead.

6.8 Acknowledgments

KR was supported by an Australian Government Research Training Program Scholarship. AP received salary support from ARC grants DP110104949 and DP150103280 and DJN from ARC grant FT110100431. Preliminary versions of this work were published in the proceedings of the 39th Annual Meeting of the Cognitive Science Society.

7 WHAT LIES BEYOND THE DATA?

My aim in this thesis has been to examine the assumptions we make about what lies beneath the data and how we use these assumptions to reason beyond it. I have investigated the effect of sampling assumptions in a diversity of settings: when the relevant basis for generalisation is clear, and when it is not; and when the generative process has been made clear and when it has not. A central tennet of this work has been the idea that data alone is prospective evidence, not evidence itself. I have considered three broad classes of prospective evidence: perceptual, conceptual and theoretical – and I have shown how sampling assumptions interact with each to alter the interpretation and the effect on inference. The work has demonstrated a wide range of effects that people's sampling assumptions may bring about. From a subtle shift in the boundary between two categories (Chapter 3), to changing the mental representation of a concept to be learned (Chapter 2). From altering the conclusions people draw from evidence in the data (Chapter 5) to altering the way they produce data as evidence (Chapter 6). Thus the studies I have described replicate and extend the core finding in a growing body of literature: namely, that our generalisations are shaped not only on the basis of the observations we sample from the world, but also on our assumptions about the processes and constraints that define the sample in the first place. Put simply, sampling assumptions can make a difference.

In this final chapter I recap the key findings of my research, and take a more critical look at the implications of the work as a whole focusing on two important aspects. Firstly, I consider what the findings say about the issue of sampling sensitivity more generally., which is the source of some debate in the literature. Secondly, I speculate about the methodological and theoretical implications of the kinds of individual differences observed across the studies. Along the way, I highlight various topics for future research.

7.1 TL;DR: The story so far...

Number of categories, category base rates, sample size and sampling assumptions interact to affect generalisation boundaries.

The original aim of the studies described in Chapters 2 and 3 was to investigate how sampling assumptions affect the breadth of generalisation on the basis of low-dimensional perceptual stimuli. By using a common experimental framework across a one-category and two-category learning task, it was possible to study how sampling assumptions interact with the learning context. The two studies manipulated the number of categories being learned, the amount of the data provided to learners, the category base rates (where two categories were involved), and the explanation given for how those examples were selected. In both experiments, people's generalisation behaviour was sensitive to the sampling explanation offered, but in a way that interacted with the other contextual variables. When learning a single category from examples purportedly sampled to facilitate learning, people reliably tightened their generalisation in response to additional (non-diverse) examples. When instead a form of censoring was apparent, the additional examples had no effect on generalisation. When learning two categories, people made use of the disparity in the relative base rates of each category to inform their generalisation decision. But they did so only when they believed that examples were sampled at random from a pool of objects rather than being sampled specifically to facilitate learning.

Sampling assumptions can interact with mental representations of a to-be-learned category to drive category generalisation.

In both the one-category and two-category experiments (Chapters 2 and 3, respectively), small samples (four examples) and large samples (twelve examples) spanned the same range on the relevant generalisation dimension (which was made explicit to participants). The original motivation for controlling sample diversity in this way was to support the inference that any change in generalisation observed in relation to additional exemples was attributable to the sampling assumption that people had adopted. However, the very notion of controlling for diversity (or equivalently, restricting examples to the region of interpolation) pre-supposes a particular representation in psychological space. Closer inspection of the results of the first one-category experiment (Chapter 2, Experiment 1) revealed that this assumption was not supported by the data. Thus the study described in Chapter 2 aimed at investigating how mental representations of the inductive problem at hand (in this case the category to be learned) and people's sampling assumptions can interact. The study revealed two important findings. Firstly, despite the relative simplicity of the low-dimensional perceptual stimuli used, there were significant individual differences in people's mental representation of the category to be learned. Secondly, additional examples led many people to adopt a different representation, but this change in representation was influenced by the sampling explanation people had been given.

Sampling assumptions may take effect during learning when stimuli are encoded in memory.

While the extant literature (as well as my own research) has focused on demonstrating important effects of sampling assumptions and the computational underpinnings, questions concerning the nature of evidence and the representation of such assumptions have received little attention. The motivation behind the work in Chapter 4 was to begin to get at one such question: do sampling assumptions affect learning or reasoning? The new experiment was an extension of the one-category learning experiment described in Chapter 2 (Experiment 2). The design involved systematically varying both the sampling cover story that people were provided, and whether it was given before or after the training stimuli were presented. In this way it was possible to examine whether sampling assumptions took effect when stimuli were first encoded during training or later, when memories were retrieved during testing. Although the study should be taken in the spirit of a proof of concept, it nonetheless revealed two interesting findings. The results suggested that people's sampling assumptions impacted category learning and not simply generalisation performance. The sampling explanation had an effect only when it was made explicit, prior to learning. When it was presented instead after learning (prior to testing), aggregate generalisation behaviour was remarkably similar irrespective of cover story manipulation. The finding casts doubt on the notion of the theoryless learner and has implications for effective pedagogy as well as the correction of misinformation.

Sampling assumptions can effect the perception of premise relevance in a way that changes the conclusion drawn from an inductive argument.

The work in Chapters 2–4 examined generalisation in what may be thought of as a *knowledge-poor* context. Because the stimuli involved were perceptual, artificial and low-dimensional there is little background knowledge that can be recruited to assist inference. Thus the relevant basis on which inductive generalisation should proceed is relatively clear. The study described in Chapter 5 considered a *knowledge-rich* scenario, where people reasoned from high-dimensional natural concepts. In such cases, the relevant basis for generalisation may not be clear. From the results of the category-based induction task involving non-diverse premises, the study revealed how sampling assumptions affect the perception of relevance. When people were told that an argument's premises were sampled by a helpful confederate they reasoned that the lack of premise diversity was relevant to the argument. When instead the sampling process appeared random (and filler trials involved negative evidence) people effectively ignored the "coincidently" non-diverse premises.

Reciprocal meta-inference shapes the production and comprehension of evidence: cooperative and misleading communication strategies may exert bi-directional influence as a result.

Previous work (N. D. Goodman & Frank, 2016; Shafto et al., 2014) had established that a form of *reciprocal meta-inference* shapes the way people reason about evidence both when producing it and interpreting it. Inspired by these findings, the work described in Chapter 6 used a "deception game" to examine what happens when cooperation is no longer explicitly or tacitly apparent. The study confirmed that meta-inferential considerations affect production and comprehension behaviour in situations where deception is motivated (but outright lying is not an option). The results suggested an intriguing bi-directional influence between cooperative and misleading communication. Because some data is more helpful than others, people look for signs of cooperation in the data and calibrate their sampling assumptions accordingly. Would-be deceivers mimic the ostensive signs of cooperation to benefit in their deceit from the stronger misleading implications such signs afford. To avoid the possibility of this form of deception, reasoners may adopt a conservative stance towards inference, placing a greater burden of evidence on their interlocutor.

7.2 On people's sensitivity to the sample

Inductive inference, characteristically, involves reasoning on the basis of limited samples of data. Understanding the implications of this data shortage remains a central challenge for cognitive science. If reasoning on the basis of limited data is to be accurate, or at least not systematically biased, then it is important for the reasoner to understand the way that the sample was composed. Yet, a widely held view is that reasoning proceeds on the basis of what is in the sample (*the data*) and not on how the data came to be. Such a view is implicit in many of the models reviewed in Chapter 1. Models which, like the GCM (Nosofsky, 1986) for example, have found considerable success in capturing people's performance in what are ostensibly inductive reasoning tasks, despite the lack of explicit mechanisms for capturing people's sampling assumptions. And the view is expressed more explicitly too. Kahneman's (2011, ch. 7) notion of *what you see is all there is*, for example, captures the idea that because people reason on the basis of evidence retrieved from memory, they fail to take account of how their own memory sample may be unrepresentative of the world in ways that bias judgments. Fiedler (2012) takes a different perspective, where empirical reality and not reasoning deficits play a significant role, but reaches a similar conclusion. According to Fiedler, while people display considerable accuracy with regard to the statistical properties of the sample itself, they are largely oblivious to the origin of the data and the processes that lie beneath it. Fiedler (2008) provides a striking demonstration of what he terms *meta-cognitive myopia*, where people are largely insensitive to sampling concerns despite having sampled the data themselves.

Somewhat in contrast to this view stands the extant literature on the role of sampling assumptions in shaping inductive inference (e.g., Gweon et al., 2010; Hayes, Banner, et al., 2019; Hayes, Navarro, et al., 2019; Lawson & Kalish, 2009; Navarro et al., 2012; Tenenbaum & Griffiths, 2001a; Voorspoels et al., 2015; Xu & Tenenbaum, 2007a), that suggests that people do take the sampling process into account when reasoning from the contents of the sample. By demonstrating further important aspects of people's sensitivity to sampling, the findings I report bolster the evidence in support of this latter view. Study 1 (Chapters 2–4) demonstrated that people used what they knew about the way in which examples had been selected to decide whether or not sample variability and category base rates were representative of the to-be-learned categories. While the sampling manipulations had only a modest effect on placement of the category boundary, the selective use of such implicit negative evidence has implications for the efficiency with which categories may be learned, as well as the representation that is ultimately acquired. Study 2 (Chapter 5) provided what is perhaps a more stark demonstration of people's selective sensitivity to sample variability. Whether further positive examples increased or decreased rates of property projection was shown to depend on people's sampling assumption. Together the results demonstrate that what you see *isn't* all there is – what you don't see may also be relevant, depending on how the sample was constructed.

An interesting avenue for future work is to investigate the discrepancies between these two bodies of literature which prima facie at least, seem to be in conflict. If people are sometimes, but not always sensitive to the origins of data, then what might distinguish these cases? What are the cues that people take account of when forming their assumptions in the first place, and which cues tend to go unnoticed? In each of my first two studies, the experiments used cover stories and manipulations that were somewhat explicit regarding the particular ways that stimuli were sampled. In keeping with the majority of studies which have explicitly manipulated sampling assumptions (but with Hayes, Banner, et al., 2019; Lawson & Kalish, 2009 as notable exceptions), my own involved the contrast between socially sampled data and something closer to "naturally" sampled data. The results provide fresh evidence that people can reliably distinguish between the sampling implications when data is "encountered at random" or when instead it is provided by another to demonstrate (Study 1), persuade (Study 2), or misinform (Study 3). But of course, everyday encounters with data are often much less clearly signposted. If reasoners are able to use the data itself to infer the appropriate assumption or to distinguish amongst alternatives, this might confer significant advantages in terms of staying calibrated with the data. So, in the absence of more explicit guidance regarding the sample's origins, what cues might people pick up on that are *embedded in the data itself*?

My own findings provide some suggestive evidence in this regard. In Study 1 (Chapters 2–4), I used the same basic training and test procedure employed in Hendrickson et al. (2019), including identical stimuli. As discussed in Chapter 3 (see Table 3.2), this provides evidence regarding the "default" sampling assumption that people adopt when learning from examples of one or more categories. In the one-category case, the default assumption appears to be consistent with a strong (or helpful) sampling assumption. Given that category labels are implicitly social constructs, and not a property of natural kinds or objects per se, it is perhaps unsurprising that the "default" assumption where no explicit guidance was given (Hendrickson et al., 2019, Expt. 2) and the assumption that people adopted when told that examples had been "helpfully sampled" (Chapter 2), produced very similar patterns of generalisation performance. Yet this explanation is not completely convincing. For if the use of category labels were to reliably signal that examples were sampled from the relevant concept (per strong sampling), then the same default assumption should apply in the two category case. But the effect of "no manipulation" (Hendrickson et al., 2019, Expt. 1) in the two-category context more closely resembled that of the "random sampling" manipulation (Chapter 3) than it did the "helpful sampling" manipulation. Further, despite being given explicit instruction that additional exemplars had been sampled with helpful intent from the category in question, the extra information had no effect on generalisation performance in the relevant condition. This suggests that the implications of "helpfully sampled" data are not always as clear as might be expected, and that the introduction of "mixed samples" (evidence of for more than one kind of thing or consequence) significantly impacts the strength of the sampling assumption adopted.

Study 2 (Chapter 5) also provides suggestive evidence concerning the cues to the sample's origin that can be extracted from its composition. The experimental investigation involved two conditions (BOTH RELEVANT and BOTH RANDOM) which manipulated both the sampling explanation with which people were presented, as well as the filler items used in control trials. Aligned with these were two further conditions (RELEVANT FILLERS and RANDOM FILLERS) that employed corresponding filler items but no cover story. The overall pattern of results suggested that the filler items (those appearing in the trials preceding the target trials) contributed to the sampling assumption that people adopted. The key difference in the filler items was that one set involved negative evidence (categories not exhibiting the property in question) while the other involved exclusively positive evidence. Like the introduction of a second category in Study 1, the use of negative filler items in Study 2 acted to weaken the sampling assumption that people adopted (see Figure 5.3).

A further intruiging possibility regarding the connection between sample composition and sampling assumptions, is that the sampling process itself is treated as another unknown and that inference proceeds jointly over the original hypotheses of interest and different assumptions about how the data might be sampled. Such a capacity, if reasoners do indeed exercise it, might go some way to explaining how context-specific default assumptions are bootstrapped in the first place via a process analogous to the way that overhypotheses are acquired. Support for related forms of joint inference have been explored in the literature. Notably, work by Goodman and colleagues (see N. D. Goodman & Frank, 2016 for a detailed discussion) has found that a version of the Rational Speech Act framework extended to express uncertainty about the speaker, has successfully accounted for how people infer the meaning of figurative forms of speech (such as hyperbole and irony, for example) as well as the use of scalar adjectives (like *tall*, for example). In my own work, Study 3 (Chapter 6) suggests two distinct (though related) strands of evidence for the notion that people jointly consider how the data was sampled at the same time as trying to draw a direct inference from it. In the comprehension task, people (or a sizeable sub-group thereof) appeared to vary the strength of the assumption they adopted based on the contents of the sample; and in the corresponding production task, people appeared to act consistently with the assumption that this is what people do.

When taken together, I think the findings in this thesis license a cautious interpretation, that we as reasoners stand somewhere between total blindness to what lies beneath the data, and fully Bayesian reasoners who integrate over multiple sources of uncertainty in the data. But although the studies have demonstrated people's sensitivity to various sampling manipulations, and for the most part the assumptions people appear to have made have sensible computational interpretations[1], just what it means to make a "sampling assumption" is less clear from my experiments. In the Bayesian computational models I have described and used throughout the thesis, sampling assumptions are captured by the likelihood function. But it is interesting to consider what the reasoner's own perception of their assumptions are. In the study of meta-inference described in Chapter 6, it is tempting to conclude that people's assumptions have a "theory-like" status and are explicit in a way that might be accessible and recountable by them. Yet, on the basis of the findings of Study 1, the "theory-like" status seems more questionable. For example, the experiment described in Chapter 4 found no effect of sampling manipulation when the sampling explanation was given after the training stimuli were presented. The possibility arises, that what we refer to as "sampling assumptions" might reflect a number of qualitatively different learning or reasoning phenomena.

In light of this uncertainty regarding the status of sampling assumptions and the wider debate regarding the extent to which people make them, the development of methods for explicitly measuring assumptions seems worthwhile. At first glance, various techniques seem worthy of investigation. For example, a simple technique would be to present people with a sampling cover story, and ask them to rate different samples for consistency with the explanation given. By systematically varying the composition of the alternative samples and examining the effect on ratings, various insights might be gained. Regarding the plausibility of proposed cover story manipulations, such information would be useful in a methodological sense. From a theoretical perspective, there is the potential to examine how tightly people hold their sampling assumptions in different circumstances. Techniques for likelihood elicitation could also prove useful. It is easy to conceive a variety of options: from asking people to provide examples of likely (or unlikely) data items; to rating alternatives for their relative likelihood. Were any such technique to prove feasible, it could provide useful insight regarding the way that learning a concept of interest and learning about the generative process may jointly affect one another. And it is possible that it may reveal a disparity between the "elicited likelihood" and one inferred on the basis of Bayesian computational analysis.

A final remark regarding people's sensitivity to the sample, concerns a matter that is both a limit to the generality of the work in Study 1, as well as an opportunity for

[1] A least, when viewed in the aggregate. It is important to note, as I discuss in section 7.3, that individual results may vary.

further research. The experiments involved in the first study demonstrated that people do consider the process that lies behind the data when attempting to learn from it. But what the long term impact of such assumptions are is an open question. Do people really remember the conditions under which data was sampled in the longer term? Fiedler (2008) suggests that *"we do not keep separate memories or archives for experiences that are subject to different sampling constraints"*. An interesting question is whether we need to keep "separate memories" in the episodic sense, or whether the sampling assumptions that people hold affect the longer term representation in memory of the "likelihood" of previous observations. The preliminary investigation I describe in Chapter 4, suggests that this may be the case, or at the very least does not preclude this possibility. A potentially interesting line of investigation would be to examine the effect of sampling assumptions in a multi-stage learning scenario, where different assumptions might reasonably apply at different stages. After all, category learning outside the lab is rarely a one-sitting affair – we frequently learn in *wicked environments* (Hogarth, Lejarraga, & Soyer, 2015) from a mix of socially sampled and naturally sampled data.

7.3 On individual differences

A recurring phenomena observed in the studies I have conducted concerns evidence of qualitative individual differences in patterns of inference. For reasons I discuss in some detail in Chapters 2 and 6, there is reason to believe that these differences are not merely experimental artefacts arising from the particular tasks employed. In any case, there is nothing unusual about such a finding – it is a commonplace in experimental studies of cognition in general, and in particular has been observed in experiments comparable to the ones I have conducted. Navarro et al. (2012), for example, observed individual differences in performance in a perceptual generalisation task comparable to the one described in Chapters 2–4, albeit with different stimuli and cover story manipulations. In a study of the effect of sampling assumptions on the perception of evidentiary diversity (closely related to the experiment described in Chapter 5), Hayes, Navarro, et al. (2019) noted considerable individual difference in ratings of inductive argument strength. And using a signalling game somewhat analogous to the *deception game* described in Chapter 6, Franke and Degen (2016) found individual differences in the production and comprehension of referential expressions.

The prevalence of individual differences in experiments like these represent a challenge for the researcher that is worthwhile acknowledging. Depending on the research question being investigated and the approach adopted, it may be preferable to attempt to reduce the extent to which such differences manifest themselves, to account for them as nuisance variables, or to fully explicate the differences in some way.[2] Regardless of the appropriate

[2] See Tauber et al. (2017) for a detailed discussion of related issues in the context of Bayesian computational modelling.

course of action, the issue of individual difference should not lightly be ignored. This point has been well made in the literature (e.g., Gigerenzer & Brighton, 2009; Webb & Lee, 2004). I raise it again here because my own studies highlight not only the prevalence but also the impact of the issue. Namely, that aggregate performance on the kinds of tasks I have used may represent a poor description of many or even the majority of individuals, masking the very behaviour that is of interest.

Notwithstanding the methodological issues, a deeper understanding of what is driving the kinds of individual differences discussed is important from a theoretical perspective. As with the study of inductive inference in general, the question of individual differences may be posed at different levels: computational, algorithmic and implementational, for example (Marr, 1982). Even at the computational level, qualitatively different explanations can be entertained. A Bayesian computational analysis admits many possibilities, for example. Differences in prior belief or sampling assumptions are an obvious candidate. Navarro et al. (2012) inferred differences in both. Computational modelling suggested that people had different prior expectations regarding the extent of the one dimensional range they were asked to infer. But as the authors conceived, there is also the possibility that people entertain alternative representations of a consequential region even in what might seem like clear cut cases. Indeed, this is just what the analysis in Chapter 2 shows. Regarding sampling assumptions, Navarro et al. (2012) suggested that people might vary on a characteristic of inferential conservatism that affects the strength of assumption they adopt. Computationally speaking, this characteristic is represented as a continuous linear dimension spanning a weak sampling assumption at one extreme to a strong sampling assumption at the other. In Bayesian terms, a value on this scale is something like a prior bias (based on experience or *ab initio*) regarding the appropriateness of following the size principle to reason beyond the data. A reasonable psychological interpretation is that this form of conservatism reflects the reasoner's degree of confidence that the observations they are seeing are representative of the concept of interest.

There are other interpretations that are worth considering. While reasoning beyond the data can be an efficient way to learn, it may negatively impact accuracy where the underlying assumptions turn out to be incorrect. It is interesting to speculate whether this might trigger (or induce) an element of loss aversion through the avoidance of regret. For example, I might feel reluctant to accept a pragmatic implication of an utterance because it feels like a commitment I am making, rather than the speaker. This idea has been discussed in the deception literature. The suggestion is that the moral distinction between outright lies and misleading implications arises in part because the listener accepts some responsibility for drawing the inference (Adler, 1997). Although it is perhaps more compelling to consider loss aversion in the context of socially generated data (because of potential for deception involved), it is interesting to consider whether a similar aversion might apply in cases where data is generated by the environment. In sum, an interesting line of future investigation would be to drill down further on the kind of stance that this form of conservatism represents.

So far, I have discussed how different sampling assumptions and different priors (including priors over assumptions) may lead to individual differences in patterns of inductive reasoning. Staying at a computational level abstraction, there is also the possibility that individuals vary in terms of the complexity of computation performed. The work in Chapter 6 raised two points on which individuals might differ. Firstly, individuals may vary in the degree to which they view the problem to be solved as one of joint inference regarding the meaning of socially generated data and the manner in which it was generated. And secondly, individuals may vary in the depth and complexity of "theory of mind" style reasoning. Further research is needed in both areas to clarify the scale and impact of such differences.

Lastly, it is worth noting that the presence of pervasive individual differences in sampling assumptions and inferential stance presents a further opportunity for research. For each kind of individual difference observed, we can ask a related question: when we collaborate and communicate on the basis of socially generated data, are we as reasoners sensitive to such individual differences? For example, to what extent do people as speakers take account of trait-like characteristics like a tendency for overly-literal interpretation (per conservative, or weak sampling). Likewise, do people track the "confidence" with which speakers state their beliefs when trying to infer the strength of their underlying evidence base? If, for example, people are able to perform the kinds of joint inference that the results in Chapter 6 suggest, then it is at least feasible that they might do so. Interestingly, if people track such tendencies over the population of speakers and listeners as a whole, the possibility for strong bi-directional influence emerges: a tendency for conservative meta-inference on the part of listeners may be influenced by a tendency to overstate the evidence on the part of speakers, and vice versa. A fully-fledged theory of meta-inference may need to take meta-inferential sensitivity to individual differences, and bi-directional influences into account.

7.4 It's a rap

I've made a little study of how we take data, elementary
and through assumptions calculate its value, evidentiary.
It has close ties to research, some that's quite historical.
I've used stimuli, perceptual and some that's categorical.
Along the way I've noted many individual differences,
which may have their consequences shaping our meta-inferences
I've shown how our assumptions affect comprehension and production.
Yet this all barely scrapes the surface of a science of induction...

Yes, this all barely scrapes the surface of a science of induction.

Ok, that's more Gilbert and Sullivan than Eminem. Perhaps I'll stick to research.

REFERENCES

Adler, J. E. (1997). Lying, deceiving, or falsely implicating. *The Journal of Philosophy*, *94*(9), 435–452.

Anderson, J. R. (1991). The adaptive nature of human categorization. *Psychological Review*, *98*(3), 409–429.

Arts, A., Maes, A., Noordman, L., & Jansen, C. (2011). Overspecification facilitates object identification. *Journal of Pragmatics*, *43*(1), 361–374.

Ashby, F. G., & Townsend, J. T. (1986). Varieties of perceptual independence. *Psychological Review*, *93*(2), 154–179.

Attneave, F. (1957). Transfer of experience with a class-schema to identification-learning of patterns and shapes. *Journal of experimental psychology*, *54*(2), 81–88.

Austerweil, J., & Griffiths, T. (2010). Learning hypothesis spaces and dimensions through concept learning. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 32).

Avrahami, J., Kareev, Y., Bogot, Y., Caspi, R., Dunaevsky, S., & Lerner, S. (1997). Teaching by examples: Implications for the process of category acquisition. *Quarterly Journal of Experimental Psychology Section A*, *50*(3), 586-–606.

Baker, C. L., Saxe, R., & Tenenbaum, J. B. (2009). Action understanding as inverse planning. *Cognition*, *113*(3), 329–349.

Barsalou, L. W. (1989). Intraconcept similarity and its implications for interconcept similarity. In S. Vosniadou & A. Ortony (Eds.), *Similarity and analogical reasoning* (p. 76–121). Cambridge: Cambridge University Press.

Bonawitz, E., Shafto, P., Gweon, H., Goodman, N. D., Spelke, E., & Schulz, L. (2011). The double-edged sword of pedagogy: Instruction limits spontaneous exploration and discovery. *Cognition*, *120*(3), 322–330.

Bond, C., & DePaulo, B. (2006). Accuracy of deception judgments. *Personality and Social Psychology Review*, *10*(3), 214-234.

Bush, R. R., & Mosteller, F. (1951). A model for stimulus generalization and discrimination. *Psychological Review*, *58*(6), 413–423. doi: 10.1037/h0054576

Carnap, R. (1967). *The logical structure of the world and Pseudoproblems in philosophy* (R. A. George, Trans.). Berkeley, CA: University of California Press.

Coleman, L., & Kay, P. (1981). Prototype semantics: The english word lie. *Language*, *57*(1), 26–44.

Coley, J. D., & Vasilyeva, N. Y. (2010). Chapter 5 - generating inductive inferences: Premise relations and property effects. In B. H. Ross (Ed.), *The psychology of learning and motivation: Advances in research and theory* (Vol. 53, pp. 183–226).

Academic Press.

Collins, G. P. (2007). The many interpretations of quantum mechanics. *Scientific American*, *297*(5), 19.

Colman, A. (2003). Depth of strategic reasoning in games. *Trends in Cognitive Sciences*, *7*(1), 2–4.

Corner, A., Harris, A., & Hahn, U. (2010). Conservatism in belief revision and participant skepticism. In S. Ohlsson & R. Catrambone (Eds.), *Proceedings of the 32nd Annual Meeting of the Cognitive Science Society* (pp. 1625–1630).

Csibra, G., & Gergely, G. (2006). Social learning and social cognition: the case for pedagogy. In Y. Munakata & M. H. Johnson (Eds.), *Processes of change in brain and cognitive development. Attention and performance xxi* (p. 249-–274). Oxford: Oxford University Press.

Csibra, G., & Gergely, G. (2009). Natural pedagogy. *Trends in cognitive sciences*, *13*(4), 148–153.

Duda, R., Hart, P., & Stork, D. (2001). *Pattern classification*. New York, NY: Wiley.

Dulcinatti, G. (2018). *Cooperation and pragmatic inferences* (Unpublished doctoral dissertation). University College London.

Durkin, K., Caglar, L. R., Bonawitz, E., & Shafto, P. (2015). Explaining choice behavior: The intentional selection assumption. In D. C. Noelle et al. (Eds.), *Proceedings of the 37th Annual Meeting of the Cognitive Science Society* (pp. 614–619).

Dynel, M. (2011). A web of deceit: A neo-gricean view on types of verbal deception. *International Review of Pragmatics*, *3*(2), 531–538.

Eaves, B. S., & Shafto, P. (2012). Unifying pedagogical reasoning and epistemic trust. In *Advances in child development and behavior* (Vol. 43, pp. 295–319). Elsevier.

Ecker, U., Lewandowsky, S., Swire, B., & Chang, D. (2011). Correcting false information in memory: Manipulating the strength of misinformation encoding and its retraction. *Psychonomic Bulletin & Review*(18), 570–578.

Edwards, W. (1968). Conservatism in human information processing. In B. Kleinmuntz (Ed.), *Formal representation of human judgment* (pp. 17–52)). New York, NY: Wiley.

Engelhardt, P. E., Bailey, K. G., & Ferreira, F. (2006). Do speakers and listeners observe the gricean maxim of quantity? *Journal of Memory and Language*, *54*(4), 554–573.

Eskritt, M., Whalen, J., & Lee, K. (2008). Preschoolers can recognize violations of the gricean maxims. *British Journal of Developmental Psychology*, *26*(3), 435–443.

Estes, W. K. (1950). Toward a statistical theory of learning. *Psychological review*, *57*(2), 94.

Feeney, A., Coley, J. D., & Crisp, A. K. (2010). The relevance framework for category-based induction: Evidence from garden-path arguments. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *36*(4), 906–919.

Feeney, A., & Heit, E. (2011). Properties of the diversity effect in category-based inductive reasoning. *Thinking & Reasoning*, *17*(2), 156–181.

Feeney, A., Scrafton, S., Duckworth, A., & Handley, S. J. (2004). The story of some: Everyday pragmatic inference by children and adults. *Canadian Journal of Experimental Psychology/Revue Canadienne de Psychologie Expérimentale*, *58*(2), 121–132.

Fernbach, P. M. (2006). Sampling assumptions and the size principle in property induction. In R. Sun, G. W. Cottrell, & N. Miyake (Eds.), *Proceedings of the 28th Annual Meeting of the Cognitive Science Society* (pp. 1287–1293).

Fiedler, K. (2008). The ultimate sampling dilemma in experience-based decision making. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *34*(1), 186–203.

Fiedler, K. (2012). Meta-cognitive myopia and the dilemmas of inductive-statistical inference. In *Psychology of learning and motivation* (Vol. 57, pp. 1–55). Elsevier.

Frank, M., & Tenenbaum, J. B. (2011). Three ideal observer models for rule learning in simple languages. *Cognition*, *120*, 360–371.

Frank, M. C., & Goodman, N. D. (2012). Predicting pragmatic reasoning in language games. *Science*, *336*(6084), 998.

Franke, M., & Degen, J. (2016). Reasoning in reference games: Individual-vs. population-level probabilistic modeling. *PloS one*, *11*(5), e0154854.

Frederick, R. I., & Speed, F. M. (2007). On the interpretation of below-chance responding in forced-choice tests. *Assessment*, *14*(1), 3–11.

Fried, L. S., & Holyoak, K. J. (1984). Induction of category distributions: A framework for classification learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *10*(2), 234–257.

Gigerenzer, G., & Brighton, H. (2009). Homo heuristicus: Why biased minds make better inferences. *Topics in Cognitive Science*, *1*(1), 107–143.

Good, I. J. (1960). The paradox of confirmation. *The British Journal for the Philosophy of Science*, *11*(42), 145–149.

Goodman, N. (1955). *Fact, fiction, & forecast*. Cambridge, MA: Harvard University Press.

Goodman, N. (1972). *Problems and projects*. Indianapolis: Bobbs-Merrill.

Goodman, N. D., & Frank, M. C. (2016). Pragmatic language interpretation as probabilistic inference. *Trends in Cognitive Sciences*, *20*(11), 818–829.

Goodman, N. D., & Stuhlmüller, A. (2013). Knowledge and implicature: Modeling language understanding as social cognition. *Topics in cognitive science*, *5*(1), 173–184.

Grice, H. P. (1989). *Studies in the way of words*. Cambridge, MA: Harvard University Press.

Griffiths, T., Canini, K., Sanborn, A., & Navarro, D. (2007). Unifying rational models of categorization via the hierarchical Dirichlet process. In M. D. S. & T. J. G.

(Eds.), *Proceedings of the 29th Annual Meeting of the Cognitive Science Society* (pp. 323–328).

Gweon, H., Tenenbaum, J. B., & Schulz, L. E. (2010). Infants consider both the sample and the sampling process in inductive generalization. *Proceedings of the National Academy of Sciences*, *107*(20), 9066–9071.

Hardin, K. J. (2010). The Spanish notion of *lie*: Revisiting Coleman and Kay. *Journal of Pragmatics*, *42*(12), 3199–3213.

Harman, G. H. (1965). The inference to the best explanation. *The Philosophical Review*, *74*(1), 88.

Harris, A. J. L., Corner, A., & Hahn, U. (2013). James is polite and punctual (and useless): A bayesian formalisation of faint praise. *Thinking & Reasoning*, *19*(3-4), 414–429.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning*. New York, NY: Springer.

Hawkins, R. X. D., Stuhlmüller, A., Degen, J., & Goodman, N. D. (2015). Why do you ask? good questions provoke informative answers. In D. C. Noelle et al. (Eds.), *Proceedings of the 37th Annual Meeting of the Cognitive Science Society* (pp. 878–883).

Hawthorne-Madell, D., & Goodman, N. D. (2015). So good it has to be true: Wishful thinking in theory of mind. In D. C. Noelle et al. (Eds.), *Proceedings of the 37th Annual Meeting of the Cognitive Science Society* (pp. 884–889).

Hayes, B. K., Banner, S., Forrester, S., & Navarro, D. J. (2019). Selective sampling and inductive inference: Drawing inferences based on observed and missing evidence. *Cognitive Psychology*, *113*.

Hayes, B. K., Banner, S., & Navarro, D. J. (2017). Sampling frames, Bayesian inference and inductive reasoning. In G. Gunzelmann, A. Howes, T. Tenbrink, & E. Davelaar (Eds.), *Proceedings of the 39th Annual Meeting of the Cognitive Science Society* (pp. 488–493). Cognitive Science Society.

Hayes, B. K., Navarro, D. J., Stephens, R. G., Ransom, K. J., & Dilevski, N. (2019). The diversity effect in inductive reasoning depends on sampling assumptions. *Psychonomic Bulletin & Review*, 1–8.

Hebb, D. O. (1949). *The organization of behavior: A neuropsychological theory*. John Wiley & Sons.

Hedden, T., & Zhang, J. (2002). What do you think i think you think?: Strategic reasoning in matrix games. *Cognition*, *85*(1), 1–36.

Heit, E. (1998). A Bayesian analysis of some forms of inductive reasoning. In M. Oaksford & N. Chater (Eds.), *Rational models of cognition* (pp. 248–274). Oxford: Oxford University Press.

Heit, E., & Rubinstein, J. (1994). Similarity and property effects in inductive reasoning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*(2), 411–422.

Hempel, C. G. (1966). *Philosophy of natural science*. Englewood Cliffs, N.J: Prentice-Hall.

Hendrickson, A. T., Perfors, A., Navarro, D. J., & Ransom, K. J. (2019). Sample size, number of categories and sampling assumptions: Exploring some differences between categorization and generalization. *Cognitive Psychology*, *111*, 80–102.

Hespanha, J. P., Ateskan, Y. S., & Kizilocak, H. (2000). Deception in non-cooperative games with partial information. In *Proceedings of the 2nd darpa-jfacc symposium on advances in enterprise control* (pp. 1–9).

Hinsey, W. C. (1963). *Identification-learning after pretraining on central and noncentral standards* (Unpublished doctoral dissertation). University of Oregon.

Hogarth, R. M., Lejarraga, T., & Soyer, E. (2015). The two settings of kind and wicked learning environments. *Current Directions in Psychological Science*, *24*(5), 379–385.

Homa, D., Sterling, S., & Trepel, L. (1981). Limitations of exemplar-based generalization and the abstraction of categorical information. *Journal of Experimental Psychology: Human Learning and Memory*, *7*(6), 418.

Horwich, P. (2016). Evidence. In *Probability and evidence* (p. 110–120). Cambridge University Press.

Hovland, C. I. (1937). The generalization of conditioned responses. iv. the effects of varying amounts of reinforcement upon the degree of generalization of conditioned responses. *Journal of Experimental Psychology*, *21*(3), 261–276.

Hsu, A., & Griffiths, T. L. (2016). Sampling assumptions affect use of indirect negative evidence in language learning. *PLoS ONE*, *11*(6), 1–20.

Hull, C. L. (1943). *Principles of behavior: An introduction to behavior theory.* New York: Appleton-Century.

Hume, D. (1748/2007). *An enquiry concerning human understanding*. Oxford: Oxford University Press.

Humphreys, L. G. (1939). Generalization as a function of method of reinforcement. *Journal of Experimental Psychology*, *25*(4), 361–372.

Jaynes, E. T. (2003). *Probability theory: The logic of science*. Cambridge, UK: Cambridge University Press.

Jern, A., Lucas, C. G., & Kemp, C. (2017). People learn other people's preferences through inverse decision-making. *Cognition*, *168*, 46–64.

Johnson, H., & Seifert, C. (1994). Sources of the continued influence effect: When misinformation in memory affects later inferences. *Jn Exp Psych: LMC*(20), 1420–1436.

Kahneman, D. (2011). *Thinking, fast and slow*. London, UK: Penguin Books.

Kalish, M. L., & Kruschke, J. K. (1997). Decision boundaries in one-dimensional categorization. *Journal of Experimental Psychology: Learning, Memory, and*

Cognition, *23*(6), 1362–1377.

Kashimori, A. (2017). *The illustrated egg handbook*. Context Publications.

Keil, F. C., Smith, W. C., Simons, D. J., & Levin, D. T. (1998). Two dogmas of conceptual empiricism: implications for hybrid models of the structure of knowledge. *Cognition*, *65*(2), 103–135.

Kemp, C., Perfors, A., & Tenenbaum, J. B. (2007). Learning overhypotheses with hierarchical bayesian models. *Developmental Science*, *10*(3), 307-321.

Kemp, C., & Tenenbaum, J. B. (2009). Structured statistical models of inductive reasoning. *Psychological Review*, *116*(1), 20–58.

Klayman, J., & Ha, Y.-W. (1987). Confirmation, disconfirmation, and information in hypothesis testing. *Psychological Review*, *94*(2), 211–218.

Klugkist, I., Laudy, O., & Hoijtink, H. (2005). Inequality constrained analysis of variance: a Bayesian approach. *Psychological Methods*, *10*(4), 477–493.

Kruschke, J. K. (1992). Alcove: An exemplar-based connectionist model of category learning. *Psychological Review*, *99*(1), 22-–44.

Landrum, A. R., Eaves, B. S., & Shafto, P. (2015). Learning to trust and trusting to learn: a theoretical framework. *Trends in Cognitive Sciences*, *19*(3), 109 - 111.

Laplace, P. S. (1825/1985). *Philosophical essays on probabilities*. Dover.

Lashley, K. S., & Wade, M. (1946). The pavlovian theory of generalization. *Psychological Review*, *53*(2), 72–87.

Lawson, C. A., & Kalish, C. W. (2009). Sample selection and inductive generalization. *Memory & Cognition*, *37*, 596–607.

Lee, J. J., & Pinker, S. (2010). Rationales for indirect speech: The theory of the strategic speaker. *Psychological Review*, *117*(3), 785–807.

Lee, M. D. (2002). Generating additive clustering models with minimal stochastic complexity. *Journal of Classification*, *19*(1), 69–85.

Levine, T. R., & McCornack, S. A. (2014). Theorizing about deception. *Journal of Language and Social Psychology*, *33*(4), 431-440.

Levinson, S. C. (2000). *Presumptive meanings: The theory of generalized conversational implicature*. Cmabridge, MA: MIT press.

Lewis, M. L., & Frank, M. C. (2016). Understanding the effect of social context on learning: A replication of Xu and Tenenbaum (2007b). *Journal of Experimental Psychology: General*, *145*(9).

Liang, F., Paulo, R., Molina, G., Clyde, M. A., & Berger, J. O. (2012). Mixtures of g priors for Bayesian variable selection. *Journal of the American Statistical Association*, 410–423.

Lloyd, S. (2002). Computational capacity of the universe. *Physical Review Letters*, *88*(23), 237901.

Lodewyckx, T., Kim, W., Lee, M. D., Tuerlinckx, F., Kuppens, P., & Wagenmakers, E.-J. (2011). A tutorial on Bayes factor estimation with the product space method.

Journal of Mathematical Psychology, *55*(5), 331–347.

Love, B. C., Medin, D. L., & Gureckis, T. M. (2004). Sustain: a network model of category learning. *Psychological Review*, *111*(2), 309–332.

Luce, R. D. (1959). *Individual choice behavior: A theoretical analysis*. John Wiley & Sons, Inc.

Margolius, G. (1955). Stimulus generalization of an instrumental response as a function of the number of reinforced trials. *Journal of Experimental Psychology*, *49*(2), 105–111.

Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information*. San Francisco, CA: W. H. Freeman.

McKenzie, C. R. M., & Mikkelsen, L. A. (2007). A bayesian view of covariation assessment. *Cognitive Psychology*, *54*(1), 33–61.

Medin, D. L., Coley, J. D., Storms, G., & Hayes, B. L. (2003). A relevance theory of induction. *Psychonomic Bulletin and Review*, *10*(3), 517–532.

Medin, D. L., Goldstone, R. L., & Gentner, D. (1993). Respects for similarity. *Psychological review*, *100*(2), 254–278.

Montague, R., Navarro, D., Perfors, A., Warner, R., & Shafto, P. (2011). To catch a liar: The effects of truthful and deceptive testimony on inferential learning. In L. Carlson, C. Holscher, & T. Shipley (Eds.), *Proceedings of the 33rd annual meeting of the cognitive science society.* Austin, TX: Cognitive Science Society.

Morey, R. D., & Rouder, J. N. (2014). BayesFactor: Computation of Bayes factors for common designs [Computer software manual]. Retrieved from http://CRAN.R-project.org/package=BayesFactor (R package version 0.9.9)

Morey, R. D., & Wagenmakers, E.-J. (2014). Simple relation between Bayesian order-restricted and point-null hypothesis tests. *Statistics and Probability Letters*, *92*, 121-124.

Murphy, G. L., & Medin, D. L. (1985). The role of theories in conceptual coherence. *Psychological Review*, *92*(3), 289–316.

Navarro, D. J. (2006). From natural kinds to complex categories. In R. Sun, G. W. Cottrell, & N. Miyake (Eds.), *Proceedings of the 28th Annual Meeting of the Cognitive Science Society* (pp. 621–626).

Navarro, D. J., Dry, M. J., & Lee, M. D. (2012). Sampling assumptions in inductive generalization. *Cognitive Science*, *36*(2), 187–223.

Navarro, D. J., & Griffiths, T. L. (2008). Latent features in similarity judgments: A nonparametric Bayesian approach. *Neural computation*, *20*(11), 2597–2628.

Navarro, D. J., & Perfors, A. F. (2010). Similarity, feature discovery and the size principle. *Acta Psychologica*, *133*, 256–268.

Navarro, D. J., & Perfors, A. F. (2011). Hypothesis generation, sparse categories, and the positive test strategy. *Psychological Review*, *118*(1), 120–134.

Nisbett, R. E., Krantz, D. H., Jepson, C., & Kunda, Z. (1983). The use of statistical heuristics in everyday inductive reasoning. *Psychological Review*, *90*(4), 339–363.

Nosofsky, R. M. (1986). Attention, similarity, and the identification–categorization relationship. *Journal of Experimental Psychology: General*, *115*(1), 39–57.

Nosofsky, R. M. (1988). Similarity, frequency, and category representations. *Journal of Experimental Psychology: learning, memory, and cognition*, *14*(1), 54.

Oaksford, M., & Chater, N. (1998). *Rational models of cognition*. Oxford: Oxford University Press.

Ohtsubo, Y., & Rapoport, A. (2006). Depth of reasoning in strategic form games. *The Journal of Socio-Economics*, *35*(1), 31–47.

Okanda, M., Asada, K., Moriguchi, Y., & Itakura, S. (2015). Understanding violations of gricean maxims in preschoolers and adults. *Frontiers in Psychology*, *6*, 901. doi: 10.3389/fpsyg.2015.00901

Osherson, D. N., Smith, E. E., Wilkie, O., Lopez, A., & Shafir, E. (1990). Category-based induction. *Psychological Review*, *97*(2), 185.

Pavlov, I. P. (1927). *Conditioned reflexes: an investigation of the physiological activity of the cerebral cortex*. Oxford: Oxford University Press.

Perfors, A., Ransom, K. J., & Navarro, D. J. (2014). People ignore token frequency when deciding how widely to generalize. In *Proceedings of the 36th Annual Meeting of the Cognitive Science Society* (pp. 2759–2764).

Phillips, L. D., & Edwards, W. (1966). Conservatism in a simple probability inference task. *Journal of Experimental Psychology*, *72*(3), 346–354.

Pierce, C. S. (1955). Abduction and induction. In J. Buchler (Ed.), *Philosophical writings of Pierce* (pp. 150–156). New York: Dover Books.

Pinker, S., Nowak, M. A., & Lee, J. J. (2008). The logic of indirect speech. *Proceedings of the National Academy of Sciences*, *105*(3), 833–838. doi: 10.1073/pnas.0707192105

Pogue, A., Kurumada, C., & Tanenhaus, M. K. (2016). Talker-specific generalization of pragmatic inferences based on under- and over-informative prenominal adjective use. *Frontiers in Psychology*, *6*, 2035.

Posner, M. I., & Keele, S. W. (1968). On the genesis of abstract ideas. *Journal of experimental psychology*, *77*(3, part 1), 353–362.

Pryslopska, A. (2013). *Implicatures in uncooperative contexts* (Unpublished master's thesis). University of Tübingen.

Qing, C., & Franke, M. (2015). Variations on a bayesian theme: Comparing bayesian models of referential reasoning. In *Bayesian natural language semantics and pragmatics* (pp. 201–220). Springer.

Quine, W. V. O. (1969). Natural kinds. In *Ontological relativity and other essays* (pp. 114–138). New York: Columbia University Press.

Ransom, K. J., Hendrickson, A., Perfors, A., & Navarro, D. J. (2018). Representational and sampling assumptions drive individual differences in single category generalisation. In C. Kalish, M. Rau, J. Zhu, & T. Rogers (Eds.), *Proceedings of the 40th Annual Meeting of the Cognitive Science Society* (pp. 930–935).

Ransom, K. J., Perfors, A., & Navarro, D. J. (2016). Leaping to conclusions: Why premise relevance affects argument strength. *Cognitive Science*, *40*(7), 1775–1796.

Ransom, K. J., Voorspoels, W., Perfors, A., & Navarro, D. (2017). A cognitive analysis of deception without lying. In *Proceedings of the 39th Annual Meeting of the Cognitive Science Society.* (pp. 992–997).

Razran, G. (1949). Stimulus generalization of conditioned responses. *Psychological Bulletin*, *46*(5), 337–365.

Reboul, A. (2017). Is implicit communication a way to escape epistemic vigilance? In S. Assimakopoulos (Ed.), *Pragmatics at its interfaces* (Vol. 17, pp. 91–112). Walter de Gruyter GmbH.

Rhodes, M., Bonawitz, E., Shafto, P., Chen, A., & Caglar, L. (2015). Controlling the message: Preschoolers' use of information to teach and deceive others. *Frontiers in psychology*, *6*, 867.

Rhodes, M., Gelman, S. A., & Brickman, D. (2010). Children's attention to sample composition in learning, teaching and discovery. *Developmental Science*, *13*(3), 421–429.

Rips, L. (1975). Inductive judgments about natural categories. *Journal of Verbal Learning and Verbal Behavior*, *14*, 665–681.

Rogers, T., Zeckhauser, R. J., Gino, F., Norton, M. I., & Schweitzer, M. E. (2017). Artful paltering: The risks and rewards of using truthful statements to mislead others. *Journal of personality and social psychology*, *112*(3), 456–473.

Rosch, E. (1973). Natural categories. *Cognitive psychology*, *4*(3), 328–350.

Rosch, E. (1975). Cognitive reference points. *Cognitive psychology*, *7*(4), 532–547.

Rosch, E., & Mervis, C. B. (1975). Family resemblances: Studies in the internal structure of categories. *Cognitive Psychology*, *7*(4), 573 - 605.

Ross, B. H., & Murphy, G. L. (1999). Food for thought: Cross-classification and category organization in a complex real-world domain. *Cognitive Psychology*, *38*(4), 495–553.

Rouder, J. N., Morey, R. D., Speckman, P. L., & Province, J. M. (2012). Default Bayes factors for ANOVA designs. *Journal of Mathematical Psychology*, *56*(5), 356–374.

Russell, B. (2006). Against grammatical computation of scalar implicatures. *Journal of semantics*, *23*(4), 361–382.

Sanjana, N. E., & Tenenbaum, J. B. (2003). Bayesian models of inductive generalization. In S. Becker, S. Thrun, & K. Obermayer (Eds.), *Advances in Neural Information Processing Systems 15* (pp. 59–66). MIT Press.

Schauer, F., & Zeckhauser, R. (2009). Paltering. In B. Harrington (Ed.), *Deception: From ancient empires to internet dating* (pp. 38–54). Stanford, CA: Stanford University Press.

Seger, C. A., & Peterson, E. J. (2013). Categorization = decision making + generalization. *Neuroscience & Biobehavioral Reviews*, *37*(7), 1187–1200.

Shafto, P., & Bonawitz, E. (2015). Choice from intentionally selected options. In B. H. Ross (Ed.), *The psychology of learning and motivation* (Vol. 63, pp. 115–139). Academic Press.

Shafto, P., Eaves, B., Navarro, D. J., & Perfors, A. (2012). Epistemic trust: Modeling children's reasoning about others' knowledge and intent. *Developmental Science*, *15*, 436–447.

Shafto, P., & Goodman, N. (2008). Teaching games: statistical sampling assumptions for learning in pedagogical situations. In V. Sloutsky, B. Love, & K. McRae (Eds.), *Proceedings of the 30th Annual Cognitive Science Society.*

Shafto, P., Goodman, N. D., & Frank, M. C. (2012). Learning from others: The consequences of psychological reasoning for human learning. *Perspectives on Psychological Science*, *7*(4), 341–351.

Shafto, P., Goodman, N. D., & Griffiths, T. L. (2014). A rational account of pedagogical reasoning: Teaching by, and learning from, examples. *Cognitive Psychology*, *71*, 55–89.

Shafto, P., Kemp, C., Mansinghka, V., & Tenenbaum, J. B. (2011). A probabilistic model of cross-categorization. *Cognition*, *120*(1), 1–25.

Shepard, R. N. (1957). Stimulus and response generalization: A stochastic model relating generalization to distance in psychological space. *Psychometrika*, *22*(4), 325–345.

Shepard, R. N. (1987). Toward a universal law of generalization for psychological science. *Science*, *237*(4820), 1317–1323.

Shepard, R. N., & Arabie, P. (1979). Additive clustering: Representation of similarities as combinations of discrete overlapping properties. *Psychological Review*, *86*(2), 87–123.

Skarakis-Doyle, E., Izaryk, K., Campbell, W., & Terry, A. (2014). Preschoolers' sensitivity to the maxims of the cooperative principle: Scaffolds and developmental trends. *Discourse Processes*, *51*(4), 333–356.

Sloman, S. A. (1993). Feature-based induction. *Cognitive Psychology*, *25*, 213–280.

Sloman, S. A. (1994). When explanations compete: The role of explanatory coherence on judgments of likelihood. *Cognition*, *52*(1), 1–21.

Smith, E. E., Shafir, E., & Osherson, D. (1993). Similarity, plausibility and judgments of probability. *Cognition*, *49*, 67–96.

Smith, J. D., & Minda, J. P. (1998). Prototypes in the mist: The early epochs of category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *24*(6), 1411.

Sperber, D., Clément, F., Heintz, C., Mascaro, O., Mercier, H., Origgi, G., & Wilson, D. (2010). Epistemic vigilance. *Mind & Language*, *25*(4), 359–393.

Stahl, D. O., & Wilson, P. W. (1995). On players' models of other players: Theory and experimental evidence. *Games and Economic Behavior*, *10*(1), 218–254.

Stiller, A. J., Goodman, N. D., & Frank, M. C. (2015). Ad-hoc implicature in preschool children. *Language Learning and Development*, *11*(2), 176-190.

Tauber, S., Navarro, D. J., Perfors, A., & Steyvers, M. (2017). Bayesian models of cognition revisited: Setting optimality aside and letting data drive psychological theory. *Psychological Review*, *124*(4), 410–441.

Tenenbaum, J. B. (1999). *A Bayesian framework for concept learning.* (Unpublished doctoral dissertation). Massachussets Institute of Technology.

Tenenbaum, J. B. (2000). Rules and similarity in concept learning. In S. A. Solla, T. K. Leen, & K.-R. Müller (Eds.), *Advances in Neural Information Processing Systems 15* (pp. 59–65). MIT Press.

Tenenbaum, J. B., & Griffiths, T. L. (2001a). Generalization, similarity, and Bayesian inference. *Behavioural and Brain Sciences*, *24*(4), 629–640.

Tenenbaum, J. B., & Griffiths, T. L. (2001b). The rational basis of representativeness. In *Proceedings of the 23rd annual conference of the cognitive science society* (pp. 1036–1041).

Topál, J., Gergely, G., Miklósi, Á., Erdőhegyi, Á., & Csibra, G. (2008). Infants' perseverative search errors are induced by pragmatic misinterpretation. *Science*, *321*(5897), 1831–1834.

Tversky, A. (1977). Features of similarity. *Psychological Review*, *84*(4), 327–352.

Ulanowicz, R. E. (2012). *Growth and development: ecosystems phenomenology*. Springer Science & Business Media.

Ullman, T., Baker, C., Macindoe, O., Evans, O., Goodman, N. D., & Tenenbaum, J. B. (2009). Help or hinder: Bayesian models of social goal inference. In *Advances in neural information processing systems* (pp. 1874–1882).

Vanpaemal, W., & Navarro, D. (2007). Representational shifts during category learning. In M. D. S. & T. J. G. (Eds.), *Proceedings of the 29th Annual Meeting of the Cognitive Science Society* (pp. 1599–1604).

Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., … Zhu, X. (2001). The sequence of the human genome. *Science*, *291*(5507), 1304–1351.

Vogel, A., Potts, C., & Jurafsky, D. (2013). Implicatures and nested beliefs in approximate Decentralized-POMDPs. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (volume 2: Short papers)* (pp. 74–80).

Vong, W. K., Hendrickson, A. T., Perfors, A., & Navarro, D. J. (2013). The role of sampling assumptions in generalization with multiple categories. In *Proceedings of the 35th Annual Meeting of the Cognitive Science Society* (pp. 3699–3704).

Voorspoels, W., Navarro, D. J., Perfors, A., Ransom, K., & Storms, G. (2015). How do people learn from negative evidence? non-monotonic generalizations and sampling assumptions in inductive reasoning. *Cognitive Psychology*, *81*, 1–25.

Voorspoels, W., Van Meel, C., & Storms, G. (2013). Negative observations, induction and the generation of hypotheses. In *Proceedings of the 35th Annual Meeting of the Cognitive Science Society* (pp. 1552–1557).

Vrij, A., Leal, S., & Fisher, R. P. (2018). Verbal deception and the model statement as a lie detection tool. *Frontiers in Psychiatry*, *9*, 492.

Walker, D., Smith, K. A., & Vul, E. (2015). The "fundamental attribution error" is rational in an uncertain world. In D. C. Noelle et al. (Eds.), *Proceedings of the 37th Annual Meeting of the Cognitive Science Society* (pp. 2547–2552).

Webb, M. R., & Lee, M. D. (2004). Modeling individual differences in category learning. In K. Forbus, D. Gentner, & T. Regier (Eds.), *Proceedings of the 26th Annual Meeting of the Cognitive Science Society* (pp. 1440–1445).

Wilson, D., & Sperber, D. (2004). Relevance theory. In L. R. Horn & G. Ward (Eds.), *Handbook of pragmatics* (pp. 607–632). Oxford: Blackwell Publishing.

Wolpert, D. H. (2008). Physical limits of inference. *Physica D: Nonlinear Phenomena*, *237*(9), 1257–1281.

Xie, B., Hayes, B. K., & Navarro, D. J. (2018). Adding types, but not tokens, affects the breadth of property induction. In C. Kalish, M. Rau, J. Zhu, & T. Rogers (Eds.), *Proceedings of the 40th Annual Meeting of the Cognitive Science Society* (pp. 1199–1204).

Xu, F., & Tenenbaum, J. B. (2007a). Sensitivity to sampling in Bayesian word learning. *Developmental Science*, *10*(3), 288–297.

Xu, F., & Tenenbaum, J. B. (2007b). Word learning as Bayesian inference. *Psychological Review*, *114*(2), 245–272.