RESEARCH STATEMENT TOM GRIFFITHS

The goal of my research is understanding the computational basis of human cognition: identifying the computational problems people face, working out how they can be solved, and using those solutions to explain human behavior. Pursuing this goal requires combining the empirical methods of cognitive psychology with the tools and concepts of computer science, particularly artificial intelligence and machine learning. My aim is to exploit the interplay between these disciplines to produce both a deeper understanding of how the mind works and new ideas that can be used to develop intelligent computing machines. Artificial intelligence and machine learning provide a powerful set of tools that can be used to develop new psychological theories, and human cognition is a rich source of challenging computational problems that often lead to new concepts relevant to computer science.

The current emphasis of my work is the question of how people solve inductive problems, in which underdetermined conclusions are drawn from limited evidence. Many of the central problems in cognitive science, such as language acquisition and categorization, are problems of this kind. I have focused on everyday inductive leaps – inductive inferences that people make rapidly and unconsciously in the course of everyday life, such as learning causal relationships, predicting the future, and deciding whether something is the work of chance. These everyday inductive leaps provide the opportunity to study how people solve inductive problems in microcosm. The key challenge in explaining these inferences is accounting for how people are able to learn so much from so little: people can infer causal relationships from a handful of datapoints, make accurate predictions from a single observation, and assess the randomness of a binary sequence only half a dozen bits long. These inferences cannot be explained in terms of traditional statistical methods or the generic algorithms used in computer science, which require large amounts of data to reach similar conclusions.

The outcome of any inductive inference is the result of combining the available data with prior knowledge. People are able to draw conclusions from limited data by exploiting strong prior knowledge. In many everyday inductive leaps, and in other inductive problems that arise in higher-level cognition, this prior knowledge takes the form of a rich, structured theory of how the world works. Attempting to explain these inferences thus establishes two projects: a psychological project of determining the content of this prior knowledge in specific settings and elucidating its general structure, and a computational project of developing methods for combining structured representations with statistical learning. The nature of these two projects can be understood via an analogy to linguistics, a discipline that has been extensively concerned with inductive problems involving structured representations. The role of prior knowledge in everyday inductive leaps is similar to the role of a grammar in sentence comprehension, allowing people to infer the structure that was responsible for generating data. Under this analogy, the psychological project is the analogue of characterizing the grammars of different languages and attempting to formalize their common structure. The computational project explores questions more similar to those explored in computational linguistics: understanding how these representations can be used to parse data, and how they might be learned.

I approach these projects using the tools of Bayesian statistics. This formalism provides the foundation of much work in contemporary artificial intelligence and machine learning, and is beginning to be used in modeling cognition. In the remainder of this research statement,

I will outline how I have applied this approach to the projects described above.

Psychological project: prior knowledge in inductive inferences

At the heart of Bayesian statistics is Bayes' rule, which provides a rational prescription for how prior knowledge should be combined with data to yield new beliefs. Modeling human inferences as Bayesian inferences provides a direct mapping from assumptions about prior knowledge to predictions about human behavior, making it possible to explore the implications of such assumptions directly. I have used this approach to study how people solve a variety of inductive problems, with a particular focus on causal induction.

Causal induction.

Inferring causal relationships from data is central to the growth of both scientific and everyday knowledge, whether it is done by a scientist inspecting a contingency table or a child observing the interaction between two objects. The computational problem of causal induction can be formulated using causal graphical models, a language for representing, reasoning, and learning about causal relationships that has been developed in computer science and statistics. In this language, the problem of inferring causal structure from data can be expressed as the problem of identifying the causal graphical model most likely to have generated those data. A number of generic algorithms exist for solving this problem. However, people are capable of inferring causal relationships from far less data than is required by these algorithms.

In collaboration with Josh Tenenbaum [6,9,16,22,28][1], I have developed a formal framework for modeling human causal induction based upon the idea that people's inferences are guided by causal theories. In this framework, a theory acts as a hypothesis-space generator: it generates hypotheses about the causal structure underlying a set of events in the same way that a grammar generates hypotheses about the syntactic structure underlying a sentence. These hypotheses are then evaluated by Bayesian inference. This framework allows theories to impose strong constraints on the causal structures considered by a learner, making it possible to infer causal relationships from only small amounts of data. The key idea behind this approach – using structured representations to define probability distributions over graphical models – shares a great deal with current research in the artificial intelligence community on causal learning and probabilistic logic. We are beginning to explore the question of how the theories themselves can be learned [26,27], and I am in the process of applying this approach to the domain of physical causality, with the aim of developing a computational account of folk physics.

I have also used this approach to investigate some "cognitive illusions" – phenomena that psychologists and statisticians traditionally consider manifestations of human irrationality, such as our intuitions about coincidences [8]. Just as optical illusions reveal the implicit assumptions made by the visual system, cognitive illusions reveal the implicit assumptions that guide inductive inference. In both cases, we can gain insight by assuming that the system is rational and then asking what set of assumptions would produce the illusions. I have argued that a coincidence can be defined as an event that provides evidence for a hypothesis, but not enough evidence to convince us that the hypothesis is true. This definition can be

[1] Numbers in brackets refer to items in the publication list of my curriculum vitae.

formalized in the language of Bayesian statistics, and provides accurate predictions of how people interpret events, as well as the strength of coincidences. I am currently using my work on coincidences as the basis of an account of what makes a good magic trick. Magic tricks are often used in experiments that explore the ontological commitments of infants and children. By making the connection between prior knowledge and magic precise, I hope to obtain a tool that can be used to assess subtle details of the ontological commitments of adults.

Other everyday inductive leaps.

While causal induction plays a central role in the acquisition of knowledge about the world, people make inductive leaps in a variety of other contexts. I have also investigated the computational basis of similarity judgments [2,3], our sense that an outcome is "representative" of a set or process [10], the inference that something was produced by a random process [8,13,17], and the human capacity for predicting the future [7]. In each of these cases, I tried to identify the underlying computational problem, and the assumptions that people make in solving that problem. For example, my work on randomness examines why people think a sequence of heads and tails like `HHTHT` is more random than `HHHHH`. Many psychologists consider this judgment irrational: for any sequence X of a particular length, the probability of that sequence under a random generating process, $P(X|\text{random})$ is the same. However, if we view the underlying computational problem not as one of evaluating the probability of the sequence, but of determining the process that produced that sequence (which is arguably a more important kind of inference) then people should actually be computing a different probability: the probability that a random generating process produced this sequence, $P(\text{random}|X)$. This probability can be computed using Bayes' rule, but requires making assumptions about the alternative generating processes that could have been involved. If the alternatives include all computable probability distributions, then we obtain a measure of randomness called Kolmogorov complexity, which is used in theoretical computer science. By exploring different subsets of these alternative processes, we can begin to map out the kinds of regularities to which people are sensitive, and the kinds of computations required to identify those regularities (see [17] for more details).

Computational project: combining structure and statistics

Human cognition poses a fundamental challenge for artificial intelligence, machine learning and statistics: developing probabilistic models that are rich enough to express the kind of complex structures that people can represent, yet are sufficiently constrained to be inferred from data. This challenge has led me to work on structured probabilistic models for language and non-parametric Bayesian models.

Structured probabilistic models for language.

Linguistic corpora provide a rich source of data that can be used to develop models that combine structure and statistics. Many statistical approaches to language are piecemeal, using special-purpose methods to get at a particular aspect of linguistic structure. Probabilistic generative models provide a general-purpose approach to modeling language in which many different aspects of linguistic structure can be integrated. A probabilistic generative model specifies a simple stochastic procedure by which data are generated. This procedure

can involve unknown parameters or latent variables. Given the data, the values of those parameters and variables can be inferred via Bayes' rule. This approach is naturally applied to language, in which latent structure (meaning) is communicated via limited observable data (sentences).

While probabilistic models for syntax have been explored in some detail, developing probabilistic models for semantics and testing them against human data remains a significant open problem. The questions of how semantic information about words should be represented and how much of this information can be extracted from text make contact with other research areas that interest me, such as the structure and learning of intuitive theories. So far, I have worked with a very simple semantic model, developed in the information retrieval literature, in which words are represented using a set of "topics". Most accounts of language comprehension stress the importance of a working memory representation that predicts which concepts are likely to be relevant and disambiguates words. This model provides a simple means of solving these problems, as well as providing information about the relationships between words and the content of documents. In joint work with Mark Steyvers, I developed a simple algorithm that can be used to apply this model to large corpora [5], and showed that it provides a better account of human semantic memory data than previous approaches [12,15,31]. Working with Dave Blei and Josh Tenenbaum, we have used this model to show how the distinction between "function" and "content" words can be learned based upon a simple characterization of the different statistical dependencies that result from syntax and semantics [23].

The simplicity of this generative model and the algorithm used to apply it to data make it easy to extend (e.g. [20]). I am currently working on probabilistic models that use more sophisticated semantic representations (such as hierarchies [18] and predicate trees). The goal of this work is not necessarily better models of language, but to exploit the availability of linguistic data as a means of deriving prior distributions over structured representations that can be used in explaining cognition. I have also begun to develop Bayesian models that are relevant to the formal analysis of language evolution (and cultural evolution more generally).

Non-parametric Bayesian statistics.

A major theoretical problem for any form of learning is the construction of a space of hypotheses that is simultaneously rich enough to contain each of the infinitely many things that one might want to learn, yet sufficiently constrained to allow anything to be learned from finite data. This is accompanied by the practical problem of representing and reasoning about infinite combinatorial hypothesis spaces. Solving these two problems is a necessary part of developing statistical models that can capture the flexibility of human cognition.

A simple example of a case in which this problem arises is deciding how to cluster a set of animals into different species. In such a setting, it does not make sense to assume that there is a strict upper bound on the total number of species: while any finite number of animals can only come from a finite number of species, any new animal could come from a species that we have never seen before. We thus need to define a model that allows an infinite number of different species, yet allows us to reason about our finite set of animals without using infinite resources.

Problems of this kind arise whenever we want to use models in which we allow for the

possibility that new data could reveal new structure, whether that structure be clusters, hidden causes, trees, features, or spatial dimensions. One approach to solving these problems comes out of nonparametric Bayesian statistics, which provides methods for constructing tractable probability distributions over infinite combinatorial hypothesis spaces. In joint work with Charles Kemp and Josh Tenenbaum, I have applied these methods to the clustering problem described above (focusing on the case where clusters are defined by the relations between objects [27]). I have also worked on extending these methods to other kinds of structured representations: trees [18], and binary matrices [29]. These extensions each make it possible to define new kinds of statistical models, and bring us closer to being able to develop the kind of combination of structure and statistics that is needed to model human cognition.