Bayesian Computational Models for Inferring Preferences

by

Owain Rhys Evans


Submitted to the Department of Philosophy
in Partial Fulfillment of the Requirements for the Degree of
Doctor of Philosophy
at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

September 2015

© 2015 Massachusetts Institute of Technology. All rights reserved.
Bayesian Computational Models for Inferring Preferences
Owain Evans

submitted to the Department of Philosophy on August 26, 2015,
in partial fulfillment of the requirements for the degree of Doctor of Philosophy

Abstract
This thesis is about learning the preferences of humans from observations of their choices. It builds on work in economics and decision theory (e.g. utility theory, revealed preference, utilities over bundles), Machine Learning (inverse reinforcement learning), and cognitive science (theory of mind and inverse planning).

Chapter 1 lays the conceptual groundwork for the thesis and introduces key challenges for learning preferences that motivate chapters 2 and 3. I adopt a technical definition of ‘preference’ that is appropriate for inferring preferences from choices. I consider what class of objects preferences should be defined over. I discuss the distinction between actual preferences and informed preferences and the distinction between basic/intrinsic and derived/instrumental preferences.

Chapter 2 focuses on the challenge of human ‘suboptimality’. A person’s choices are a function of their beliefs and plans, as well as their preferences. If they have inaccurate beliefs or make inefficient plans, then it will generally be more difficult to infer their preferences from choices. It is also more difficult if some of their beliefs might be inaccurate and some of their plans might be inefficient.

I develop models for learning the preferences of agents subject to false beliefs and to time inconsistency. I use probabilistic programming to provide a concise, extendable implementation of preference inference for suboptimal agents. Agents performing suboptimal sequential planning are represented as functional programs.

Chapter 3 considers how preferences vary under different combinations (or ‘compositions’) of outcomes. I use simple mathematical functional forms to model composition. These forms are standard in microeconomics, where the outcomes in question are quantities of goods or services. These goods may provide the same purpose (and be substitutes for one another). Alternatively, they may combine together to perform some useful function (as with complements). I implement Bayesian inference for learning the preferences of agents making choices between different combinations of goods. I compare this procedure to empirical data for two different applications.

Thesis supervisor: Roger White
Title: Professor of Philosophy, Massachusetts Institute of Technology.
Collaborations

Chapter 2 is based on joint work in collaboration with Andreas Stuhlmueller and Daniel Hawthorne. Chapter 3 is based on two projects. One (which I acknowledge in the text) was in collaboration with Leon Bergen and Josh Tenenbaum. The other (on the Knobe effect) was in collaboration with Josh Tenenbaum. All mistakes are my own.
Bayesian Computational Models for Inferring Preferences

Owain Evans

Long Abstract

This thesis is about learning the preferences of humans by observing their choices. These choices might be in the context of psychology experiments in a lab, where the decision problems that people face are simple and controlled. They could be choices in some real-world task, such as navigating a city or building a website. Preferences inferred from observed choices are sometimes called ‘revealed preferences’ in economics. A related idea appears in the field of Machine Learning, under the label ‘Inverse Reinforcement Learning’ (Ng and Russell 2000), and also in Cognitive Science, where it’s called ‘Inverse-Planning’ (Baker and Tenenbaum 2014).

Inferring things from observed choices is not the only way to learn about someone’s preferences. It’s possible that in daily life we learn more about preferences from verbal testimony than from observed choices. Nevertheless, this thesis will only consider inference of preferences from choices¹.

I treat the problem of learning preferences from choices as an inductive, statistical inference problem. One goal of this work is to develop algorithms for automatically learning an agent’s preferences from data that records their choices. However, my principal concern in this thesis is with questions that are prior to building such algorithms. Two of these major questions, which are explored in Chapters 2 and 3, are the following:

¹ One can construe verbal testimony as a choice (much like construing speech as an ‘action’) and treat it in the framework I develop in this thesis. This is beyond the scope of this work.
(1) What is the causal and evidential connection between people's preferences and their choices? If people do not always choose what they prefer, what kind of evidence do their choices provide about their preferences? Can we learn about their preferences even if people don’t reliably choose based on them?

(2) How should human preferences be formally represented so as to facilitate inference from choices and to capture the full range and structure of human preferences?

Chapter 1 lays the conceptual groundwork for the rest of the thesis and introduces the key challenges for learning preferences that motivate chapters 2 and 3. Chapter 1 begins with a discussion of the concept of ‘preference’ itself. My focus is on what I call ‘economic preferences’, which are defined as ‘total subjective comparative evaluations’ of world histories. This definition aims to capture the notion of preference used in economics and decision theory. Since economic preferences are total evaluations, the question of whether \( A \) is preferred to \( B \) depends on a comparison across all evaluative dimensions that are relevant to the agent. So it’s not just a question of whether the agent likes \( A \) more than \( B \). It might also depend on whether \( A \) is morally better than \( B \), whether \( A \) satisfies an obligation or commitment (e.g. a promise or agreement) that \( B \) does not, and so on for other evaluative dimensions.

Having explained the notion of economic preferences, I discuss two distinctions that are important in the rest of this thesis. First, there is the distinction from the theory of well-being between actual preferences and informed or enlightened preferences (Railton 1986, Sobel 1994). Actual preferences have a closer link to observed choices but a weaker link to what is good for the agent. While much more difficult to infer from choices, I argue that enlightened preferences should be the ultimate target of the project of preference inference. The second distinction is between basic preferences and instrumental or derived preferences. Some world states are preferred for their intrinsic or immediate properties (this is a basic preference), while others are preferred only because they cause the former kind of state (a derived preference). It is often difficult to infer from observed choices whether a preference is

---

2 This analysis of economic preferences is taken wholesale from Hausman (2012).
derived or basic. Yet basic preferences will provide more concise and robust explanations of behavior. They should generally be the aim of preference inference.

Having made some conceptual distinctions about preferences, I explain why the problem of learning preferences is important. I consider its relevance to the social sciences, to normative ethics and value theory, and to cognitive science.

The final section of Chapter 1 introduces two general challenges for learning preferences from observations of human choices. The first challenge is the problem of human 'suboptimality'. An idealized agent has complete knowledge of the world and computes optimal plans to realize its preferences. The problem of inferring the preferences of such an agent has been studied in economics and Machine Learning. In real-world situations, humans are not omniscient and are not optimal in their planning. So existing formal techniques for inferring preferences don't straightforwardly carry over to humans. In Chapter 2, I develop models for learning the preferences of agents with false beliefs and for agents subject to time inconsistency. These models apply to sequential decision problems. I make use of recent advances in probabilistic programming to provide a concise, easily extendable implementation of preference inference for suboptimal agents.

The second challenge is the problem of finding formal representations for preferences that facilitate learning and are able to capture (as far as possible) the richness and complexity of human preferences. Human preferences seem to be 'sparse'. That is, most macro-level properties are irrelevant to whether people prefer an outcome. Few people care, for example, whether the number of blades of grass in their lawn is odd or even. It seems that preference learning would be easier and more efficient, if we could make some general assumptions about which properties are likely to matter for preferences.

A related question is how preferences vary under different combinations (or 'compositions') of properties. Suppose someone values a number of distinct and independent goods (e.g. friendship, art, community, hedonic pleasures). It's plausible that while these goods are independent, they are not perfectly substitutable. That is, having a large quantity of just one of these goods would not be preferred to having a

---

3 When I write 'complete knowledge', I always mean 'complete knowledge up to non-deterministic features of the world that are not possible to know'.

4 Time inconsistency has been used to explain behaviors like procrastination, temptation, and addiction (Ainslie 2001).
more equitable balance. In Chapter 3, I explore representations of preferences that use simple mathematical functions to model how different properties compose. These functions are standard utility functions in microeconomics, where the properties in question are quantities of goods or services that a consumer might purchase. These goods may serve the same purpose (and so be substitutes for one another), they may combine together to perform some useful function (as with complements), or they may be entirely independent. I describe and implement a Bayesian inference procedure for inferring the preferences of agents making choices between different combinations of goods. I compare this procedure to empirical data for two different applications.

---

5 I use the term 'goods'. This might sound like a discussion of 'objective goods' for an individual, rather than a discussion of the structure of someone's preferences. However, I'm just using 'goods' as a shorthand for 'a property or quality of a future state of affairs that someone prefers/values'.
Chapter 1:

Key Concepts and Challenges for Learning Preferences

1. Concepts and Distinctions for Preference Learning

1.1. Preferences as total subjective comparative evaluations

Throughout this thesis I do not use ‘preference’ in its everyday sense. Instead I use the notion of ‘preference’ found often in economic theory or decision theory. This notion is close to the everyday notion but differs in important ways. Hausman (2012) argues that preferences in both economic and everyday senses are ‘subjective comparative evaluations’ of states or outcomes. These evaluations involve multiple outcomes (in contrast to desires) and they involve multiple dimensions of evaluation. The key difference is that the economic notion of preference involves all dimensions of evaluation that the subject finds relevant. The everyday notion, in contrast, includes some but not all dimensions. So on Hausman’s terminology, economic preferences are ‘total subjective comparative evaluations’, while everyday preferences are ‘overall subjective comparative evaluations’.

The everyday notion of preference includes someone’s ‘likes’, their predilections, and their aesthetic evaluations. It typically excludes moral considerations, obligations,

---

6 I’m not suggesting that work in economic theory or decision theory always uses this notion of ‘preference’. Some work uses this notion but it’s often hard to tell exactly which notion is being used. See Hausman’s discussion of this in his (2012).

7 Everyday preferences are ‘overall evaluations’ because they involve comparing outcomes across a number of dimensions and then coming to an overall, holistic judgment. For example, a preference for one restaurant over another will take into account the food, the cost, the décor, the service, and so on.
and commitments. Someone might prefer (in the everyday sense) to drink wine at dinner but abstain because they promised their partner to cut down or because they agreed to drive people home. In this case their economic preference is not to drink wine (Hausman 2012). Economic preferences consider all relevant dimensions and weigh them against one another to reach a total evaluation.

Why work with economic preferences instead of everyday preferences? Here are two major motivations:

1. The economic notion of preference has a tighter connection to choice than the everyday notion. In domains where moral considerations and commitments are important, people's preferences (in the everyday sense) won't much influence their choices. The tighter connection between preference and choice makes prediction of choices from preferences more reliable. I discuss the importance of predicting choices below when explaining the relevance of preference learning to the social sciences and cognitive science.

2. The economic notion of preferences is more relevant to the task of helping someone by providing normative advice or normative recommendations. If you only learn about someone’s preferences in the everyday sense, you won’t provide useful advice in situations where the moral stakes are high. By definition, economic preferences include all dimensions of evaluation relevant to the agent. If you learn someone’s economic preferences, you can provide advice that includes everything the subject deems relevant to their making a decision. For this reason, someone's economic preferences are similar to their ‘values’ or to ‘everything they consider important’. The difference is that economic preferences encompass not only the list of evaluative dimensions that are important to someone but also the precise way in which these dimensions are weighted or traded-off in different contexts.

Using a notion of preferences that combines or ‘collapses’ different evaluative dimensions does not imply that distinctions among these dimensions are unimportant. On the contrary, it’s plausible that there are important philosophical and

---

8 Advice based on economic preferences may not be the best advice from an objective moral standpoint. For one, some individuals may give no weight to moral considerations in their preferences.
psychological differences between moral constraints or commitments and someone’s personal tastes or predilections. However, it can be hard to determine (a) whether someone’s choice was influenced by moral considerations or by a personal preference and (b) if both influences were present how those influences interact/compete to result in the choice. In this thesis the focus is on learning preferences from observed choices. Observed choices typically do not provide evidence about whether the evaluations underlying the choice were moral or were due to personal preference or predilection. So while the distinctions between evaluative dimensions are important, it won’t be something that the preference learning ideas I explore here can directly address.

It’s important to note that the economic notion of preference does not collapse all subjective influences on choice. I discuss below the fact that choices are influenced by cognitive biases such as Loss Aversion and hyperbolic discounting. In Chapter 2, I consider ways to learn preferences from observed choices without collapsing these biases into the preferences. Building on the ideas in Chapter 2, future work could try to infer preferences while avoiding collapsing different evaluative dimensions (e.g. avoid collapsing together someone’s likes and someone’s moral concerns)⁹.

1.2. The Objects of Preferences

If economic preferences are ‘total evaluations’, what are they evaluations of? In everyday speech, we sometimes talk about preferences over types of goods or over types of activities. For example, one might prefer Coke to Pepsi or prefer playing tennis to hiking. We also talk of preferences over tokens of these types. For example, one might prefer to play tennis this afternoon (but not necessarily on other occasions). I discuss preferences over things like goods and activities in Chapter 3. In this chapter I will not consider preferences over goods or activities (in either the type and token sense). Instead I consider preferences to be over more ‘fine-grained’ or ‘specific’ entities, namely, complete world histories (or ‘possible worlds’).

---

⁹ The idea is to make some structural assumptions about tastes and moral constraints and how they influence choice. If there is some difference in how they influence choice, you might be able to distinguish them from observing choices alone. One example is that people will talk differently about tastes and moral constraints. Another is that moral constraints might be more consistent across individuals and across time. People might also be less likely to make mistakes or be subject to biases for some dimensions rather than others, and so on.
To make it clear what I mean by 'preferences over world histories', I will make some terminological distinctions. In this chapter I use the term 'world state' to mean the state of the world at a particular time. I use 'world history' to mean the sequence of all world states, i.e. the set of all facts about a world throughout time as well as space. When rational agents are choosing actions, they evaluate the world histories that would result from taking each possible action. (Note that the world history includes the past and future world states that result from taking an action. In some cases, the agents will only consider the future consequences of an action when evaluating it.\(^\text{10}\))

The motivation for defining preferences over world states is that it leads to a simpler connection between choices and preferences. If preferences were over token goods or activities, there would be a more complex relation between preferences and choices. This is because rational agents make choices based on the ensuing complete world history, rather than on the good or activity that ensues in the short-term. For example, someone might enjoy drinking wine more than water but choose water because of the likely consequences of drinking wine. These consequences could be short-term (e.g. a hangover) or long-term (e.g. acquiring a social reputation as a drinker or adverse long-term health effects).

The assumption that people consider the entire future resulting from each action is unrealistic. People have cognitive/computational limits. Moreover, it's clear that many properties of a world history will be irrelevant to most people's preferences. For example, the choice of saying one word vs. another word will influence the movement of a particular oxygen molecule in the room. But no one would care about this molecule's movements in itself. I will consider each of these issues later in this chapter.

A related question is whether we should consider world states to be the primary objects of preferences, as opposed to world histories. Some would claim that people's preferences on world histories are a conceptually and computationally simple function of their preferences on world states (e.g. a discounted sum of world-state utilities)\(^\text{11}\). I

\(^{10}\) In the models in Chapter 2, the agents only take into account the value of the future resulting from an action. For some purposes (e.g. maintaining pre-commitments) it might be important to also take into account the past.

\(^{11}\) This claim is made both by economists and computer scientists (Schwarz 2009). People who are hedonists about preferences (i.e. who think that people's only basic or fundamental preferences are for pleasure over pain) might also have this view. If one is modeling preferences in a narrow domain, it
will not address this claim here. In this chapter I will stick to world histories, which are the more fine-grained object. If it turns out human preferences can be fully characterized in terms of preferences on world states, then this will make little difference to the points I raise here.

1.3. Actual vs. enlightened preferences

Consider the following example:

Fred is offered the choice between a cup of ice water and a cup of hot tea. Because it’s hot and humid, Fred would like the water. Unbeknownst to Fred, the water is poisoned and would kill Fred if he drank it. There is no poison in the tea.

This kind of example has been used in philosophical discussions of welfare or well-being to illustrate the distinction between 'actual' and 'informed'/enlightened' preferences. The idea is that Fred's actual preference for the cup of water differs from the preference he’d have if he knew the water was poisoned (his 'informed' preference). According to the notion of preference I've introduced, Fred does not have an actual preference for the cup of water. I define preferences over world histories, rather than objects like 'the cup of water' or the state of drinking the water. The relevant world histories are:

1. 'feel refreshed (from water), alive'
2. 'not refreshed (due to drinking tea), alive'
3. 'feel refreshed, dead from poison'

Fred prefers the first and second histories to the third one. Unfortunately for Fred, the first (and most preferred) outcome is not a realistic outcome given his choice.

might be plausible that this assumption holds. And even if the assumption is not plausible, it might be a modeling assumption that simplifies computation.

12 I only include the first part of these world histories and not the entire future extending beyond what happens to Fred. We can assume the histories don't differ in important ways after the events in the first part.
On my notion of preference, Fred does not prefer the cup of water. However he does plan to get the water and does believe this will cause a good outcome. This is relevant to tightness of the connection between preference and choice. Fred chooses the water despite its causing a world history he prefers least. The virtue of my view is that preferences are robust over time. Fred's preference wouldn't change on learning about the poison. It's plausible that many preferences (e.g. for survival, bodily integrity and basic comforts, or for social interaction) are much more robust over time than beliefs about which choices will help realize these preferences.

I've just argued that if preferences are over world histories, there is no difference between actual and informed/enlightened preferences in the example of Fred. This argument did not depend on Fred being logically omniscient or having any unrealistic computational powers. If we asked Fred about his preferences over outcomes (1) – (3) above, he would rank the outcomes with no difficulty. Moreover, we could read his preferences off his choices if he didn't have false beliefs. So even though Fred's preferences are defined over complex, fine-grained objects (i.e. entire world histories), there is not a practical difficulty here in learning those preferences.

While the actual vs. informed/enlightened distinction isn't applicable in Fred's case, it is still applicable to my account of preferences. It's often hard to say which of two world histories is preferable. This is the case in moral dilemmas or in the kind of examples used to argue for incommensurability or incomparability (Chang 1998). In these examples the difficulty seems to lie in trying to compare different dimensions of value. In other examples where comparison is difficult, the problem is not just multiple dimensions of value but also the cognitive complexity of the comparison.

I will illustrate the distinction between actual and enlightened preferences with an example. Suppose Bill, a teenager, considers moving to a foreign country permanently vs. staying in his home country. Possible futures resulting from each choice will differ in many ways. If Bill were able to see each future laid out in full detail (e.g. in an exhaustive, novelistic description of his possible life stories), it might be very hard to choose between them. Making this choice might require lots of deliberation and empirical investigation (e.g. trying out different experiences or hearing the testimony of other people). When it's hard to decide between two world histories, people's judgments will vary over time as a function of their deliberation and inquiry. In such cases we can distinguish between the person's actual preferences
and their enlightened preferences. The enlightened preferences are what the person would converge to under further deliberation and inquiry\textsuperscript{13}.

This actual/enlightened distinction is important for the problem of learning preferences from observed choices. Suppose we observe someone making a choice and we know that they have no false beliefs and no failings of rationality. Yet for the reasons given above, we also know that they might revise the actual preference implied by this choice. How likely are they to revise it? Answering this would require having a model of (a) how likely they are to deliberate and investigate this preference, and (b) how likely they are to change their mind if they do deliberate on it. I won't discuss such models here. Incorporating them into my current framework is a challenge for future work.

2. Why preferences? Why learning preferences?

The previous section introduced the idea of economic preference and made some basic distinctions about preferences. Why should we care about economic preferences? And why care about learning them?

2.1. Social Science

The social sciences aim to predict and explain human behavior that is relevant to social, political and economic entities. People's behavior is a function of many things. It depends on their beliefs, their abilities to plan and to strategize, and on constraints imposed by law and convention. Individual preferences are a crucial determinant of behavior. Suppose, for example, we want to predict how people will behave during a major recession. This will depend on their preferences for moving to a new city, their preferences for being employed (even at lower wages), their preferences for consumption, and the preferences of employers to hold on to employees (even if times are bad), as well as many other more specific preferences.

\textsuperscript{13} One might argue that the actual preferences are likely to be incoherent. So they won't have the formal structure that the enlightened preferences would have. I won't pursue this here.
Why care about learning preferences? As a thought experiment, we can imagine humans having homogeneous preferences that researchers could write down by doing some armchair introspection. It could be that humans prefer only a certain simple-to-identify pleasure state and all human behavior is the pursuit of this state. Alternatively, it could be that human preferences are easy to deduce from our evolutionary history. In actual fact, it seems that human preferences are complex. People seem to care about a wide array of apparently distinct things. Preferences are also heterogeneous across times and places and across individuals. So techniques for learning preferences have two key roles. First, they are needed to learn the complex structure constituting human preferences (some of which is probably common to nearly all humans). Second, they are needed to learn particular preferences of each individual.

This argument for the importance of preference learning doesn't explain my choice of focusing on learning from observed human choices. An obvious alternative way of learning preferences is to give people a big survey about their preferences. While surveys are valuable in learning about preferences, I want to highlight some advantages of learning from actual choices. First, people may intentionally misreport some of their preferences (e.g. to conceal unpopular preferences). Second, people may be bad at accurately reporting their enlightened preferences on a survey. If people actually face a choice between outcomes, they are likely to think carefully about each option, to consult wise friends, and to gather any empirical evidence that is relevant. A well-constructed survey might provoke some of this thinking but it's unlikely to match the real-life decision.

2.2. Normative Ethics and Value Theory

One theory of individual welfare/well-being is that someone's well-being is a function of whether their preferences are satisfied (Parfit 1984). I will call this the 'preference-satisfaction' theory. Versions of this theory do not always construe preferences as economic preferences that include all relevant evaluative dimensions between outcomes. Nevertheless, whether an agent's economic preferences are satisfied is likely to be highly relevant to their well-being on such theories.
There are various objections to preference-satisfaction theories. One objection is that a person preferring an outcome is insufficient for the outcome to confer well-being (Hausman 2012). The person also has to find the outcome rewarding or gratifying in some way. Another objection focuses on the problem (discussed above) that people can have actual preferences for outcomes they will later not prefer. Even if these objections undermine the claim that preferences are all that matters for individual well-being, it is still plausible that satisfying one's preferences is an important part of well-being.

If we assume that preferences are an important component of individual well-being, then the problem of learning individual preferences will be important both practically and philosophically. On the practical side, making people's lives better will depend on learning their preferences. For the reasons discussed in the section above on social science, there are good grounds for learning preferences from observed choices and not just from surveys or from introspection on one's own preferences.

On the philosophical side, if preferences are important for well-being, we need to understand better the content of those preferences. This means not just learning preferences in order to predict future choices but developing ways to represent the learned preference in some comprehensible form. It's plausible that we already have a good understanding of some important preferences (e.g. for survival and for good health). But some preferences may be more subtle. For instance, most people have a preference for having children. Some of these people don't end up having children and fertility rates in many rich countries are much lower than in the past. People's preferences here don't seem easy to capture in a simple formula or simple logical principle.

Moving on from the preference-satisfaction theory, there are other ways in which preferences are relevant to value theory and normative ethics. First, consider the importance of respecting people's freedom of choice and autonomy in modern societies. If there was a drug that made people less likely to commit violent crimes, many philosophers would judge it wrong for a government to secretly put the drug in

---

14 In principle there could be an algorithm that infers preferences and represents them in a form that is very hard for us to interpret.

15 People don't desire to maximize their number of children. People don't want exactly one or two children (otherwise there'd be no people who wanted children but didn't have them). Many only want children if the right partner comes along or if it won't get in the way too much of other projects. But how can we formalize these conditions? The claim is not that it's impossible but that it's a substantive task to formalize human preferences.
the water supply. This judgment would be stronger if the drug had negative side-effects – even if the benefits of the drug overall would outweigh the side-effects. More generally, governments need to obtain consent from individuals before taking actions that affect them in substantive ways. An individual will give consent if they prefer (in the economic-preference sense) the outcome. Individuals can be influenced in their judgments about what they prefer by evidence and argumentation: it's not just their immediate gut feeling that matters. But people's preferences will be an important determiner of whether moral or political ideas are able to influence actual behavior.  

2.3. Cognitive Science

As I noted in my discussion of social science (above), preferences are important for predicting human behavior. This makes the problems of learning and representing preferences of obvious relevance to cognitive science. Preference learning is also important in cognitive science because it's an important cognitive task that humans perform. People's ability to reason about other minds, often known as 'Theory of Mind', has been extensively studied in recent decades. It has been shown that Theory of Mind and preference inference develop very early, with sophisticated abilities observed before one year of age (Hamlin et al. 2007 and Hamlin et al. 2013). It is plausible that preference learning plays an important role in young children's ability to learn about the world during childhood. For adults, inferring the preferences of others is important in many spheres of life. It's important in social and family life, but also in doing deals (in business or politics), in designing new products or organizations, and in creating or understanding art.

Improved theories of the structure of human preferences and new techniques for learning human preferences can be applied to modeling Theory of Mind. In the other direction, we might better understand how to learn human preferences from studying how humans do it.

---
16 Note that this is analogous to the relevance of Pareto improvements in economics. Policies should be easier to bring about if no one is worse off from them. (One final way in which preferences relate to normative ethics: If people can be swayed by argument to economically prefer what is objectively good for them, then economic preferences will ultimately converge to preferences for objectively good things.)
3. Challenges for Learning Preferences from Observed Choices

The previous section motivates the general problem of learning preferences from observed choices. This section introduces two specific challenges that are part of this general problem. I address these challenges in greater detail in Chapters 2 and 3. In those chapters, I define and implement formal models that learn preferences in ways that attempt to deal with these challenges. This section discusses the challenges from a big-picture perspective.

3.1. False beliefs, suboptimal cognition and failures of rationality

3.1.1. Suboptimality and false beliefs

The choices of optimal, omniscient agents will generally be more informative about their preferences than those of suboptimal agents. Imagine an omniscient, optimal agent with human-like preferences is navigating a busy shopping district. (Optimal agents love shopping.) This agent knows about every product in every store, without having to do any browsing. So if it buys something, then it must have bought the best item relative to its preferences. Imagine now that a normal human is thrown into the same shopping district with no ability to see or hear (e.g. due to a blindfold) and having never visited the district before. If this person manages to stumble into a store, their doing so would not provide any information about their shopping preferences. This example shows that even if we (as observers of the person doing the choosing) had full knowledge of the person's epistemic state, we would not be able to infer anything about their shopping preferences from their choices.

There are similar problems with inferring preferences when we (as observers) are uncertain about the epistemic state of the agent choosing. If we don't know whether or not the person wears a blindfold, then knowing which store they visit will not allow us to identify their preferences—though it will provide some information. Consider
again the example of Fred and the poisoned water (above). If Fred chooses the hot tea, it's ambiguous whether he prefers the tea (and doesn't know about the poison) or prefers the water (and knows about the poison).

The examples of the blindfolded person and of Fred involve agents with uncertain or false beliefs. It's important to note that the difficulty in inferring their preferences is not that their choices are 'random' or 'noisy'. Random errors in someone's choices will make it more difficult to infer their preferences. But if the choices are repeated and are not completely random, errors will average out over time and preferences can still be inferred. By contrast, inaccurate beliefs can make it impossible to infer preferences - even if observations are repeated. Here's an example. Suppose Sue regularly eats candy containing gelatin. One possible inference is that Sue is happy to eat animal products. Another is that she doesn't know about the gelatin in the candy (or doesn't know that gelatin is an animal product). If Sue never receives any evidence about the gelatin, then we (as observers) can never distinguish between these two possibilities.

I have given examples that show how the agent's uncertainty or false beliefs can make preference inference more difficult. Other kinds of deviation from an ideally rational, omniscient agent will generate similar difficulties. One such deviation is suboptimal planning. For example, without computer assistance, people will plan travel routes suboptimally. The deviation from optimality will be bigger for a longer and more complex route. Suppose someone works out a route from $A$ to $B$ that is very inefficient. We observe part of their route. From this evidence, we might infer that they intend to go from $A$ to a different place $C$, for their route might be an efficient way to get to $C$. (Also note that if someone's ability to plan routes is completely non-existent, their choices would tell us nothing about their intended destination. This is like the example of the blindfolded person.)

3.1.2. Cognitive Biases

Suboptimal planning is a deviation from optimality but not one that humans could be blamed for. After all, many planning problems are known to be computationally intractable. Psychologists have argued that humans are also subject to an array of

---

17 Note that suboptimal planning is consistent with the agent knowing all the relevant empirical facts. Constructing the optimal plan is a computational task that might require elaborate calculations.
'cognitive biases'. These are deviations from ideal rationality that arise from a contingent shortcoming of our cognitive design -- analogous to our visual blind spot. I will give two examples of biases and discuss how they interact with preference inference.

**Loss Aversion**

You are offered a gamble where you get $200 if the coin comes up heads and you lose $100 if it comes up tails. Many Americans (for whom losing $100 is a small, inconsequential loss) reject this gamble, despite its expected value (Rabin 2000). Psychologists explain this as a specific aversion to outcomes that are framed or interpreted as a loss.

How does Loss Aversion interact with learning preferences? Studies show that not all subjects are equally loss averse. If we observed someone reject this gamble and didn't know whether they were subject to Loss Aversion, we would be uncertain about their preferences. They might reject it because losing $100 would have drastic consequences relative to winning $200. While this interpretation is unlikely to describe a typical US college professor, it is important that a general-purpose preference learning technique include this possibility.

Loss Aversion is described as a bias because it leads to 'preference reversals', i.e. to sets of preferences over outcomes that are inconsistent. It can't be explained away as a basic preference for avoiding loss over realizing gains.

**Time Inconsistency**

Time inconsistency (also called 'dynamic inconsistency') is one of the best-studied cognitive biases. This inconsistency arises when someone's desired choice between two options changes over time without them receiving any new pertinent information about the options. For example, someone might desire to get up early the night before and then reverse their desire when their alarm rings in the morning. One of the most influential formal models of time inconsistency is the 'hyperbolic discounting' model (Ainslie 2001). I will consider time inconsistency in much greater detail in Chapter 2.
I've noted that deviations from optimality and omniscience create difficulties for learning preferences from choices. The general problem is roughly that suboptimal agents will not always end up with outcomes that they prefer. This means that outcomes will either be uninformative about preferences or they will be highly ambiguous. One idea for getting around this problem is to try to learn preferences only from simple settings where human behavior is likely to be optimal. This could be a natural setting or it could be a lab setting which is specifically constructed to be simple. Simplified settings seem to be a valuable way of learning about preferences while avoiding the problems of suboptimality discussed here. There are, however, some shortcomings with this approach. First, it's often hard to communicate to subjects the precise structure of a very simple decision problem. Consider the Prisoner's Dilemma. Formally the game is defined by a pay-off matrix with four entries. Yet when subjects without a background in decision theory play against each other, they are likely to place some value on their opponent's utility and on equitable or 'fair' outcomes. This can make the 'cooperate' option more appealing than it is in payoff matrix.

A different, more fundamental problem is that some important human preferences may be hard to test in a simplified setting. For one example, consider again the Prisoner's Dilemma and also related games involving cooperation, coordination and conflicts between selfish and altruistic behavior. Even if the payoff matrix in these games is simple, the game is intrinsically complex because simulating the game involves simulating another human being. Moreover, it's not clear one could really model and understand human behavior in game-theoretic settings without including our ability to simulate others.

Another example of preferences that are hard to investigate in simplified settings are those involving processes unfolding over long periods of time and correspondingly large time investments. It might be hard to learn someone's preferences for friendships, family relationships, or for skills that take years to develop, in a lab experiment lasting only a few hours. Suppose it takes five years of training or experience to be in a position to evaluate two life options A and B. Then a lab experiment won't be able to do a simplified or controlled version of this training period and so won't be helpful in trying to learn about this preference.
I've argued that simplified settings are not sufficient to get around the problem of humans being suboptimal in their planning and being non-omniscient. In Chapter 2, I develop a different approach to learning preferences despite these deviations from optimality. This approach explicitly models the deviations, learns from observed behavior whether they apply to a given individual, and then uses this accurate model of an individual to infer their preferences. There are limitations to this approach also. You can only learn someone's deviations from optimality if there is sufficient behavioral evidence to do so. Moreover, if someone has very limited knowledge or planning ability (as in the case of the blindfolded person), then inferring the precise nature of their suboptimality will not help with inferring their preferences. So it seems likely that a hybrid approach to inferring preferences (e.g. combining inference in simplified settings with explicit modeling of suboptimality in the style of Chapter 2) will be needed.

3.2. Representing preferences

The second challenge with learning preferences that I explore in this thesis is the problem of how to formally represent human preferences.

3.2.1 Representing preferences as a series of high-level properties

In Section 1 of this chapter (above) I specified preferences as taking *world histories* as their objects. From the perspective of everyday human cognition, world histories are extremely fine-grained entities. Miniscule differences in the physical properties of single atoms (even for a very short period of time) anywhere in the universe suffice to differentiate world histories. Yet most humans, if actually living through these two world histories, would find them indistinguishable. Humans are typically aware of only very coarse, high-level or macro-level properties of a world history.
Moreover, human preferences seem to mostly concern a small subset of these high-level properties\textsuperscript{18}. The relevant properties concern pleasure and pain, good physical and mental health, rewarding artistic and leisure activities, family and friends, and so on for many more. While we might have some practical or scientific interest in how many grains of sand there are on a beach, or in whether a particular tree has more mass in its roots or its bark, few people will have a preference either way. I summarize this point by saying that preferences are 'sparse' over the set of useful macro-level properties of the world. That is, if we collect a long, diverse list of high/macro-level properties of world histories then most of these predicates would be irrelevant to the preferences of most people. It's not easy to spell out this 'sparsity' claim formally. The intuition is that if we took a non-gerrymandered and extensive list of macro-level properties that are useful in describing the world, then most of these properties would not be useful for characterizing human preferences. An example of a non-gerrymandered list would be all properties expressed by single words in a big dictionary or all properties used in Wikipedia articles.

This sparsity property is important when it comes to learning human preferences from observations. In most cases the two possible world histories resulting from a choice will differ in a large number of their high-level properties. For instance, suppose Mary is choosing between a holiday in Puerto Rico and one in Mexico\textsuperscript{19}. This choice clearly has consequences that are relevant to Mary's preferences. But the choice also has many consequences that Mary would not even think about. In Mexico Mary would minutely influence currency exchange rates and would also influence prices for Mexican goods and labor. She might kill some individual blades of grass by treading on them. She might bring over some pollen from the US, and so on. If we observe Mary choose to go to Mexico, then this could be because of a preference for any of these high-level properties on which the Mexico and Puerto Rico options differ. All we know is that one of these thousands of properties speaks in favor of Mexico. It could also be the interaction of different properties that is important (see below). If we assume that each of these different properties is equally important, we would scarcely be able to generalize having observed Mary's choice. Any similar choices in the future would themselves be between world histories which differ in

\textsuperscript{18} By 'properties' I mean any function of world histories, not just those expressed by one-place predicates. This can include arbitrary relations.

\textsuperscript{19} In practice, Mary would usually have lots of uncertainty about how the holidays would go. But for this example I assume (implausibly) that there is no such uncertainty.
thousands of mostly inconsequential ways from each other and from the Puerto Rico and Mexico world histories.

One obvious way around this problem of generalization is to limit the set of relevant properties from the beginning. That is, we just assume that most of the consequences of choosing Mexico as opposed to Puerto Rico are irrelevant and so we don't even consider them. This still leaves us the problem of which properties to include. Ideally, we would develop algorithms that include a large space of properties and winnow them down in response to data. It's also not obvious what properties to include in the initial 'large space of properties'. It's clear that we can ignore micro-physical properties for the most part (e.g. the acceleration of a random nitrogen molecule). But what kind of high-level properties do we need to include? There are two problems here. One is the problem of casting our net wide enough that we capture all relevant properties. The other is finding good formalizations of the properties that we are pretty sure should be included. Various feelings of pleasure, reward, pride, belonging, familiarity, novelty and so on are plausibly important to us. While we have familiar and frequently used words for these concepts, it's not clear that everyday language is good at precisely specifying the states that are valuable. It might be that we can communicate about them with each other because we have enough common experience of them to be able to use vague terms\textsuperscript{19}.

3.2.2. Basic/intrinsic preferences vs. derived/instrumental preferences

I've suggested that we can represent preferences by a long list of high-level properties of world histories. It's a theoretical possibility that this long list turns out to be unnecessary. For instance, it could turn out that all humans are in fact psychological hedonists and the only relevant properties are amounts of pleasure and pain. But it seems plausible that a long and diverse list of properties is needed\textsuperscript{21}.

\textsuperscript{19} So we might want to include some related concepts from philosophy, cognitive science or neuroscience that might help precisely formulate these states.

\textsuperscript{21} The claim of sparsity (above) is that most high-level properties useful for describing the world are not useful for characterizing human preferences. This is consistent with there being a fairly long and diverse list of properties that are needed to characterize human preferences.
One general issue is that there will be semantically distinct properties that agree or coincide across a large space of possible world histories. For example, consider the following properties:

- 'being in pain'
- 'having activation in pain circuits in the brain'
- 'having activation in pain receptors (which normally leads to feeling of pain)'
- 'having a cut on the part of the skin where the pain receptors are'

While some of these properties are semantically distinct, they would coincide across many world histories. This has important consequences for learning preferences. Suppose we (as observers) see someone choosing between options, where some of the options involve getting a cut and feeling pain. From this evidence alone, we cannot tell which of the list of properties above is relevant to determining the person's choice. It's intuitively clear that some of these properties are what people actually care about in a 'basic', 'primitive', or 'intrinsic' way, while the importance of the rest is only a consequence of their robust connection to those we care about intrinsically. For example, we don't think people care intrinsically about their pain receptors firing. But people do care about the experience of pain and about damage to their body.

This suggests a distinction between preferences that are 'intrinsic' or 'basic' and those that are 'instrumental' or 'derived'. In explaining someone's choice, basic preferences will give a stronger and deeper explanation. Derived preferences are contingent, and would not be present in all possible world histories.

---

22 This distinction, for a given individual, can be read off from individual's preferences on world histories. If someone doesn't have a basic preference for their pain receptors being activated then they won't care about this happening in strange world histories where it happens without being caused by body damage and without causing pain.

23 Note that basic/intrinsic preferences are not the same thing as intrinsic objective goods. If humans are able to converge on normative truths, then we'd hope that they end up having preferences for the intrinsic objective goods. But it's possible that humans have basic/intrinsic preferences for objectively bad things and that these preferences are very robust.
3.2.3. Composition of properties

I've considered representing preferences in terms of a long list of high-level properties of world histories. An individual's preferences would then be represented by a utility function from this list of properties to the real numbers. But what kind of function is this? I'll refer to this as the question of how properties compose to produce preferences or utilities.

The simplest functional form for a utility function would be a weighted sum of each property, where the property will have a value of '0' or '1' depending on whether it holds of the world history. We could also consider multiplicative functions, functions with 'max' or 'min' operators, and functions with terms with diminishing marginal utility, and so on. As I discuss in Chapter 3, such functions have been explored in microeconomics. From a theoretical standpoint, it is desirable to explore the most expressive class of functions. So we could consider all functions satisfying some smoothness property or else all functions satisfying computability or some tractability constraint.

The distinction (above) between basic and derived preferences relates to the issue of composition. The compositional structure discussed in microeconomics can sometimes be eliminated by considering only properties for which the agent has a basic, rather than derived, preference. Consider the following example. As chili is added to a dish it first gets better (i.e. tastier) and then gets worse (as the quantity of chili makes the dish too hot). One can plot the function for the quality of the dish as a function of amount of chili. But this function is not actually needed to represent someone's preferences. The basic preference here is for a certain richness of flavoring and for the overall gustatory experience. Varying the amount of chili is only relevant to the person's preferences in so far as it changes the gustatory experience. Whereas the chili had a non-monotonic effect on the desirability of a dish (more chili made the dish better then worse), the property of gustatory experience is monotonic in its desirability.

It's plausible that lots of rich compositional structure (like the non-monotonic influence of chili on a dish) that appears to be relevant to preferences can actually be eliminated by switching to basic preferences. This leaves open the question of whether there is any rich and interesting compositional structure between properties.
for which there are basic preferences. One candidate would be the desirability of a balance between distinct sources of value\(^{24}\). People might prefer a balance between success or satisfaction (a) in work, (b) in family life, and (c) in personal hobbies and leisure. That is, people would prefer attaining some decent standard in all three categories over huge success in just one of them. For this to be an aspect of their basic preferences, the preference for balance cannot arise from the causal effect of failure in one category on the others. Some minimal level of success in work (or in making money) might be a causal pre-requisite for success in other spheres of life. The question is whether a certain level of success in work intrinsically enhances the value of success in other areas\(^{25}\).

4. Conclusion

This chapter introduced the notion of economic preference that I use in this thesis. Economic preferences are total subjective comparative evaluations of world histories. Since economic preferences depend on all dimensions of normative evaluation of world histories, they are robustly relevant both to predicting human choices and to providing normative advice or recommendations about how someone should choose. Economic preferences that satisfy basic rationality constraints have been formalized and studied extensively in economics and decision theory. I make use of the utility-function representation of economic preferences in the models I develop for learning preferences.

In Section of 2 of this chapter, I considered two challenges for inferring preferences from observed choices. The first challenge is due to humans deviating from omniscience and from optimal rationality. Because of human suboptimality, human choices are less informative about preferences than they’d otherwise be. I discussed

\(^{24}\) Some version of this idea goes back to Aristotle's *Nichomachean Ethics*.

\(^{25}\) Here are some quite different examples. The same pain experience could be bad (in normal contexts) or good (for an exercise freak who grows to like the pain of training or for a masochist). Likewise, one might normally prefer to treat people with respect. But one might prefer to withhold respect for someone who doesn't deserve it. These are cases where the value of an event varies conditionally on the context.
some strategies for dealing with suboptimality and I will look at a different strategy in Chapter 2.

The second challenge I looked at was the questions of the form and representation of human preferences. This is a fundamental question for thinking about preferences in general. The question is, 'What is a precise description (in regimented natural language or a formal language) of human preferences or of the kinds of preferences that humans have?' I also considered the related question of which preferences are basic/intrinsic as opposed to derived/instrumental. These questions are important for inferring preferences from choices. If we rule out many kinds of preferences a priori (because they seem too implausible) we risk failing to capture some surprising and counter-intuitive feature of human preferences. On the other hand, the wider we cast our net (i.e. considering a very broad class of possible preferences) the more difficult we make learning. This problem is discussed in a more applied setting in Chapter 3.
Chapter 2:

Learning the Preferences of Ignorant, Inconsistent Agents

0. Overview

Chapter 1 gave an overview of conceptual issues related to learning preferences from observed choices. In this chapter, I introduce a Bayesian statistical model for inferring preferences and describe my implementation of this model using probabilistic programming. The model intends to deal with one of the challenges for preference inference that I raised in Chapter 1. This is the problem of 'suboptimality'.

I consider two kinds of suboptimality in this chapter. First, humans often have uncertain or inaccurate beliefs about the empirical world (e.g. believing that a restaurant is open when it's already closed). Second, humans are subject to time inconsistency (e.g. staying in bed after having planned to get up early). My approach is to define a model of decision making that includes both inaccurate beliefs (which can be updated on evidence) and time inconsistency (formalized as hyperbolic discounting). I show that it is possible to learn each of the following properties of an agent from observed choices:

1. whether an agent has inaccurate beliefs and the values of the credences or subjective probabilities,
2. whether the agent is time inconsistent and (if so) the value of the discount rate,
3. the agent’s preferences.

The learning approach is Bayesian joint inference (implemented via probabilistic programming), returning a posterior distribution on all parameters characterizing the agent and their preferences. I test the inference algorithm's performance on simple scenarios involving planning and choosing a restaurant. This performance is also compared to inferences from human subjects asked to make preference inferences from the same stimuli.
1. Introduction

1.1. Examples illustrating inaccurate beliefs and time inconsistency

In Chapter 1, I gave some examples of how uncertain or false beliefs can make it more difficult to infer someone's preferences. I will now consider an additional example and discuss in more detail its implications for inference. The example is chosen to match the kind of decision problem I'll focus on in this chapter.

*Example 1*

John is choosing between restaurants for lunch. There are two restaurants close by. The Noodle Bar is a five-minute walk to John's left and the Vegetarian Cafe is five minutes to his right. Each is open and has plenty of tables. John goes directly to the Vegetarian Cafe and eats there.

If John is fully informed about the restaurants, then it's a plausible inference that he prefers the Vegetarian Cafe. However, John might not know that the Noodle Bar even exists. He might think it's further away than it is. He might falsely believe that it is closed today or that it has no free tables. He might have accurate beliefs about the Noodle Bar but falsely believe the Vegetarian Cafe to be closer or less busy or better in the quality of its food than it actually is. He might know nothing about either restaurant and have chosen the Vegetarian Cafe by flipping a coin.

---

1 As I discuss below, it's also possible that he has a long-term preference for eating at the Noodle Bar but is close enough to the Vegetarian Cafe that he feels tempted to eat there.
In the terminology of Chapter 1, John chose the world history where he eats at the Vegetarian Cafe. But this choice is compatible with his preference being for the world history where he eats at the Noodle Bar. There are many plausible inaccurate beliefs that could have led John to pick an outcome he does not prefer. I now consider a similar example, intended to illustrate how the possibility of time inconsistency can make inferring preferences more difficult.

Example 2

John is choosing between restaurants for lunch. There are two options nearby. There is a Burger Joint seven minutes away and a Salad Bar another five minutes further down the street. John walks down the street and eats at the Burger Joint.

![Figure 2. John’s choice between restaurants in Example 2.](image)

Suppose we assume that John has full information about each restaurant. Can we infer that he prefers the Burger Joint? Here is an alternative explanation. John planned to go to the Salad Bar. In walking to the Salad Bar, he had walk right past the Burger Joint. From close proximity, he found the Burger Joint so tempting that he abandoned his plan and ate there instead. This kind of explanation is common in everyday reasoning about people's actions. It ascribes a temporal or ‘dynamic’ inconsistency to John. When making his plan, John prefers that he eat at the Salad Bar over the Burger Joint. But when John stands right in front of the Burger Joint he has the reverse preference. One explanation for the preference reversal is a change in John's knowledge. John might see how delicious the burgers look and change his mind on this basis. But I want to focus on the case where there is no new information of this kind. John might feel tempted on seeing the burgers but not because he gets any new information about them.

If this temptation explanation is right, should we say that John has a preference for the Burger Joint? We can imagine arguing either side. One side is that John’s real preference is for the Salad Bar and that he only went to the Burger Joint because of weakness of will. He planned to avoid eating there and he'll regret having eaten there
afterwards. On the other side, we argue that John's failure was in his planning to go to the Salad Bar when his true preference was for the Burger Joint. When John is right in front of the Burger Joint he sees sense and realizes that he should have planned to eat there all along.

I will not enter into this debate. Instead I assume John has longer-term and shorter-term preferences that differ. When considering his choice of restaurants from his starting point, John is not close to either restaurant and so he acts on his longer-term preferences. When John is outside the Burger Joint, a few minutes away from sinking his teeth into a burger, his shorter-term preference is for the burger and he acts on this (where some people would stick to their longer-term preferences). In this chapter, I'll show that both longer- and shorter-term preferences can be inferred from observed choices.

1.2. The Wider Significance of Inaccurate Beliefs and Time Inconsistency

The examples above show how inaccurate beliefs and time inconsistency can make choices less informative about preferences. I will now consider the wider significance of each of these deviations from optimality.

1.2.1. Inaccurate beliefs

Humans often have false beliefs about concrete everyday matters of fact (e.g. falsely believing a cafe is open). Many such false beliefs will be corrected by direct observation or by the testimony of other people. Beliefs are more likely to be corrected if they are directly relevant to an important choice. However, even if humans get evidence that conflicts with a false belief, they do not always abandon the belief. Superstitions of various kinds survive counter-evidence. Beliefs about emotionally charged topics (e.g. whether one is a bad person, whether one is still physically attractive despite one's advancing years, and so on) can also survive evidence.

Human choices also depend on beliefs about the causal and probabilistic structure of the environment. This includes everything from the fundamental physical laws of
the universe to the chance of winning a state lottery. Beliefs about causal and probabilistic structure are also crucial important for practical tasks. Consider the example of someone driving on an icy road. The driver will make use of a mostly implicit model of how the tires interact with the icy road\(^2\). The accuracy of this model will depend on the driver's experience with this particular car and with icy conditions in general.

On many technical questions about causal/probabilistic structure, laypeople will be agnostic and defer to authority. But on certain questions laypeople may oppose the scientific consensus\(^3\). Improving the accuracy of such beliefs is not just a question of getting more evidence. Sometimes both laypeople and scientists will need to gain new concepts in order to understand how existing evidence relates to a particular question. For example, understanding the expected value of playing different casino games depends on an understanding of probability theory.

I have described some ways in which human beliefs can be inaccurate. How are these ways relevant to learning human preferences from choices? As I discuss below, false beliefs about everyday matters of fact (like falsely believing a cafe is open) can be dealt with in some existing frameworks for inferring preferences. One challenge here is that beliefs vary across individuals. Different individuals will have different false beliefs and so inference models will have to infer an individual's inaccurate beliefs and their preferences simultaneously. If certain false beliefs are robust over time, it might not be possible to learn certain preferences with confidence (as choices based on the preferences could also result from false beliefs).

Inaccurate beliefs about the causal and probabilistic structure of the world will generally make inferring preferences even more difficult. In standard models of optimal planning (see 'Formal Model' section), a distinction is made between the causal structure of the world (which the agent knows fully up to non-determinism in the world) and contingent elements of the world state of which the agent might be ignorant. If we (as observers trying to learn preferences) know the exact false causal model that the agent is using, then inferring preferences may be possible. The hard

\(^2\) I'm not suggesting that the driver's mental model is something like a physics simulation used by the engineers who design the car. But the driver is able to make some estimates of how the car will respond to the brakes or to a sharp turn.

\(^3\) When people decide whether or not to have their child vaccinated, they invoke beliefs about the causal consequences of the vaccine and the base rates of the relevant diseases. These beliefs sometimes go against the scientific consensus.
case is where (1) the agent has a false causal model, and (2) we (as observers) don't know what the agent's model is. Unlike beliefs about contingent parts of the world state, a false causal model can lead deviations from optimality that are both large and counterfactually robust⁴. Moreover, the space of possible causal structures that an agent could falsely believe is large.

There are many historical examples of strange but persistent behavior that arguably stemmed from inaccurate causal models. This includes much of Western medicine up until the last two centuries and the treatment of the dead in many cultures (e.g. tombs filled with goods for the afterlife). Without much experience in anthropology or history, it is a non-trivial task to work out what some ancient behavioral practice is actually about. Though we have a decent intuitive idea of the preferences of ancient people, it's difficult to work out their causal models of the world.

1.2.2. Inconsistencies and cognitive biases

Example 2 (above) illustrated the phenomenon of time inconsistency in human preferences. Many behaviors have been explained in terms of an underlying time inconsistency. First, there are many examples that have a similar form to the tempting Burger Joint in Example 2. These examples involve some compelling activity with bad longer-term consequences, e.g. overspending on a credit card, taking drugs, gambling, staying out too late, getting violently angry, and so on. People plan to avoid this activity but abandon the plan at a later time. A second example is procrastination. Consider someone who procrastinates on writing a referee report and puts it off until it's too late. On day one, they prefer to do other things today and write the review tomorrow. When tomorrow comes around, their preference has changed and they now prefer to do other things again rather than write the review.

There is work trying to explain these apparent inconsistencies as resulting from purely rational choices (Becker 1976). However, one of the most influential psychological accounts of these phenomena (Ainslie 2001), which I draw in this

---

⁴ By counterfactually robust, I mean that even if the decision problem is repeated many times with small variations the suboptimal choices will persist.
chapter, argues that this behavior is not rational and provides formal models of such inconsistencies.

The behaviors above that involve time inconsistency (temptation, procrastination, addiction, impulsive violence) are just as familiar from everyday experience as false or inaccurate beliefs. By contrast, the literature in psychology and economics on 'cognitive biases' has uncovered many subtle deviations from optimal cognition. Kahneman and Tversky (1979) provided experimental evidence that people are subject to 'framing effects', where evaluations of outcomes are influenced by irrelevant details of how the outcomes are presented. They also argued that human inferences violate the probability axioms (e.g. see the Conjunction Fallacy) and that the basic human representation of probabilities is flawed (Prospect Theory).

There is a huge literature on biases and inconsistencies that influence choice and decision-making (Kahneman and Tversky 1974). Some of this work provides mathematical models of choice for agents subject to a bias (e.g. Prospect Theory or the Hyperbolic Discounting theory). However, there is little work on how to infer the preferences of people who are subject to these biases. The problem for inference is the same as in Example 2. If someone chooses $X$ over $Y$, this could be because of a preference for $X$ or because there is a bias that inflates their estimation of $X$ over $Y$ (or both). If a choice task has been studied extensively in the lab, we may be confident that a particular subject will display the typical bias. But in the general case, it is extremely difficult to say whether someone will be subject to a bias. Even in the lab, none of the prominent cognitive biases are universal. Outside the lab, people have a stronger incentive to avoid certain biases and they are able to consult experts before making decisions. So we will often be in the same problematic situation as in the case of inaccurate beliefs. That is, we know that some individuals will be subject to inconsistencies and biases. But we'll have to infer this from people's behavior at the same time as we are trying to infer their preferences. As with inaccurate beliefs, this sometimes makes it impossible to pin down someone's preferences. Yet in other cases, even a small number of choices can sometimes reveal both someone's biases and their preferences (as I will demonstrate below)\textsuperscript{5}.

\textsuperscript{5} Some will argue that certain biases (e.g. framing effects or temporal inconsistencies) undermine the idea that there are preferences that are stable over time and independent of context. I think this argument is important but I will not discuss it here.
2. Informal Description of a Model for Inferring Preferences

My goal in this chapter is to develop a statistical model for inferring preferences for agents with inaccurate beliefs and time inconsistency. The data for such inferences will come from choices on a decision problem. There are many different kinds of decision problem. Psychologists and economists studying preferences have often used decisions that are easy to specify formally and easy to run in a lab setting. These include choices between different lotteries (as in Prospect Theory) or simple two-player games such as the Prisoner's Dilemma, Rock-Paper-Scissors or the Dictator Game. In this chapter I work with decision problems that have a quite different flavor. These problems are not intrinsically non-deterministic (unlike lotteries) and they involve only a single player. They are complex because of the need to take a sequence of coordinated choices to achieve a good outcome. The sequential nature of these problems makes them analogous to many real-world problems and also provides a richer framework in which to explore a phenomenon like time inconsistency (which intrinsically involves a temporal sequence).

This section begins with background material on sequential decision problems. I then provide background on time-inconsistent planning and hyperbolic discounting. With these components in place, I give an informal description of my model for inferring preferences. The final subsection provides a formal description of the model and describes my implementation of the model.

2.1. Sequential Decision Theory

2.1.1. The Basic Idea of Sequential Decision Theory

The theory of optimal sequential action and choice, *sequential decision theory*, has been studied extensively in economics, computer science, statistics, and operations research (Russell and Norvig 2003, Ng and Russell 2000, Sutton and Barto 1998...
Bernardo and Smith (2000). Sequential decision theory is a generalization of one-shot decision theory to situations where an agent faces a series of choices instead of a single choice. Typically an agent's earlier choices and their outcomes influence future choices (as opposed to a series of independent one-shot problems). Moreover, the agent knows at the outset that they face a sequence of dependent choices and they can plan accordingly.

The mathematical objective in sequential decision theory is essentially the same as in the one-shot case. For each decision, the agent takes the action that maximizes expected utility of the resulting sequence of world states. The key difference is that the resulting sequence of world states will also depend on future actions the agent takes. I will give a formal description of this in the section 4.

2.1.2. Why consider SDT in preference inference

The main motivation for my considering sequential decision problems is that they allow a wider range of human choices to be formally modeled. In standard one-shot decision problems (e.g. lotteries, Prisoner's Dilemma, Rock-Paper-Scissors) the agent's choices lead directly or immediately to outcomes of varying expected utility. The outcome might be non-deterministic (as in the case of lotteries) but it doesn't depend on any subsequent actions of the agent. By contrast, most human choices will lead to utility only if the person makes appropriate choices in the future. Sequential decision-making is somewhat like a coordination game with one's future self. Each choice cannot be considered in isolation: it must be considered relative to how it facilitates future choices.

6 In philosophy I know of little work that focuses on sequential decision theory in particular. This is probably because many philosophical questions about decision theory can be illustrated by focusing on a one-shot decision problem. Consider, for example, the Representation Theorem for Expected Utility (von Neumann and Morgenstern 1947), the distinction between EDT and CDT (via Newcomb's Problem and related one-shot decision problems), questions of self-indication and indexicals (via Sleeping Beauty), and problems with large or infinite quantities (via Pascal's Wager, the St. Petersburg Paradox, etc.). One exception is the discussion of the iterated Prisoner's Dilemma. However such iterated games have a very simple sequential structure. Much of their complexity is already in the one-shot game itself.

7 In this chapter, I will usually talk about sequences of world states rather than world histories. Moreover, the utility functions I consider are defined primarily on world states rather than on world histories. This simplifies modeling and inferences but is inessential to the setup. Further work could explore the more general case of utility functions defined primarily on world histories.
Consider John's choice in Example 1 (see Fig. 3) of where to eat lunch. From a high-level planning perspective, John's choice is whether to eat at the Noodle Bar or the Vegetarian Cafe. But from a lower-level perspective he faces a sequence of choices. His first choice is whether to take the left or right fork in the road. There is no immediate benefit to John of taking the right over the left fork. John takes the right fork because it will put him in a location where subsequent choices will take him more efficiently to his restaurant of choice. (If John started by taking the left fork and then switched back to the right, his total path would be needlessly inefficient.)

It might seem odd to model this lower-level choice of whether to take the left or right fork. Given that planning the most efficient route is so trivial, why not just model the high-level choice of restaurant? For this simple problem, it would be fine to model the high-level choice. However, for slightly more complex problems, it becomes important to take into account the lower-level choices.

The first thing to note is that there is not always a 'most efficient' route. Suppose there are two routes to a restaurant. The shorter route is usually faster but is sometimes more congested than the other. In this case, a natural strategy is to try the shorter route and switch over if it's too congested. There's not an optimal route: there's an optimal policy (a function from states to actions) that's best in expectation. This policy wouldn't work if there was no way to switch routes. That is, the optimal strategy/policy here depends on which choices are available to the agent throughout the sequential problem. If the agent can switch routes midway, then it might be better for them to try the shorter route first. Otherwise, it might be best to always take the longer route (as it's better in expectation).

---

8 It's really more like a continuum of choices, because John could stop walking at any moment and walk the other way. But I'll imagine that space and time are discretized and so John has a finite series of choice points.
2.1.3. Sequential decision problems, inaccurate beliefs and time inconsistency

The difference between one-shot and sequential decision problems is significant in the case of problems where the agent has inaccurate beliefs. In a one-shot decision problem, an agent with inaccurate beliefs might overestimate the expected value of a choice \( X \) and choose it over \( Y \). Without further information, we (as observers) won't be able to distinguish this from a preference for \( X \). In the sequential case, the consequences of a false belief will often be protracted and easy to recognize from observations. A single false belief (if uncorrected) could lead the agent to a long sequence of actions that result in a very poor outcome. On the other hand, such a sequence of actions will often provide sufficient information to infer that the agent has a false belief. For example, if someone walks a very long way to a Starbucks (instead of going to the Starbucks nearby) we might infer both that they prefer Starbucks and that they didn't know about the closer Starbucks.

Sequential decision problems also help bring out the interesting properties of time-inconsistent planning. This will be discussed in detail in the section REF on 'time inconsistent planning', where I give a general introduction to time inconsistent planning.

2.1.4. The Restaurant-Choice Decision Problem

This section introduces the kind of sequential decision problem that I model in this chapter. The overall goal is to observe an agent playing this decision problem and then infer their preferences from these observations. The decision problem I focus on is an elaboration of the examples of John choosing a restaurant in Section 1. I will refer to this problem as the 'Restaurant-Choice problem'.

Consider the decision problem depicted in Figure 4a. The backstory is that John gets out of a meeting and needs to find somewhere to get some food. We assume that, all things being equal, he prefers walking less. The green trail indicates the path that John took. These are not real human actions choices: I created the paths based on what seemed to me a plausible sequence of actions. Rather than view John's path as a single choice (e.g. 'take the most efficient route to the Vegetarian Cafe'), I model John as making a series of discrete choices. One way of doing this is indicated by the grid
of Figure 4b. For each square on the grid, if John stands in that square then he has a choice of moving to any of the neighboring squares. Once John enters a square with a restaurant, we assume he eats at that restaurant and the decision problem ends.

Although I've described the Restaurant-Choice problem as a sequential decision problem, we could just as well call it a planning problem or a problem of instrumental ('means-ends') reasoning. John has some goal (initially unknown to the observer) of eating at his preferred restaurant but he is subject to the constraint that he'd prefer to walk as small as a distance as possible. The restaurants themselves have terminal value for John. The squares near to the restaurants have instrumental value because they are close to the states with terminal value.

![Figure 4a](left). Shows John's choice of restaurant (Veg Cafè) from his starting point in the bottom left. Fig 4b (right) shows a discrete 'grid' representation of the state space.

The sequential decision problem shown in Figure 4a is just one of a number of variant decision problems I consider in this chapter. To generate variants, I vary (1) John's starting point (which influences how close different restaurants are to him), (2) which restaurants are open, and (3) which streets are open or closed. I refer to each variant as a scenario. Figures 5a and 5b show two additional scenarios and John's path on those scenarios. Below I'll consider cases where an agent with different preferences than John plays the same sequence of scenarios.
It's clear that Restaurant-Choice problem abstracts away lots of the complexity of the real-life task of choosing somewhere to eat. I assumed that the streets themselves are homogeneous, which precludes preferences for safer or quieter routes. I also assumed that the agent has full knowledge of the street layout. While the Restaurant-Choice problem is minimalist, it does allow us to explore the implications of John having uncertain or false beliefs and of John being time inconsistent. I consider cases where (a) John does not know about a certain restaurant, and (b) John has a false belief (or uncertainty) about whether or not a certain restaurant is open or closed. I also consider different ways in which John takes actions in a time-inconsistent manner (see the following subsection).

The Restaurant-Choice problem I've described may seem quite specific to the case of navigating a city. In fact, the formal model of this problem I introduce in Section 3 is very general. The formal model consists of a set of discrete world states (which are the grid squares in Figure 4b). For each world state, there is a fixed, finite set of world states that are reachable from it (viz. the neighboring grid squares). The agent has preferences defined over these world states and also over movement between states. A sequence of world states is generated by physical/dynamical laws that update the world state given the agent's choice. In the literature on sequential decision problems, a large and diverse array of problems have been usefully encoded in this form. These include playing backgammon and other games, flying a helicopter, financial trading, and deciding which ads to show to web users.
2.2. Time Inconsistency and Hyperbolic Discounting

2.2.1. Hyperbolic Discounting in one-shot decisions

Having explained the Restaurant-Choice problem, I now introduce a standard mathematical model of time-inconsistent behavior and show how it leads to two distinct kinds of agent for sequential problems.

In Example 2 (above), I illustrated time inconsistency by considering how John might be tempted when walking past the Burger Joint on his way to the Salad Bar. A different kind of example (which has been extensively studied in lab experiments) concerns preferences over cash rewards at different times. Subjects in a psychology experiment are asked whether they'd prefer to receive $100 immediately or $110 the next day. They are also asked if they'd prefer $100 in 30 days over $110 in 31 days. Many more subjects prefer the immediate $100 than the $100 after 30 days. If people aren't able to 'lock-in' or 'pre-commit' to holding on till the 31st day for the $110, then this pair of preferences will lead to inconsistent behavior.

Similar experiments have been done with non-cash rewards (e.g. rewards of food, or negative rewards such as small electric shocks) and with animals, and the same inconsistency is frequently observed. The general pattern is that a 'smaller-sooner' reward is preferred to a 'larger-later' reward when it can be obtained quickly. But this preference reverses if each reward is shunted (by an equal amount) into the future.

There are various theories and formal models of time inconsistency. I'm going to focus on the model of hyperbolic discounting advocated by George Ainslie (2001). Ainslie explains time inconsistency in preferences in terms of a 'time preference' or 'discount function' that is applied whenever humans evaluate different outcomes. Humans prefer the same reward to come earlier rather than later: so the later reward is discounted as a function of time. If such discount functions were exponential, then this would not lead to dynamic inconsistency. However, the discount function for humans is better captured by a hyperbola, with a steep slope for very small delays and
a progressively shallower slope for longer delays\(^9\). Figure 6 illustrates the difference between exponential and hyperbolic discount curves.

![Discount factor vs. Time from present](image)

**Fig. 6.** Graphs show how much a reward is discounted as a function of when it is received.

Ainslie's claim is not that discounting only applies to cash rewards or to sensory rewards like pleasure or pain. The idea is that discounting would apply to any world state that humans find intrinsically (non-instrumentally) valuable. It would apply to socializing with friends or family, to enjoying art, and to learning and developing valuable skills\(^{10}\).

Hyperbolic discounting predicts the inconsistency I described where $100 is preferred over a delayed $110 when it's immediate (but not when it's 30 days in the

\(^9\) Various functions have these properties. The hyperbola is just a simple and analytically tractable example.

\(^{10}\) Using the terminology of Chapter 1, Ainslie gives a theory of how preferences over outcomes vary as a function of time. It's simplest to specify the theory if preferences are over world states (not world histories) and if they are represented by a utility function. The theory then has a simple compositional structure. There are timeless utilities over world states. (These capture what the agent would choose if all states could be reached at the same time). Then if one world state is received after another, it's utility is discounted by the appropriate amount.
future). Suppose we use the follow hyperbolic function to compute the discount factor (where \( \text{delay} \) is measured in days):

\[
\text{discount factor} = \frac{1}{1 + \text{delay}}
\]

Then at \( \text{delay}=0 \), the value of the options is 100 vs. 55. At \( \text{delay}=30 \), the values are 3.22 (for $100 at 30 days) vs. 3.42 (for $100 after 31 days).

2.2.2 Hyperbolic Discounting in Sequential Decision Problems

2.2.2.1. Naive vs. Sophisticated Agents

The example with cash rewards ($100 now vs. $110 tomorrow) illustrates the basic idea of hyperbolic discounting. It does not, however, demonstrate how hyperbolic discounting interacts with sequential decision-making. In the calculation above, I simply computed the discounted value of each of the outcomes. I didn't take into account the possibility of planning, which could allow one to pre-commit to waiting till the 31st day for the $110.

As I explained above, in a sequential decision problem an agent's choices need to take into account the choices they'll make in the future. So to take the best action now, the agent will benefit from having an accurate model of their future choices. (Analogously, each player in a coordination game benefits from an accurate model of the other players.) There are two natural ways for a time-inconsistent agent to model its future choices:

1. A \textit{Sophisticated} agent has a fully accurate model of its own future decisions. So if it \textit{would} take $100 on day 30, it will predict this on day 0.

2. A \textit{Naive} agent models its future self as evaluating options in the same way as its present self. So if it currently prefers to 'delay gratification' on the 30th day and wait for the $110, then it predicts (incorrectly) that its future self will do the same.

\[\text{References}\] See references in Kleinberg and Oren 2014 for more on this distinction.
Note that both Sophisticated and Naive agents will fail to delay gratification if they reach the 30th day and get to choose between the $100 now or the $110 later (Otherwise they wouldn't be time-inconsistent agents!). The Sophisticated agent has the advantage that it can foresee such behavior and potentially take steps to avoid ever being in the situation where it's tempted. This is the ability Odysseus displays when he ties himself to the mast. (Note that I'll sometimes talk about sophisticated planning. This is the planning a Sophisticated agent does, i.e. planning that takes into account accurate (rather than inaccurate) predictions about the agent's future actions. Similarly, naive planning is just planning a Naive agent does.)

2.2.2.2. Different Behavior of Naive and Sophisticated Agents

Naive and Sophisticated agents will behave differently on some of the decision problems that I model in this chapter. Consider the setup in Fig. 7a. The green line shows the path of an agent, John, through the streets. He moves along the most direct path to the Vegetarian Cafe but then goes left and ends up at the Donut Store. If he generally preferred donuts, he could've taken the shorter route to the other branch. Instead he took a path which is suboptimal regardless of his preferences.

Given the right balance of utilities for each restaurant, the behavior in Fig. 7a can be predicted by assuming that John is a Naive agent. Suppose that each restaurant has an 'immediate' utility (which John experiences immediately on eating the meal) and a 'delayed' utility (which John experiences after eating the meal). Suppose that for John donuts provide more immediate utility (or 'pleasure in the moment') than vegetables, but that vegetables provide significantly more delayed utility. (For instance, if John is trying to eat healthily, then he'll feel bad after eating the donuts and good after eating the vegetables). Now consider the choices that John faces. Early on in his path, John could turn left and go to Donut Store D1. At that point, D1 and the Vegetarian Cafe are each more than 5 minutes' walk away. From this distance, due to hyperbolic discounting, the greater immediate utility of the Donut Store doesn't lead it to win out overall. So John walks along the shortest route to the Vegetarian Cafe. This takes him right past the second Donut Store D2. Since he now gets very close to D2, it wins out over the Vegetarian Cafe. So John's being Naive can explain this behavior.
Figure 7a (left) is a path consistent with a Naive agent who prefers the Veg Cafe overall but gets high immediate utility from the Donut Store. Figure 7b (right) is a path consistent with a Sophisticated agent with the same preference structure. NB: You should assume that it takes around 10 minutes to walk the path in Fig 7a.

Suppose instead that John was a Sophisticated agent. Consider John’s decision of whether to (a) turn right onto the street that contains the Noodle Store, or (b) continue north towards Donut Store D2 and the Vegetarian Cafe. When standing at the intersection of the street with the Noodle Store, John will predict that if he went north he would end up at the Donut Store D2 (due to being tempted when he is very close to it). By contrast, if he goes right, he will avoid temptation and take a long route to the Vegetarian Cafe (assuming he doesn’t like the Noodle Store). This route is shown in Figure 7b. To get this behavior, John would have to get significantly more overall utility (i.e. delayed utility + immediate utility) from the Vegetarian Cafe than from the Donut Store D2. If the Donut Store is not much worse overall than the Vegetarian Cafe, then he’ll go to the Donut Store (since it’s closer). However, he wouldn’t go to Donut Store D2. He would instead realize that going north will not take him to the Vegetarian Cafe (due to temptation) and so instead of taking this inefficient route to a Donut Store, he’d take the more efficient route and go to D1.

Naive and Sophisticated agents behave differently on this decision problem. An agent without hyperbolic discounting behaves differently again. This agent either takes Donut Store D1 or will take the path shown in Fig 4a (above). More generally, the hyperbolically discounting agents produce patterns of behavior that are hard to
explain for an agent who doesn’t discount in this way. Fig 7a and 7b are both inefficient ways to reach a state. Without time inconsistency, ignorance, or some kind of suboptimal planning, agents will not reach goals inefficiently. Since hyperbolically discounting agents produce distinct behavior from time-consistent agents, we can infer from behavior whether or not an agent hyperbolically discounts. Moreover, we can often infer whether the agent is Naive or Sophisticated. These are the key ideas behind the algorithms for learning preferences I present in this chapter.

As a final point, I want to emphasize that the Naive/Sophisticated distinction does not just apply to hyperbolic discounters. The distinction lies in whether the agent predicts their own preference reversals (and so is able to plan to avoid them). This distinction is important in explaining a range of human behavior. Consider pre-commitment strategies, which are actions that cause a later action to be worse (or more costly) without any direct offsetting reward. John’s taking the long route in Fig. 7b is a pre-commitment strategy, as it keeps John far enough from the Donut Store that he’s not tempted. Here are some everyday examples of pre-commitment strategies:

- Spending time in places where temptations (TV, web access, alcohol) are hard to obtain, e.g. a cabin in the woods.
- Asking a boss/advisor for a hard deadline to provide an extra motivation to finish a project.
- Putting money in a savings account from which withdrawals are costly or time-consuming.

Although pre-commitment strategies are common, we still often observe time inconsistency in human actions. It’s plausible that Naive planning is part of the explanation for this.

---

12 By making the later action more costly, one’s future self is less tempted to take it. An extreme version, familiar from Odysseus and the Sirens, is when the later action is made almost possible to take.
2.2.3. Learning preferences for Agents with false beliefs and hyperbolic discounting

2.2.3.1. Inference: Bayesian generative models

I've introduced the Restaurant-Choice problem and the kind of agents that will face this problem (i.e. agents with inaccurate beliefs and with Naive/Sophisticated hyperbolic discounting). The overall goal of this chapter is to learn an agent's preferences from their choices on these decision problems. I employ a Bayesian generative-models approach (Tenenbaum et al. 2011). I define a large hypothesis space of possible agents and use Bayesian inference to infer both the preferences of the agent and their beliefs and discounting. The space of possible agents is defined by considering all possible combinations of the following properties:

(a) Cardinal preferences over restaurants (represented by a utility function), which includes immediate and delayed utilities for each restaurant.

(b) Degree of hyperbolic discounting (including no hyperbolic discounting at all), which is represented by a discount rate parameter (see Formal Model section).

(c) The agent's 'planning type', which can be Naive or Sophisticated.

(d) The agent's prior belief distribution. For example, whether the agent thinks it likely that a particular restaurant is closed (before getting close enough to see).

2.2.3.2. Evaluating algorithms by comparing to human judgments

I've outlined an approach to inferring preferences using Bayesian inference on a space of agents. Implementing this Bayesian inference procedure provides an algorithm for preference learning. Given observations of actual human behavior, this algorithm would infer some set of preferences for the agent, but how would we evaluate whether the inference is correct? This is a deep problem and (in my opinion) not one that has a single straightforward solution. Preferences cannot be directly observed from behavior and we are not infallible when it comes to accurately reporting our own preferences.

For this chapter, I adopt a simple approach. I consider a variety of behaviors on the Restaurant-Choice problem, including as a subset the examples illustrated in Figure 5 and Figure 7. I present these behaviors to a pool of human subjects (usually more than

---

13 In other words, what if the model is mis-specified and doesn't contain any model of a human that fits the data observed?
50 subjects) and have them make inferences about the preferences. I then compare the human inferences to the algorithm's inferences. The idea is that if the algorithm is doing a good job at inferring preferences, then it will give results similar to those of human subjects.

One might object that humans may not be good at this kind of 'third-person' preference inference. That is, maybe people are good at reporting on preferences that lead to their own choices but bad at inferring the preferences of others. For my experiments, I don’t this issue will arise. In the Restaurant-Choice problems, the choices take place over a short time-period and are between only a few simple options. Preferences over restaurants are familiar from everyday experience and so I expect that human third-person inferences will be reliable.

2.3.2.3. Dealing with non-omniscience and suboptimality

Deviations from omniscience and optimality mean that human choices can be consistent with many possible preferences. When Jill chooses X over Y, it could be because of a genuine preference for X or because of some bias or suboptimal evaluation that makes X look better than Y. My approach addresses this problem by explicitly modeling inaccurate beliefs and hyperbolic discounting. The idea is to use observed choices to jointly infer the biases and false beliefs of an agent as well as their preferences. With sufficiently varied decision problems, inaccurate beliefs and hyperbolic discounting predict behaviors that differ from more optimal agents. These differing predictions allow inaccurate beliefs and discounting to be inferred by Bayesian inference. The overall idea is that if you know exactly how a person is biased, you can use that knowledge to infer their preferences.

This approach can be extended to other biases or deviations from optimality. One problem with this approach is that a bias or deviations may be very difficult to model accurately. It might be relatively simple mathematically but hard to discover in the first place. Alternatively, it might be a model that requires a large number of fitted parameters. Ultimately, it seems hard to model choices without incorporated high-

---

14 In future work, I'd want to (a) record behavior from humans -- rather than making up plausible behavior myself, (b) get judgments from those individuals about their preferences. Then I could compare the inferences of the algorithm to the preferences as stated by each individual. I could get further information by also asking additional people to make impartial third-person judgments about the preferences.
level reasoning (e.g. moral deliberation where someone thinks deeply about a dilemma they face), and such high-level reasoning is difficult to model at present.

2.3.2.4. Distinguishing the normative and description in preference inference

Another general difficulty with inferring preferences from behavior is the issue of distinguishing the preferences from the normatively irrelevant causes of choice. My approach involves defining a large space of possible agents and then finding the agent features (beliefs, discount rate, preferences) that best predict the observed choices. But consider this general problem: 'Given a good predictive model for human choices, how do we extract from this model the preferences?' I will focus on a closely related problem: 'Given a good predictive model for a human's choices, how do we extract normative advice/directives for that person?'.

Consider this problem applied to agents with inaccurate beliefs and hyperbolic discounting. It's clear that in making normative recommendations to an agent we should ignore their false beliefs. For example, if John falsely believes his favorite restaurant is closed, we'd still say that John should go to the restaurant. (To get John to follow the advice, we need to convince him first that his belief is false. But I leave aside this practical dimension.)

What about an agent who hyperbolically discounts? It seems clear that the behavior of the Naive agent shown in Fig 7a is irrational, as the agent can do strictly better by taking other options (as we see with the Sophisticated agent in Fig. 7b). One idea is to make normative recommendations to a time-inconsistent agent based on what the most similar time-consistent agent would do. But there is still a question of whether to preserve the time preference that the hyperbolic discounting agents have. These agents prefer getting utility sooner rather than later. Should we treat this as part of their preference or as a bias that we ignore when making recommendations about what they should do?

There has been discussion about this issue (or closely related issues) from philosophers (Rawls, Parfit, Hausman 2012) and economists (Nordhaus 2007, Hanson 2008). While there is consensus that everyday false beliefs should be ignored when making recommendations, there is not a consensus about whether some kind of pure time preference is rational or not. How much of a problem is it that we can't say definitively whether time preference is rational or not? This depends on our goals. If
our goal is to write down the true preferences of an individual, then we can't do so without resolving this problem. However, we are still able to give normative recommendations conditional on each possibility (non-discounting and discounting) and the individual could simply choose for themselves which to follow.

3. Formal Model Description

3.1. Overview

I have given an informal description of my approach to preference inference. I now give a mathematical description of the model and explain my use of probabilistic programming to implement the model.

My approach infers an agent's preferences from their observed choices by inverting a model of (possibly suboptimal) sequential decision-making. I consider discrete sets of possible states and actions. The environment's dynamics are captured by a stochastic transition function from state-action pairs to probability distributions on states. Such transitions functions are standard in Markov Decision Problem (MDPs) and in Reinforcement Learning more generally (Sutton and Barto 1998). A Markov Decision Problem (MDP) is a sequential decision problem in which the current state (which is fully known to the agent) screens off all the previous states. The Restaurant-Choice problem naturally forms an MDP because the only information relevant to a decision is where the agent is currently positioned, not how the agent reached that position. If it matters how the agent reached the position then the problem is not an MDP.\(^{15}\) MDPs are decision problems. In what follows, I abuse terminology somewhat by talking about 'MDP agents'. By this I mean an agent with full knowledge of the current state, which is something agents in MDPs always have.

I consider three different agent models that vary in their assumptions about the agent's knowledge and time preference:

\(^{15}\) Though one can shoehorn the problem into the MDP form by enriching the state space with the information about how the agent reached his position.
**MDP agent**

This is an agent that does expected-utility maximization given full knowledge of its world state and of the transition function. The MDP agent discounts future utility in a dynamically consistent way (which means discounting is exponential or there's no discounting). The sequential planning problem for this agent can be solved efficiently via Dynamic Programming.

**POMDP agent**

A POMDP is a Partially Observable Markov Decision Problem. The difference between a POMDP and MDP is that in a POMDP the agent has uncertainty over which state it's in at a given time-step. The POMDP agent updates its beliefs based on observations. In evaluating actions, it takes into account the information it might gain by visiting certain states.

In a POMDP all of the agent's uncertainty is uncertainty over the state it is in. This does not actually restrict the kind of facts the agent can be uncertain about. For the Restaurant-Choice problem, the agent always has full knowledge of its location but it may be uncertain about whether a restaurant is open or not. We can represent this in a POMDP by defining the state to include both the agent's location and whether or not the restaurant is open or closed.

Belief updating for the POMDP agent is simply Bayesian inference. This depends on a likelihood function, which defines the probability of different observations in different states, as well as a prior on the agent's state.

**Time Inconsistent (TI) agent**

The TI agent is an MDP or POMDP agent that is subject to dynamic inconsistency. I use the hyperbolic discounting model of dynamic inconsistency (described above).

### 3.2. Formal Model Description

We begin by defining optimal sequential decision making for an MDP agent and then generalize to POMDP and TI agents. The formal setting is familiar from textbook examples of MDPs and POMDPs (Russell and Norvig 2003).
Let $S$ be a discrete state space and let $A$ be a discrete action space$^{16}$. Let $T$ be a stochastic transition function such that $T(s,a)$ is a probability distribution on $S$. The transition function species how the state of the world evolves as a function of the current state and the agent's current action. In the Restaurant-Choice problem the function $T$ is deterministic. (If the agent chooses to move right, then it moves right, and so on). However, I consider the more general case where $T$ is stochastic.

Sequences of world states are generated by repeated application of $T$ via the recursion:

*Equation (1):*

$$s_{t+1} = T(s_t, C(s_t))$$

In state $s_0$ the agent chooses the action that maximizes the expected total (discounted) utility of the future state sequence. This depends on a real-valued utility function $U$ on the state space $S$ and the standard exponential discount rate $\gamma$, and is given by$^{17}$:

*Equation (2):*

$$EU(a; s_0) = \mathbb{E} \left( \sum_t \gamma^t U(s_t) \right)$$

This expectation is over the distribution on state sequences $s_0, \ldots, s_k$ that results from action $a$, which is generated by Equation (1). So the expected utility of $a$ depends both on $T$ and on the agent's future actions after performing $a$. Hence in the MDP case, rational planning consists of a simple recursion. The agent simulates the entire world forward, conditional on each action. Simulating the world forward means recursively calling the agent to take future actions. This simple recursion does not hold in the case of TI agents. (Note that the inference problem for MDP agents is Inverse Reinforcement Learning).

Planning for POMDP agents is significantly more complex$^{18}$. Unlike in the MDP setting, the agent is now uncertain about the state it is in. The agent updates a belief

---

$^{16}$ I focus on a finite state space and a fixed, finite time horizon. Many aspects of my approach generalize to the un-bounded cases.

$^{17}$ I actually work with a more general model where the agent does not take the arg-max action but instead samples actions according to the standard soft-max distribution. This noise in the agent's choice of actions has to be taken into account when planning. This does not complicate the model or the implementation in a significant way.

53
distribution over its state on receiving observations from the world. Equation (2) still holds but the agent also has to integrate over its belief distribution.

Defining an agent with hyperbolic discounting requires a small change to the expected utility definition in Equation (2). We need to add a constant $k$, which controls the slope of the discount curve. The new equation is:

**Equation (3):**

$$EU(a; s_0) = \mathbb{E} \left( \sum_d \frac{U(s_d)}{1 + kd} \right)$$

I have replaced the time index $t$ with $d$ (which stands for 'delay'). This is to highlight the dynamic inconsistency of hyperbolic discounting. The hyperbolic discounter evaluates future states based on their delay from the present (not from where they are in the absolute time sequence). As the present moves forward, evaluations of future states change in ways that lead to inconsistencies.

As in the MDP case (Equation 1) the expectation in Equation (3) depends the agent's simulation of its own future actions. For a Sophisticated hyperbolic discounter, this 'self-model' is accurate and its simulations match what would happen in reality. So the Sophisticated agent's actions depend on the same recursion as the MDP agent.

By contrast, the Naive hyperbolic discounting agent simulates its own future actions using an **inaccurate** self-model. Let 'Mr. Naive' be a Naive agent. Then Mr. Naive will assume at time $t$ that his future selves at any $t'>t$ evaluate options with hyperbolic delays starting from time $t$, rather than from zero. So Mr. Naive wrongly assumes that his future selves are time-consistent with his present self and with each other. The assumption that the future selves are consistent is computationally simple to implement and can lead to efficient algorithms for planning for the case of deterministic planning (Kleinberg and Oren 2014). The Naive agent can plan cheaply. While this is an advantage, keep in mind that it will sometimes abandon these plans!

I have described the two main independent dimensions on which agents can vary. They can have uncertainty over their state (POMDP agents) and they can have time-
inconsistent discounting (TI agents). For inference, I define a hypothesis space of possible agents as the product of all these dimensions. First, all agents will have a utility function $U$ which represents their undiscounted preferences over all states. Second, each agent will have a probability distribution $B$ (for 'belief') on states which represents their prior probability that each state is the actual initial state. Finally, we let $Y$ be a discrete variable for the agent type, which can be ‘non-discounting’ (i.e. time consistent), ‘hyperbolically discounting and Naive’, or ‘hyperbolically discounting and Sophisticated’. The variable $k$ is the discount rate for discounting agents. So an agent is defined by a quadruple ($B$, $U$, $Y$, $k$) and our goal is to perform Bayesian inference over the space of such agents. I note that apart from the addition of TI agents, this inference problem is the problem of Inverse Reinforcement Learning (Russell and Ng 2000) or Inverse-Planning (Baker and Tenenbaum 2014).

In the most general formulation of this inference problem, only state sequences (rather than agent choices/actions) are observed. Hence the variables describing the agent are inferred by marginalizing over action sequences. The posterior joint distribution on agents, conditioned on state sequence $s_{0:T}$ and given transition function $T$ is:

**Equation (4):**

$$P(B, U, Y, k \mid s_{0:T}, T) \propto P(s_{0:T} \mid B, U, Y, k, T) P(B, U, Y, k)$$

The likelihood function here is given by Equation (2) or (3), with marginalization over actions. The prior distribution $P(B,U,Y,k)$ will reflect prior assumptions or background knowledge about which kinds of agent are more likely.

---

\[ ^{19} \text{For the Restaurant-Choice POMDP case, we can represent all the agent uncertainty we want by considering only the agent's prior on the start state.} \]
3.3. Implementation: Agents as Probabilistic Programs

3.3.1. Overview

The previous section gives a formal specification of the Bayesian inference we want to perform. This section explains my implementation of this inference for the Restaurant-Choice problem. The implementation uses recent ideas from probabilistic programming and is one of the contributions of this work. I explain some key ideas behind this implementation and then outline the implementation itself.

The Formal Model section (above) describes a hypothesis space of possible agents which vary in terms of their preferences, their beliefs and their discounting. Conceptually, Bayesian inference over this hypothesis space is straightforward. However, actually implementing this computation is not trivial. In order to compute the likelihood of observations for a given agent, we need to compute the actions that agent will take. For MDP agents, there are simple, efficient algorithms for computing an agent's actions. However, I do not know of any existing algorithms for the general case of an agent with uncertain beliefs and with either Naive or Sophisticated hyperbolic discounting. In this section I exhibit such an algorithm and discuss how it can be made efficient. Techniques from probabilistic programming are used both to implement this algorithm and (with minimal extra work) to do full Bayesian inference over agents implemented with this algorithm.

I use the ability of probabilistic programming to support models with recursion. My model of an agent can simulate its own future choices, by recursively calling itself to make a simulated future choice. The program expressing this model is short and many variant models of decision-making can be incorporated in a few lines of additional code.

3.3.2. Agents as Programs

Probabilistic programming languages (PPLs) provide automated inference over a space of rich, complex probabilistic models. All PPLs automate some parts of inference. Some provide default settings that the user can tweak (STAN, Church). Some allow the automated inference primitives to be composed to produce complex
inference algorithms (Venture). I work with the probabilistic programming language Webppl (Goodman and Stuhlmueller 2015). Webppl was designed in part to facilitate Bayesian inference over models that contain decision-theoretic agents. In particular, Webppl allows for Bayesian inference over models that themselves involve instances of Bayesian inference. This feature has been used to formulate models of two-player 'Schelling-style' coordination games, models of two-player adversarial games where agents' simulate each other to varying depths, and models of scalar implicatures for pragmatic language understanding (Stuhlmueller 2015).

To explain my implementation, I will start with simpler programs that capture one-shot decision-making with no deviations from optimality, and then build up to the full program that captures sequential decision making with hyperbolic discounting. I do not have space to give a detailed explanation of how these programs work, but an excellent tutorial introduction is available (Goodman and Stuhlmueller 2015).

I start with a program for one-shot decision making with full knowledge and no discounting. This programs illustrates a modeling technique that I use throughout, known as 'planning as inference' (Botvinick and Toussaint 2012). The idea is to convert decision making into an inference problem. The standard formulation of decision theory involves computing the consequences (either evidential or causal) of each possible action and picking the action with the best consequences (in expectation). In 'planning as inference', you condition instead on the best outcome (or on the highest expected utility) and then infer the action that leads to this outcome from a set of possible actions.

3.3.3. Generative Model: model for sequential decision-making

Figure 8 shows a model of an agent who chooses between actions 'a' and 'b'. The 'condition' statement specifies Bayesian conditioning. In this overly simple example, the agent conditions on getting the highest utility outcome. When inference is run on this program, it returns action 'a' (which yields this high utility outcome) with probability 1. Figure 9 shows a model of a slightly more complex agent, who chooses not based on maximal utility but based on maximal expected utility. This involves

---

20 This is analogous to the idea of backwards chaining, as found in work on logic-based planning. The motivation for using 'planning as inference' is that it transforms the planning problem into a different form. This form might be easier to capture with a short program or it might be amenable to techniques that make planning more computationally efficient.
computing the expected utility of each action and then using the statement 'factor' to
do a soft conditioning on this expected utility. This corresponds to soft-max decision-
making. For our purposes, this is a continuous relaxation of maximization that tends
towards maximization as the parameter 'alpha' tends to infinity.

```javascript
var consequence = function(state){
  if (state=='a'){return 20;}
  if (state=='b'){return 10;}
};

var agent = function(){
  var action = uniformDraw(['a','b']);
  condition( consequence(action) == 20);
  return action;
};
```

Figure 8 (above): one-shot, goal-based, planning as inference.

```javascript
var noisyConsequence = function(action){
  if (action=='a'){return categorical([.5,.5],[20,0]);}
  if (action=='b'){return categorical([.9,.1],[15,0]);}
};

var agent = function(){
  var action = uniformDraw(['a','b']);
  var expectedUtility = expectation(
    Enumerate(
      function()
      { return noisyConsequence(action); }
    ));
  factor(alpha*expectedUtility);
  return action;
};
```

Figure 9 (below): one-shot, EU-maximizing, planning as inference.

Figure 10 generalizes the previous models to a model for sequential decision-making
for the case of an MDP agent. This agent has full knowledge about the environment
up to features assumed to be non-deterministic. The agent model is essentially the
same as before. However, to compute the consequences of an action the agent now
invokes the function 'simulateWorld'. This function runs the whole world forward
until the decision problem (or 'scenario') terminates. Since this is a sequential decision
problem, running the whole world forward means computing the future actions of the
agent. So 'simulateWorld' will make repeated calls to the agent. So in computing the
consequences of an action 'a', the agent simulates its future actions given that it first takes action 'a'.

```javascript
var agent = function(state, numActionsLeft){
    var action = uniformDraw(['a', 'b']);
    var expectedUtility =
        expectation(
            Infer(function(){
                return discountedSum(
                    simulateWorld(transition(state, action)), numActionsLeft);
            }));
    factor(alpha*expectedUtility);
    return action;
};

var simulateWorld = function(state, history, numActionsLeft){
    if (numActionsLeft == 0){
        return history;
    } else {
        var action = agent(state);
        var nextState = transition(state, action);
        var nextHistory = push(history, nextState);
        return simulateWorld(nextState, nextHistory, numActionsLeft-1);
    }
};
```

Figure 10. Agent that solves MDP via planning as inference, conditioning on expected utility.

This program (Fig. 10) also includes a discount function 'discountedSum', which corresponds to the discounting function in the mathematical model description. By varying this function we can obtain different kinds of agent. If there is no discounting (or exponential discounting), we get the standard MDP agent (i.e. an agent with no false beliefs and no hyperbolic discounting). If we use the hyperbolic discounting function, then we get a Sophisticated agent.

I won’t provide the code here but a simple modification of the code in Fig. 10 gives us the Naive agent. The Naive agent has a model of its future self that is inaccurate. So the Naive agent needs to call 'simulateWorld' with an additional argument 'selfModel' by which it passes its own model of itself. This self-model discounts utility in the same way as the top-level agent (while the actual future self will discount in a way that is inconsistent with the top-level agent).
The final program (Fig. 11) allows the agent to have uncertainty over the start state. This requires two changes from Fig. 10. First, the world must generate observations (which will vary depending on the actual world state). This is done in the line 'var nextObs ...'. Second, the agent must update its beliefs on these observations. This is done in the assignment to variable 'stateDistribution' and is show in a schematic (simplified) form. The basic idea is that the agent samples the world forward from its prior distribution and then conditions the outcome on the observations. This distribution on the current state is then used when computing the expected value of the different actions.

I digress briefly to make a note on computational efficiency. The programs given here will be exponential time in the number of actions available to the agent. For the MDP agent without discounting, we can use standard dynamic programming techniques to get an algorithm quadratic in the size of the state space. For the POMDP case, we can use standard approximation techniques to get reasonable performance on small instances. For the Naive agent case, we can use a transformation of the state-space to turn this into a series of MDP problems. We form new states by taking the product of states and time-steps. The utility of a pair \((state, time)\) is just its discounted utility under the original state space. To choose an action, we solve an MDP over this state-space. The process is then repeated for every subsequent action. This algorithm will have runtime \(number\_actions \times (number\_states)^2\).
var agent = function(state, observations, numAct){
    var stateDistribution = Infer(function(){
        var startState = startStatePrior();
        var outcome = sample( simulateWorld(startState, [], t, [], random));
        condition( outcome.observation.slice(t) == observations);
        return outcome.observation[t-1]
    });
    var action = uniformDraw(['a', 'b']);
    var expectedUtility = E(function(){
        var state = sample(stateDistribution);
        return discountedSum( simulateWorld( ... ) );
    });
    factor(alpha*expectedUtility);
    return action;
};

var simulateWorld = function(state, history, numAct, observations, _agent).
    if (numAct == 0){
        return {history:history, observations:observations};
    } else {
        var action = _agent(state, observations);
        var nextState = transition(state, action);
        var nextHistory = push(history, nextState);
        var nextObs = push(observations, observe(nextState));
        return simulateWorld(nextState, nextHistory, numAct-1, nextObs);
}

Figure 11. Agent with potentially false beliefs, Bayesian belief updating on observations and hyperbolic discounting (implementing as in Fig 10). This is a POMDP agent with hyperbolic discounting (in the earlier terminology).

3.3.4. Inference model
The programs discussed above implement the model of decision making for different kinds of agents. For inference over these agents, I define a prior on the variables in Equation (4) (above) and use the built-in inference methods provided by Webppl. I show some actual inference results in the next section.
4. Experiments

In the previous sections, I outlined my Bayesian inference model and its implementation as a probabilistic program. This section describes the experiments I carried out to test the model's inferences of preferences against human inferences. The human subjects were recruited from Amazon's Mechanical Turk and were shown the stimuli via a web interface. Both the human subjects and my inference model were presented with sequences of choices from agents playing the Restaurant-Choice problem. The goal was to infer the agent's preferences and whether the agent had false beliefs or was time inconsistent.

4.1. Experiment 1

Experiment 1 considers only MDP and hyperbolic discounting agents and excludes agents with uncertainty (POMDP agents). The goal is to test the model's ability to capture key qualitative features of the human subjects' inferences. This includes inference about temptation (or immediate vs. delayed desirability of an outcome), and also prediction and explanation of actions.

Experiment 1 has a between-subjects design with three separate conditions. In each condition, subjects are shown three ‘scenarios’, each showing an individual ('John') choosing a restaurant in the same neighborhood. Subjects have to infer John's preferences by observing his movements across the three scenarios. Across the scenarios, there is variation in (1) John's starting location (which determines which restaurants he can get to quickly), and (2) which restaurants are open. Note that John will always know whether a restaurant is open or closed.

The stimuli for the Naive and Sophisticated conditions are shown in Figure 12. The Neutral Condition was already shown in Figures 4 and 5.
Figure 12. The three scenarios above and left make up the Naive condition. Subjects in the experiment were presented with these exact images and John’s backstory.

The Sophisticated condition differs only on scenario 3. Its version of scenario 3 is below left.

The three conditions agree on the first two scenarios, which show John selecting the Vegetarian Cafe. They differ on Scenario 3. The Naive and Sophisticated conditions exhibit actions that fit with Naive and Sophisticated agents respectively. In the Neutral condition, John chooses the same restaurant (Veg Cafe) every time and there is no evidence of inconsistency.

After viewing the scenarios, subjects answered three kinds of question. Part I was inferring John's preferences. Part II was predicting John's actions on additional scenarios (which added new restaurants to the neighborhood).
4.2. Model Predictions

We first describe the general results of our model inference and then discuss how well our model captures the inference of the human subjects on Parts I-II of Experiment 1.

4.2.1. Computing Model Predictions

I constructed a prior over the space of agent models described in Section 3.2 but leaving out the POMDP agents. I assume that John gets ‘immediate' and ‘delayed' utility for each restaurant. If a restaurant is tempting or a 'guilty pleasure', it might provide positive immediate utility and negative delayed utility. In place of continuous variables for the utilities and discount constant $k$, I use a discrete grid approximation. I did joint inference over all parameters using Webppl's built-in inference by enumeration (Goodman and Stuhlmueller 2015) and then computed posterior marginals.

To summarize inference over preferences for restaurants, I first introduce two properties of an agent's utility function that capture the kind of information that I’m interested in:

**Property 1: 'X is tempting relative to Y'**

We say that the Donut Store is tempting relative to the Vegetarian Cafe (or Noodle Store) if its immediate utility is greater than the Cafe's discounted utility (with one unit of delay).

**Property 2: 'Agent has an undiscounted preference for X over Y'**

We say John has an undiscounted preference for the Donut Store over the Cafe or Noodle Place if its summed, undiscounted utilities ('immediate' + 'delayed') are greater.

4.2.2. Model Posterior Inference

I present first the model posterior inferences for each condition. After that I compare these inferences to those of the human subjects.
The model does inference over the discount rate \( k \) of the agent, where \( k=0 \) means that the agent does not discount utility at all. For agents with \( k>0 \), it infers whether they are Sophisticated or Naive. The model also infers the agent's delayed and immediate utilities for the restaurants. (I present results only for the Donut Store and Vegetarian Cafe and don't report the results for Noodle Store).

![Figure 13. Graph showing the change from prior to posterior for each value of \( k \) and for each condition.](image)

Figure 13 shows the difference between the posterior and prior for a given value of \( k \). The prior was uniform over the values \([0, 1, 2, 3, 4]\) and so was 0.2 for each. We see that in the Naive and Sophisticated conditions, \( k=0 \) is ruled out with full confidence. In the Neutral condition, the model updates strongly in favor of no discounting. Figure 14 shows the inference about whether the agent was Naive or Sophisticated. Here the model strongly infers the appropriate agent type in each condition. There is no evidence either way in the Neutral condition. This is what we expect, as there's no evidence here in favor of discounting.
Inference for whether agent is Naive or Sophisticated

Figure 14. Graph showing posterior inference for whether agent is Naive or Sophisticated.

Posterior inference for whether Donut Store is tempting and whether it has higher overall (or 'undiscounted') utility compared to Vegetarian Cafe

Figure 15. Graph showing posterior inference for utilities of Donut vs. Veg Café.
Figure 15 (previous page) shows that an overall or undiscounted preference for the Donut Store was ruled out in all three conditions. By contrast, in the Naive and Sophisticated conditions there was a strong inference towards the Donut Store being tempting relative to the Cafe.

4.2.3. Model vs. Human Subjects

In Part I, subjects were asked to judge how pleasurable John finds each food option. In Figure 16, I compare this to model inference for the immediate utility of each option. Each bar shows the ratio of the pleasure rating for the Donut Store vs. the Vegetarian Cafe. The most noteworthy feature of human and model results is the high rating given to Donut Store in the Naive and Sophisticated conditions. In these conditions, John chose the Vegetarian Cafe two out of three times, and was willing to walk a fairly long distance for it. Yet the Donut Store is inferred to have an immediate pleasure close to that of the Vegetarian Cafe. This suggests inferences are sensitive to the specific context of John's choices, e.g. that donuts are tempting when up close. (Note that the Neutral condition, where the Donut Store gets a low score, shows that the model and humans did not have strong priors in favor of Donut Store being more pleasurable.)
The model differs from humans in assigning a higher immediate utility to Donut Store than to the Vegetarian Cafe. It's not clear how much to read into this discrepancy. First, subjects were asked how pleasurable the food was, which doesn't map exactly onto the 'immediate utility', because something with lots of negative delayed utility could easily be judged less pleasurable. Second, the ratios shown are point estimates and omit uncertainty information. Both subject responses and model inferences had significant variation, and we shouldn't take small differences in these point estimates too seriously.

In Part II, subjects predicted John's choices in two additional scenarios. In Scenario 4, John starts off far from both the Donut Store and Vegetarian Cafe. In Scenario 5, John starts off close to a new branch of the Donut Store. In the Neutral condition, people and the model strongly predict the Cafe in both scenarios. For Scenario 4, people and the model strongly predict the Cafe in both Naive and Sophisticated Conditions (see Fig. 17). These results show that people and the model infer a 'longer-term' preference for the Cafe across all conditions. For Scenario 5, people and the model both predict a much higher chance of John taking the tempting option (see Fig
18). This confirms the result in Part I that people and the model infer that the Donut Store is appealing to John, despite his only choosing it once.

Figure 17. Model vs. People for Prediction Task (Experiment 1, Part II).

Figure 18. Model vs. People for Prediction Task (Experiment 1, Part II).

4.3. Experiment 2

In Experiment 1, John was shown playing three different decision problems. Combined with the assumption of John having full knowledge, this enabled strong posterior inferences about John's preferences and his discounting. If instead, John is shown playing only a single decision problem, and inaccurate beliefs are not ruled
out, then there will be multiple explanations of John's actions and broader posteriors on parameter values. Experiment 2 tests the model's inferences from choices on a single scenario. I test whether the model can produce the same kinds of explanations as humans and whether it can capture the relative strength of different explanations. This time model includes POMDP as well as MDP agents (and it still includes hyperbolic discounting).

For Experiment 2, subjects were shown Scenario 3 in Fig 12 (Naive version), without being shown the other scenarios. Subjects are told that John generally prefers to go to somewhere closer to his starting point. They are also told that the two branches of the Donut Store are similar. Subjects were then asked to explain John's behavior and were allowed to write multiple explanations. (Subjects varied in how many distinct explanations they gave.) After looking through the explanations, I devised a series of categories that covered all the explanations given. I then categorized each subject's response, with some responses belonging to multiple categories. The proportion of subjects giving explanations in each category is shown in Fig 19. Note that the categories are not mutually exclusive. For example, the explanation 'Donut Store is closer' is likely to correlate with the explanation '[John] planned to go to the Donut Store'. Likewise, the explanation '[John] changed his mind about eating at the Vegetarian Cafe' is likely to correlate with '[John] planned to eat at the Vegetarian Cafe'.

Figure 19. Subjects' explanations for John's behavior in Naive scenario 3.
To compare the model inferences to these results, I computed the changes between the prior and posterior for various statistics of the agent model (Fig 20). The 'naive (temptation)' statistic corresponds to the explanation in terms of a Naive agent being tempted when walking right by Donut Store D2. The statistic 'ignorant of D1 / believes closed' tracks whether the agent has a prior belief that D1 is closed or whether the agent is ignorant of D1's being there. The statistic 'D2 is preferred to Vegetarian Cafe' measures whether D2 is preferred to the Cafe from John's starting point. This takes into account the utility of each restaurant as well as how far each restaurant is from John. (So the Donut Store could be slightly less preferred on its own merits, but John would choose it because it's closer than the Vegetarian Cafe).

4.3.1. Comparing results

We cannot directly map explanations like 'John changed his mind about eating at the Vegetarian Cafe' onto inferences from the model. It's possible that the change in mind is because of John being tempted by the Donut Store. But another possibility is that John was uncertain about the distance of the Cafe and changed his mind after
realizing it was too much of a walk. My model does not include uncertainty about distances.

Still, most of the explanations given by people are captured by the model in some form. Moreover, the strength of posterior inferences for the model roughly track the frequency of explanations given by people. It's plausible that subjects gave the explanations that seemed most plausible to them. Follow up experiments, where subjects were asked to choose between explanations, generally confirmed this.

There are two salient things about John's behavior. First, he doesn't choose the Vegetarian Cafe (or Noodle Store). Second, of the two Donut Stores, which are assumed to be similar, he chooses the further away branch (D2). The model makes a strong inference for a preference for D2 over the Vegetarian Cafe. It was not common for people to say directly that John prefers D2 to the Cafe (as they presumably thought this too obvious to be part of a good explanation). However, the categories 'Donut Store is closer' and 'John planned to go to the Donut Store' both fit will with this inference.

In terms of why John did not choose D1, the strongest answer for both people and the model is a preference for D2 over D1. The next strongest answer is that John doesn't know about D1 or thinks it is closed. These two explanations are judged much stronger than the explanation discussed above, on which John is a Naive agent and is tempted by D2 when walking towards the Vegetarian Cafe (and where D2 and D1 have very similar utility and John has full knowledge of D1). The model does capture this explanation (its posterior probability is higher than its prior). It does comparatively poorly as an explanation because it only works with a fairly narrow range of parameters values [This is an instance of the Bayesian Occam's Razor.]. For most parameter settings, a Naive agent would either (a) go straight to D1 (because the discounted utility of the Vegetarian Cafe is low compared to the much closer D1) or (b) go to the Vegetarian Cafe (because D2 isn't that tempting). The basic problem is that if D2 is really tempting, it must have high utility (especially immediate utility) and this will also make D1 more appealing.

In Experiment 1, the model infers the temptation explanation with very high probability. There are two reasons for this. First, I did not include the 'ignorance of D1' explanation as a possibility in that case. Second, in that case the model was doing inference based on three scenarios. Scenario 1 (Fig 12) shows John start off closer to
D2 than to the Vegetarian Cafe and still choose the Vegetarian Cafe. In Scenario 2, John starts closer to D1 but chooses the Cafe. These scenarios suggest that John has an overall preference for the Vegetarian Cafe. This now leaves the question of why he didn't go all the way to the Cafe in Scenario 3, and his being Naive gives an explanation.

Conclusion

This chapter addresses a key difficulty for inferring human preferences from observed behavior, viz. that humans have inaccurate beliefs, cognitive biases (such as time inconsistency) and cognitive bounds (that mean they so approximate planning and inference). The chapter looked at two particular deviations from optimal *homo economicus*: inaccurate beliefs and time inconsistency. The approach was to explicitly model these deviations from optimality and do inference over a large space of agents, where agents can include these deviations or not. Although the presence of these deviations from optimality can sometimes make preference inference more difficult, I showed that it is possible from short action sequences to diagnoses these deviations from optimality and to infer a significant amount about the agent's preferences.

This approach is quite general. If a deviation from optimality can be formalized, then it can be added to a Bayesian model. In this chapter, I implemented this inference model using a probabilistic program. I was able to use a single concise program (less than 150 lines of code for the agent model) to model a variety of different agents (MDP, POMDP and hyperbolic discounting versions of each). This program can be easily extended to capture other kinds of decision-making behavior. For instance, agents could have make plans of limited complexity, agents could have preferences that vary over time, or agents could do inference to learn about their own properties or parameters. One shortcoming of this probabilistic program is that it builds in inflexible assumptions about the structure of the agent's decision making. (Agents can vary in whether they hyperbolically discount and how they discount, but the structures that implement discounting and planning are fixed). Further work could explore more
flexible models that are able to learn more what kind of structures are useful for explaining human behavior (rather than assuming them).

The performance of my inference model was demonstrated by doing inference on simple sequences of plausible human actions. I compared the inferences of the model to those of human subjects. My Experiment 1 was for inferences about MDP hyperbolically discounting agents, where the agent is observed over multiple decision problems. The model produced qualitatively similar inferences to the human subjects on tests of (1) inferring the preferences of the agent and (2) predicting the agent's choices on new scenarios. In Experiment 2, inferences were based on a single, ambiguous decision problem and the space of possible agents included POMDP agents (which can have inaccurate beliefs). I showed that the model was able to produce much of the range of explanations that humans gave for behavior and that the model captured the relative plausibility of different explanations.

While these results testing the model are promising, it would be valuable to extend the model to richer applications. One project is to run an experiment where subjects both generate behavior and self-report their preferences. For example, subjects could be recorded using the web for a variety of tasks. Then we test the algorithm’s ability to infer their self-reported preferences.
Chapter 3:
Multivariable utility functions for representing and learning preferences

Introduction and Motivation

This section introduces some of the key ideas of this chapter. The basic approach of this chapter is to look at how preferences are modeled in economics and decision theory and consider whether this approach can scale up to learning all of human preferences. Preferences are often learned for a very specific task. A model that predicts well on such a task need not generalize well to novel tasks. Compositionality in utility functions is a simple way to facilitate such generalization.

This section begins by considering non-compositional utility functions and the problems they run into. The key point is that without compositionality, a utility function would need to contain a numerical utility for every way in which world histories can differ. This is implausible as an account of human cognition and would also make learning a complete utility function impossible. Along the way, I introduce the ideas of independent goods and complement goods, which play an important role throughout the chapter. These concepts correspond to simple functional forms for compositional utility functions.

In the final part of this section, I consider approaches that learn the utility function from a much more restricted class of functions. The idea is to work out the utility function a priori (or at least not to learn it by observing human choices). Such a utility function will presumably be defined in terms of properties for which humans have basic preferences (or at least properties that are closely related to those which are preferred in a
I discuss the difficulty of building formal techniques that make predictions based on such functions.

1. Utility function over objects of choice

Imagine an agent making the choice between some pieces of fruit (e.g. an apple, a pear, an orange, etc.). The agent chooses exactly one item from the list. When looking only at this single choice situation, we could perfectly predict the agent's choice from a utility function that assigns utilities only to the objects of choice (i.e. to each kind of fruit).

We can generalize beyond this example. In any situation, the agent has a number of choices available. If we learn a utility function over all objects of choice, then we can predict actions in any situation. Learning such a utility function is conceptually simple: you infer ordinal utilities from choices.

However, this approach has obvious shortcomings. If we only ever made choices between a small fixed set of kinds of fruit, this approach would be successful in making predictions. But we have an enormous set of options available, many of which are not choices of consumer goods. As a concrete example, consider the complexity of a utility function that assigns a unique utility to every item on Amazon.com or to items and services on offer at Craigslist. This approach makes computing predictions and plans from the utility function simple, but at the cost of making the utility function too complex to be learned from a feasible quantity of data. Such an approach can only be useful in very narrow domains, where the set of choices is fixed and small.

There are some obvious ways to patch up this approach. We can group together various choices for which the utility function has underlying structure. For example, instead of allowing unique, independent utilities to be assigned to "1 orange", "2 oranges", "3 oranges", and so on, we can instead specify a function $U(orange, k)$, which will be some increasing concave function of $k$. This corresponds to a standard practice in

---

1 This utility function is intuitively wrong. We don't have basic preferences over most of the objects of choice. For instance, I might choose one restaurant over another because I like that kind of food better (or because it's closer to my house) and not because of some sui generis preference. Moreover, there are things we value that we never directly choose, e.g. being a good parent or a good friend. In any case, this approach is very useful to describe in order to delineate the kinds of approaches one can take.
economics, where the utilities for different quantities of the same good are all systematically related by a simple function (usually with diminishing marginal utility).

We could also simplify the function by making it compositional. For example, instead of having a unique utility for the option “a Porsche and a swimming pool” we can compute this utility from the utilities of its components (“Porsche” and “swimming pool”). In this example, it’s natural to view the utilities as additive:

\[
U("\text{Porsche and swimming pool}") = U("\text{Porsche}") + U("\text{swimming pool}")
\]

These goods are independent, in the sense I introduce in Section 1 of this chapter (below). If many objects of choice were independent, then a utility function defined on objects of choice could be made much simpler by exploiting independence. Instead of having a unique utility value for each “bundle” or combination of objects of choice, you can compute utility values by summing the utilities of the components. Thus, the planning element has not become more complex in any significant way, but the utility function has fewer free parameters.

Even if many objects of choice were independent, this approach could still come unstuck. You still need to learn a utility for each of the independent objects, which would include all the items and services on Amazon and Craigslist. However, there can be compositionality without independence. Whenever you have \(n\) bike frames and \(2n\) bike wheels, you have \(n\) working bikes. Bundles of the form “\(n\) bike frames and \(2n\) bike wheels” will be systematically related in utility. But clearly these goods are not independent. The value of either two bike wheels or a frame on its own is minimal (ignoring resale value), and so the value of bike and two wheels is not equal to the sum of its components. Economists call such goods complements. I will discuss complement goods in Section 1.

One point to recognize is that there’s not going to be a general way to decide how the utility of a composite depends on the utility of the components. You need knowledge about the causal structure of the world\(^2\). This is the same kind of knowledge that you need

\(^2\) For example, the fact that two wheels are necessary for a working bike.
to compute plans from a utility function defined on more abstract objects. So simplifying the utility function by finding compositional structure in it will either require many observations (e.g. to determine exactly from observations how the utility of a bundle is related to the utility of its components) or will require the world knowledge needed for planning.

Note that some objects of choice could be decomposed into their properties. For example, the value of a cup of coffee for an agent may derive only from its caffeine content. This could be helpful in simplifying a utility function. However, it also requires world knowledge and it won’t work as a way to decompose all objects of choice (such as in the example bike wheels discussed above).

2. Hedonistic utility functions

On the approach discussed above, utility functions are defined on the objects of choice. This makes it easy to predict what someone will do given his or her utility function. However, the approach has serious shortcomings. Either the utility function is defined on a small number of objects of choice, in which case it won’t generalize to many different choice situations. Or the utility function is defined over a very large number of objects of choice, in which case it will be too hard to learn.

I will now discuss one example of a very different approach to modeling preferences. Instead of having utilities defined only on the objects of choice, we define a utility function that depends only on the agent’s hedonic state. For instance, we could pick out a particular brain state, and let utility be an increasing function of the intensity of this state. In this case, there is no need to learn the utility function, as we have specified it a priori. So we avoid any of the difficulties of learning the utility function from observing human behavior.

The choice of a hedonistic utility function here was just for the sake of having a concrete example. The general point is that we can select a utility function by some method that doesn’t involve learning it from human choice data. We can also use a hybrid

---

3 If my utility function was defined on having a working bike, then a rational plan to increase my utility would be to choose two wheels and a frame from the bike store.

4 One would need too many observations to find the utilities of every possible object of choice.
approach. We can choose a restricted class of utility functions (e.g. all functions that depend on some set of distinct hedonic states) and then select between this restricted class using data from human choices. This will make the learning problem much more straightforward.

The problem is how to compute predictions of choices from this utility function. Again I focus on the hedonistic function for concreteness. The difficulty is that many pleasure experiences result from elaborate sequences of actions (e.g. completing a hard assignment, re-uniting with an old friend, experiencing of great work of art). Computing a rational plan for maximizing utility will involving (a) considering this range of diverse events that would cause pleasure, (b) computing a sequence of actions that would cause these events. Mimicking the human ability to make such plans will require the kind of extensive and nuanced causal understanding of the world that humans have.

The general lesson is as follows. On one approach, we learn a utility function over a very large space of ‘objects of choice’. This will require many observations of choices in order to fit all the parameters. The advantage is that we don’t need an elaborate causal model of the world. The other approach is to choose a restricted class of utility functions. We now need much less data to pin down the utility function. The problem is that predicting behavior requires sophisticated causal knowledge.

It’s important to point out that in the former approach (where the utility function class is less restricted) we essentially have to learn the causal knowledge in an implicit way. If diverse events cause the same pleasure state, then we’ll have to learn a similar utility value for each of these states. (In this sense, there is some analogy between learning a utility function over a large set of states and model-free reinforcement learning).

3. Money maximization

Economists sometimes model humans as “money maximizers”. These are agents have utility functions that are some simple increasing (concave) function of the money in their possession. This utility function is a clear misrepresentation of human preferences. However, as a purely instrumental theory with the aim of achieving good predictive power, it has virtues. First, there’s no utility function to learn: all agents are assumed to
have the same simple utility function. Second, it's much easier to work out profit maximizing actions than actions that bring about some specific brain state. So it's easier to compute the plans for different agents, by using economic methods to work out profit maximizing strategies. Finally, the theory will give correct predictions to the extent that everything that people do in fact have preferences over can be bought and sold. There are lots of details here that I won’t discuss. I mainly want to note the structural features of this model of human preference.

Overview of Chapter

In Section 1 of the chapter, I consider in detail the approach above on which utility functions are defined on the object of choice (the first of the three examples above). I consider the kinds of compositional functions that economists use and how they relate to the structure of human preferences. In later parts of this chapter, I consider some novel practical applications of this approach to modeling human theory of mind. The question is how well this approach matches the way in which humans carry out “intuitive economics”, inferring utility functions from observations. This can also be applied to modeling the Knobe effect, where the inferences are not about preferences but about intentions.

5 As I've noted above, there are serious limitations to the approach where utilities are defined over objects of choice. However, it's important to understand the details of this very simple approach before moving on to richer but more complex approaches.
Section 1. Microeconomic models of human preference

1.1. Utility functions over bundles

1.1.1. Utility functions of a single variable

I start by considering simple models of preference based on utility functions. First, consider a utility function for money. If we make certain rationality assumptions, we can infer an agent’s utility function over money by observing their choices over lotteries. We expect money to have diminishing marginal utility for individuals, and thus for utility functions representing these preferences to be concave.

If we observe an agent making some choices between lotteries with cash prizes, we can use the information about their utility function to predict choices among different cash lotteries (which could be much more complex than the previously seen lotteries). The simplicity of the model will make it relatively easy to learn and compute predictions from the utility function. The accuracy of the model is an empirical question. What’s important for now is that the utility function we learn for money will be of limited use. We can predict choices between new cash lotteries, but not choices between lotteries over oranges or pencils (or separate information about the cash value of these objects). To predict choices over oranges, we’d need to observe some choices over oranges and infer a separate utility function over oranges.

---

6 Experimental evidence (Kahneman and Tversky 1981) suggests that people violate the rationality assumptions needed for the expected utility representation theorem and their choices over lotteries seem very hard to capture with simple functions (Rabin 2005). I will touch on these issues later on in the chapter.

7 The function need not be concave at all points. Even if it is always concave, it’s hard to see any a priori restrictions on the complexity of the function. For modeling purposes, it will often be easiest to assume that the utility function has a simple functional form and then fit the parameters of that functional form to the observed data.

8 The success of Prospect Theory could be seen as strong evidence against the idea that a utility-based model could be useful in making predictions. However, Prospect Theory tests subjects on a set of unfamiliar lottery choices in the lab. We might get different results by testing human experts in making decisions under different risk conditions (professional gamblers, traders, business people). In any case, the accuracy of the utility model for money is not my concern here.
1.1.1.1. Orange juice example

Things become more complex when we consider utility functions over multiple kinds of good. Suppose I am at the supermarket, deciding which brand of orange juice to buy. There are four brands and I have tried all of them before. We assume that I have diminishing marginal utility in each brand. But some brands will be better than others, and by varying degrees.

One starting point would be to observe choices (lotteries or other choice situations) that just involve at most one carton of any brand. From this we can put each of the four brands on a utility scale and predict choices over lotteries (or over deals in which brands sell as different prices). For instance, imagine I’m observed buying brand A over brand B, even though A is $0.50 more expensive per quantity. The inferred utility for A would be appropriately higher than for B. We’d thus predict that I’d take A over B if at some future time there was a change of prices which made brand A slight cheaper. (If A became slightly more expensive, we’d still expect it to be chosen. This depends on prior assumptions about how likely different utilis are.)

Having a single utility function defined on each of the four brands of orange juice will be adequate if all choices involve at most one unit of each brand. This is not realistic. We’d like to be able to predict whether someone would take two units of brand A over one unit of each of brands B and C. The naive way to do this would be to infer a utility function over each brand (as we described doing for money above) and then add up these utilities. This assumes the independence property introduced above. Let ‘1A’ mean ‘1 unit of A’, ‘2A’ mean ‘2 units of A’, and so on for other quantities and brands. Let $U_A$ be the single-variable utility function for brand A (and similarly for other brands). Let $U$ be the overall utility function. Then we have:

$$U(2A) = U_A(2A)$$
$$U(1B, 1C) = U_B(1B) + U_C(1C)$$

Yet this is clearly not right. Unless there are huge differences between the juice brands, there will be diminishing utility for any additional orange juice, not just for additional bottles of the same brand. The independence property does not hold here.
1.1.1.2. Printers, cartridges, and a pair of speakers

The phenomenon just described is not particular to orange juice. Similar examples are ubiquitous. I want to focus on another example where these simple utility models fail and where there's a quite intuitive solution. Suppose that Sue goes to an electronics store with $200 to spend. She is observed buying two HP print cartridges. Two weeks later she is in the store again with another $200 and buys some audio speakers. If we had to represent the print cartridges and the speakers as two points on a common utility scale (as in the orange juice case), we would have to infer that Sue's preferences had changed. Her first choice gave evidence that the print cartridges had more utility than the speakers and so would predict that two weeks later she would choose the print cartridge again.

For informed observers, the obvious explanation of Sue's behavior is not that her preferences have changed. Instead, we'd explain that if you have just one printer, you only need two cartridges and those cartridges will last for months. Extra cartridges will be of very little value to you if you already have two. This is not just a case of diminishing marginal utility. If you had no HP printer, then there'd be almost no utility (save for resale value) in having a single HP cartridge. If you had ten printers and needed to be using all of them at once, then a single cartridge would have no value and you'd need instead to have two for each printer. The general point is that the utility you get from a good depends on what other goods you have in your possession. We saw this in a less obvious example with the brands of orange juice. The marginal bottle of juice is less valuable depending on how many bottles of juice you have, even if the bottles are of different brands.

There is a diverse array of further cases showing the sensitivity of the utility of one good to other "goods" (broadly construed) in the agent's possession. A season ticket for the Metropolitan Opera becomes less valuable to someone if they take a job in California. A book in French becomes more valuable after you've learned French. An app for the iPhone becomes more valuable once you own an iPhone and again more valuable if you have other apps that complement its functionality or if the network service improves in
your area and speeds up your internet connection. In all these cases, it is plausible that the benefit to the agent of acquiring the good in question varies depending on the other goods the agent possesses, without there being anything we’d describe as a change in the agent’s preference.

1.1.2. Utility functions on bundles (multivariable utility functions)
What are the consequences of this kind of example for formal modeling of preferences? Let’s go back to the print cartridge case. When Sue enters the store with $200, it’s natural to construe all items in the store that cost under $200 as her options, because these are the things that she could feasibly buy from the shop. However, if we model her preferences as a utility function that assigns a utility to each one of these individual items (as we considered a function assigning each orange juice brand a point on the utility scale), then we won’t be able to make reliable predictions about her future choices.

One idea is to take the utility function to be defined not on the items in the store, but instead on sets or bundles of items that could be in Sue’s possession at a given time. We considered this approach above when asking about the utility of 2 units of brand A juice vs. one unit of each of brands B and C. I’ll illustrate with an example. We can imagine that Sue can only possess three kinds of goods (in varying quantities) at a time: printers, cartridges, and speakers. Bundles will have the form \((P, C, M)\), where \(P\) is the number of printers, \(C\) the number of cartridges and \(M\) the number of pairs of speakers. Each bundle \((P,C,M)\) will have a real-valued utility \(U(P,C,M)\). I’ll call this multivariable utility function a bundle utility function (and sometimes I’ll just say utility function).

The first time Sue enters the store, she buys two cartridges rather than one set of speakers (where both cost the same). This gives the information that:

\[ U(1,2,0) > U(1,0,1) \]

When she later buys the speakers over more cartridges, we get the information that:

\[ U(1,2,1) > U(1,4,0) \]

---

9 For contrast: At age 13, John is a self-described hater of opera. At age 23, after maturing in various ways, John enjoys going to the opera. This is naturally described as a change in preferences. It’s not clear in general how to decide between attributing a preference change vs. the acquisition of some good or ability that changes the utility of some option.

10 More generally we could consider sets of items whose total price is less than $200 budget.
Sue probably doesn’t explicitly construe her choice in the store as being between (1 printer, 2 cartridges, 1 pair of speakers) and (1 printer, 4 cartridges, 0 pair of speakers), but we presume that her choices in the store will be sensitive to possessions of hers like the printer. This will enable us to predict that in a future situation where Sue has lost her two print cartridges and her speakers, she is likely to first buy more cartridges and then speakers.

As with the single-variable utility function for money, we could infer an agent’s utility function over bundles by observing many choices over bundles. In the juice example, we considered a two-variable utility function over quantities of brands \( B \) and \( C \) that was just the sum of the single-variable utility functions over each brand. For the current example of Sue, an additive utility function, where the utility of a bundle \((P, C, M)\) is just \( U(P) + U(C) + U(M) \), is a worse fit than it is for orange juice. The additive model assumes that possession of one good doesn’t change the value of any other. Yet it’s clear that possession of a printer dramatically changes the value of cartridges. The general virtue of the additive model is its computational simplicity. On the additive model, the utility of a bundle of goods is just the sum of the utilities of its components. So adding more goods means only a linear increase in the number of objects that the utility function has to assign values to. If we allow goods to have arbitrary dependencies, then the number of objects will grow exponentially, because we need utilities for every possible combination of goods\(^{11}\). Thus in the printer example, it could be that \( U(0, 1, 0) \) is very low (no printer, one cartridge), \( U(0, 2, 0) \) involves an extra cartridge but zero extra utility, and \( U(1, 2, 0) \) is suddenly much higher (now you have both printer and cartridges).

### 1.1.2.2. Complement Goods

I’ve noted the space of bundles increases exponentially as we add more goods, and that in principle bundle utility functions could become that much more complex as the bundle size increases. To make such functions tractable, however, economists use multivariable functions that take simple functional forms. These forms include the additive function above but other forms are possible.

---

\(^{11}\) Having a domain that grows linearly is not sufficient for the function to be simple. You could still have a very complex function with a domain that is "small". However, other things being equal, smaller domains will make for a simpler function.
The printer example will provide a useful illustration. Suppose we just focus on bundles of two goods: printers and print cartridges. For concreteness, we’ll imagine that Sue is now working in an office that needs to print a large volume of pages in as short a time as possible. Sue would like to be running lots of printers in parallel in order to finish the job quickly, and each printer needs exactly two cartridges. (We also assume that no cartridges will run out during the time period). For Sue’s task, the function $U(P,C)$ will have a fairly simple structure. Two cartridges will lead to an increase in utility iff there’s at least one printer already in possession that lacks cartridges. Thus we get a function of the following form:

$U(P,C) = \min(P, 0.5C)$

This function does not model the diminishing marginal utility of working printers for Sue. However, assuming that this depends only on the number of working printers (and not on printers or cartridges independently), it would be straightforward to add extra parameters for this. Function (1) is usefully visualized using indifference curves (see Diagram 1 below). Each indifference curve is the set of all bundles that have a given amount of utility. (And so they pick out a set of bundles the agent is indifferent between in choosing).

Diagram 1 (left)

Indifference curves for $U(P,C)$ above, with quantity of printers on x-axis and cartridges on the y-axis. Colored lines are bundles with utility 1 (navy), 2 (cyan), 3 (yellow) and 4 (brown) respectively. Note $U(1,2)=1$, but $U(1+c,2)=1$ for any positive constant $c$. This is because adding a printer without cartridges does not increase the number of printers in operation.

\[ \text{Diagram 1 (left)} \]

Indifference curves for $U(P,C)$ above, with quantity of printers on x-axis and cartridges on the y-axis. Colored lines are bundles with utility 1 (navy), 2 (cyan), 3 (yellow) and 4 (brown) respectively. Note $U(1,2)=1$, but $U(1+c,2)=1$ for any positive constant $c$. This is because adding a printer without cartridges does not increase the number of printers in operation.

\[ 12 \] The “min” function returns the minimum of its two arguments. To be rigorous, we need to correct for the case where $P$ and $C$ are less than 1 and 2 respectively.
We can generalize the printer example. In the case of the printer, the source of utility is a working printer, and a certain fixed ratio of printers to cartridges is needed for a working printer. More generally, when a fixed ratio of goods $A:B$ is needed to produce additional utility we call the goods complements. The utility function over these goods is the Complements utility function and has the form:\(^{13}\):

\[
U(A,B) = \min(A, aB)
\]

The parameter $a$, which I’ll call the mixing ratio controls the ratio of $A:B$ needed to produce additional units of utility.

The example I’ve given of a Complements utility function only takes two variables (printers and cartridges). However, it is straightforward to generalize to more variables. For example, if each printer needs a special power cable as well as two cartridges, we’d get a “min” function on three variables. Yet it is clear that not every pair of goods interacts in the way complements do. For example, having both a pair of speakers and a print cartridge does not provide any extra utility over the utility provided by these goods independently. Thus, if we return to the bundle utility function above, $U(P,C,M)$, we can imagine it having the following simple functional form:

\[
(3) \quad U(P,C,M) = \min(P, 0.5C) + aU(M)
\]

The parameter $a$ in equation (3) controls the utility of a working printer relative to a pair of speakers. Here we have a kind of independence between goods that did not obtain between brands of orange juice or between printers and print cartridges. The intuition behind independent goods is almost opposite to the intuition behind complements. If goods are complements, then both are necessary in some specific ratio in order to produce something of value (e.g. a working printer). Diminishing utility for each complement good will generally be a function of diminishing utility for the product of the complements (e.g. working printers). Independent goods produce utility by independent means. Hence, diminishing utility for each good will also (in many cases) be the result of the diminishing ability of the good itself to produce utility. For example, most people will get little value from additional pairs of speakers because one set enough. This is completely independent of how many printers one has. Thus in the general case, a utility

\[^{13}\] Again, I’m not going go into the details of ordinal vs. cardinal utility functions and of diminishing marginal utility.
function over a bundle of independent goods will decompose into a sum of single-variable utility functions over each of the goods, each of which will have its own independent marginal behavior.

1.1.2.3. Substitute Goods

The notion of independent goods is related to another standard notion in microeconomics. Economists use the term *substitutes* for goods that perform the same function. The idea is that if the price of one good increases, some consumers will use the other good as a substitute. For example, when the price of butter increases, demand for margarine increases (all things being equal). Two substitute goods could perform their function equally well and thus be so-called *perfect* substitutes. For example, I can't distinguish Coke from Pepsi and so they are perfect substitutes for me. However, if you have a strong preference for Coke over Pepsi, they will be imperfect substitutes for you. You will always choose Coke over Pepsi at the same price and the price of Coke will have to get high before you switch to Pepsi.

Substitute goods differ in an important way from complement goods. Complement goods \(A\) and \(B\) produce utility only when combined in a fixed ratio: they are useless alone or in the wrong ratio. Substitute goods each perform the same function. So the value of having units of each good at the same time will just be the sum of their value independently (modulo diminishing utility). For example, suppose I have a slight preference for Coke over Pepsi and rate Pepsi as having 90% of the value of Coke. Given a Coke and a Pepsi, I drink the Coke today and the Pepsi tomorrow. This is no different from summing up the utility of the world where I just have a Coke today with the utility of the world where I just have a Pepsi tomorrow. Thus the *Substitutes* utility function will have the following functional form (indifference curves are below):

\[
(4) \quad U(A, B) = A + \alpha B
\]

The \(\alpha\) parameter, or mixing ratio, now controls how much utility a unit of \(B\) provides relative to \(A\). This parameter might take different values for different individuals. For example, if Pepsi and Coke are perfect substitutes for me, then I'll have a utility function for which \(\alpha=1\). Someone who is willing to pay twice as much for Coke as Pepsi (an unusually strong preference) would have \(\alpha=0.5\). Note that the mixing ratio for the
*Complements* function could also take different values for different individuals, though this wouldn’t happen in the printer case.

**Diagram 2 (left)**
Indifference curves for \(U(A,B)\) above, with quantity of \(A\) on the x-axis and \(B\) on the y-axis and with \(\alpha=0.5\). Colored lines are bundles with utility 1 (navy), 2 (green), 3 (brown) and 4 (cyan) respectively. Note that if we hold fixed the value of \(A\) at \(A=1\), and then vary \(B\), there is no change in the marginal value of increasing \(A\). (Taking partial derivatives w.r.t. \(A\), we can ignore \(a\beta\) as it’s a constant).

Equation (4) shows that Substitutes have the same additive functional form as independent goods. This makes sense. Coke or Pepsi provides the same “function” of being a cola drink of the same size, same sweetness level, same caffeine content, and so on. Moreover, they each provide this function independently. They don’t combine in the way that the printer and print cartridge do. In this way, they are similar to the pair of speakers and the printer. However, since the cans of Coke and Pepsi provide the same function (unlike the speakers and printer) and there is diminishing utility in this function, they are not independent in the way they contribute to diminishing marginal utility. Thus, as with the Complements function (2), we will need to add parameters outside of function (4) in order to model diminishing utility. The intuition here is very simple. One’s enjoyment of cola will drop quickly after a couple of cans. Whether the cola is Coke or Pepsi will make little difference to this drop in enjoyment. In contrast, additional Porsche cars won’t change the value of additional swimming pools. Additional printers don’t change the value of additional pairs of speakers.

Here is a summary of the functional forms, if we include parameters for diminishing marginal utility. For concreteness, I have chosen a particular functional form for the diminishing marginal utility, viz. raising to an exponent. The exponents below \((r\text{ and }s)\) are all bounded between 0 and 1. One example is the square root function, for which we’d have \(r=0.5\).
\[(\text{independent}) \quad U(A,B) = A' + aB'\]

\[(\text{substitutes}) \quad U(A,B) = (A + aB)'

\[(\text{complements}) \quad U(A,B) = \min(A, aB)'

\subsection*{1.1.3 Summary}

For the example of Sue making purchases in the electronics store, we considered representing Sue’s choices in terms of three single-variable utility functions, one for each good:

\[U(P,C,M) = U(P) + U(C) + U(M)\]

This functional form, although very simple, fails to capture the dependence between printers and cartridges, which produce additional utility only from an additional one printer and two cartridges. It is clear that if we are able to consider any possible function on \((P,C,M)\), then we can find a function that allows us to capture the dependency between printers and cartridges. However, we would like to work with a function that is both analytically and computationally tractable. At this point, we introduce non-additive multivariable utility functions that are simple and tractable but also capture the way in which different goods can interact to produce utility. We introduce complementary goods and substitute goods along with their respective utility functions. We then give a plausible Complements utility function for \((P,C)\). We note that \((P,C)\) and \(M\) are independent in the way they produce utility. Hence the three-variable utility function on this bundle will be the following:

\[(5) \quad U(P,C,M) = \min(P,0.5C) + aU(M)\]

We can then add diminishing marginal utility to each of the independent components of this sum. This example of the printer involves only three goods. We’d like to consider agents who, like modern consumers, make choices among a vast number of goods. Let
\( A_1, A_2, \ldots, A_n \) be the quantities of \( n \) goods that an agent is choosing between. Then we can straightforwardly generalize the approach taken in the printer example. The \( n \)-variable function \( U(A_1, A_2, \ldots, A_n) \) will decompose into various functions on fewer variables, due to dependencies between the goods. For instance, if \((A_1,A_2)\) are complements and are independent of every other variable, we’ll get:

\[
U(A_1, A_2, \ldots, A_n) = \min(A_1, a_1A_2) + a_2 U(A_3, A_4, \ldots, A_n)
\]

As I will discuss below, there are dependencies between goods other than those captured by the notions of substitutes or complements. However, we have the outline of a general approach to modeling preferences. Given a set of goods that can be purchased in arbitrary quantities, we model an agent’s preferences over bundles of goods using a multivariable utility function. This function decomposes (or "factors") into smaller functions, capturing dependencies or interactions between subsets of the set of all goods.

So far I have not given any empirical evidence in support of this approach to modeling preferences. Empirical testing of this model will be explored in Section 2 of this chapter. In the last subsection of Section, I will explore extensions to this bundle utility approach. Before then, I want to consider at some length a theoretical objection to the approach to modeling suggested so far.

### 1.2. Interlude: Objecting to the Bundle Utility framework

The move from single- to multi-variable utility functions is a small departure from our starting point. We start by using simple functions on single goods, such as different quantities of oranges. We switch to using simple functions on bundled goods like combinations of printers and cartridges, which we write as \((\text{number of printers, number of cartridges})\). The utility function is first defined on the object of choice (e.g. an orange, a print cartridge, etc.) and then defined on a bundle of objects (which may or may not be chosen as a bundle). What the printer example really points to is that we get value not from printers or from cartridges, but from having a working printer. The value we assign
to having a working printer will explain our choices over printers and cartridges. This knowledge also explains why, having been given a Canon (rather than HP) printer as a gift, we would suddenly become interested in buying Canon print cartridges.

The transition to bundle utility functions suggests a general approach that could be taken much further than we have gone so far. The approach is to try to define the utility function over the world-states that are “more” intrinsically valuable. I’m using the word “intrinsic” in a limited sense here, avoiding any controversial philosophical claims. I don’t mean that to model someone’s preferences we must first solve the problem of discovering the sources of intrinsic value for an individual. I’m supposing that there is some set of goods or outcomes that we consider defining the utility function over. We can then talk about outcomes being intrinsically or instrumentally valuable relative to this set.

If the set contains just “printer”, “print cartridge”, “working printer”, and “pair of speakers” then we might say that (relative to this set) the outcome “working printer” has intrinsic utility for an agent. Of course not many people find a working printer intrinsically valuable in absolute terms. So in this example, it will easy to add items to the set such that “working printer” is no longer intrinsically valuable relative to the set. (If the printer is used for printing a term paper, then we could add to the set the outcome ‘printed term paper’ or ‘term paper that got a good grade’ or ‘getting a good grade for the course overall’. Any of these would displace the printer in terms of having intrinsic value relative to the set).

The virtue of assigning utilities to more intrinsically valuable outcomes is that these utility functions should lead to more robust predictions. As noted above, if HP print cartridges are assigned independent value, then we’ll predict that Sue will buy cartridges even if she is dispossessed of her printer and lacks funds for a new one. Likewise, if having a working printer is assigned independent value, we’ll predict that Sue will invest in printer equipment even if her class now allows for email submission of assignments. It is not surprising that the “intrinsic values” approach should lead to more robust predictions. It is a move towards a model that better captures the actual causal dynamics of the situation. These causal dynamics may be extremely complicated. Intrinsic or terminal values for a single human may be diverse and may change over time (e.g.
outcomes previously having only instrumental value coming to be intrinsically valuable). A further complication is weakness of will.

This "intrinsic values" approach seems to me worth pursuing both for predictive purposes in psychology and economics and (even more so) as a proposal about how folk-psychology makes predictions based on preferences. To have a flexible model for predicting human actions across a wide range of situations, this approach seems necessary. Humans can take a huge range of actions (or choose from a huge range of options) and it’s not feasible to construct a utility function that includes a utility for all of them (or for all bundles of them).

However, the “intrinsic values” approach is will be difficult to get to work. First, as the printer example suggests, the outcomes that have more intrinsic value tend to be more abstract and further from observable or easily recordable choice. Compare “having a print cartridge” to “communicating my assignment to the professor”. This can make the connection between preference and action much more complex. Consider that there could be various kinds of action that satisfy an abstract goal like “communicate ideas to professor”. (In the microeconomic bundle models considered above, this connection is very simple: you will choose a unit of a good if the resulting bundle is utility maximizing).

There is also a challenge in learning which outcomes have more intrinsic utility for an agent. For instance, suppose Sue needs a working printer to print an assignment for a class. Given our folk knowledge of the situation, we’d think it unlikely that Sue gets any intrinsic value from being able to print the essay. If, hypothetically, she were offered to submit the paper electronically, she would probably do so$^{14}$. A further question is whether Sue gets any utility from producing the essay (as some students certainly would). We can easily imagine a case where Sue is merely taking the class to fulfill a requirement and that she places no value at all on the activity of completing the essay. If she could fulfill the requirement by going to see a movie, she would much rather do that. Fulfilling the requirement for a degree may also be merely instrumental. If Sue were offered the degree

---

$^{14}$ Sue may also take (at least some) pleasure in producing an attractive physical instantiation of her work. Human preferences are complex in this way: something with mainly instrumental value can also have some intrinsic value. (That Sue would not continue to print papers were electronic submission allowed does not imply that she gets no intrinsic value from printing papers).
(or the employment advantage of the degree) without having to fulfill the requirements, she might take it.

In general, it’s hard to work out which states are merely instrumentally valuable because it’s hard to get access to the relevant discriminating evidence. Usually, there’s no way to fulfill the class requirement other than by writing the paper. We can look at whether Sue writes papers just for fun. But her not doing so is inconclusive because she may be too busy writing papers for class or because of akrasia. We can ask her whether she’d skip writing the paper if given the option. But it’s not a given that her answer would track her hypothetical behavior. So this extra information gained by interviewing Sue and examining her general paper-writing habits would not necessarily give us that much information. If we did have good information about Sue’s intrinsic values, we could still have difficulty predicting her actions based on this. We face the double challenge of coming up with a formal model for making rational plans and then adapting the model to take into account ways in which humans might deviate from the rational-planning model.

The microeconomic bundle-utility approach we have sketched above, in which utility functions are defined over sets of dependent goods, can be seen as a small first step towards defining utility functions over the states that have intrinsic value to the agent. Learning these utility functions won’t allow us to generalize in the way that we’d ultimately like to. (Example: If Sue buys print cartridges today, she probably has some things she wants to print in the next couple of days. If her printer malfunctions and she is offered a cheap printer by an acquaintance, she will probably take the offer. On the other hand, we might be able to learn these simple utility functions quickly and be able to make reasonably good predictions about choices made under similar underlying conditions. However if the world changes, e.g. a university switches to only accept electronic submission of papers, then we won’t always be able to predict choices very well.)
1.3. Bundle Utility: Preference for variety and other challenges

1.3.1. Multiplicative, Cobb-Douglas Utility functions

The previous section discussed some of the limitations of bundle utility functions and a general approach to remedying these limitations by defining utility functions over more abstract, more intrinsically valuable outcomes or states. In this section, I return to the bundle-utility model itself. I want to discuss some other ways in which goods can interact to produce utility and discuss how such interactions are problematic for this formal model.

I start with an example. Suppose that Scott is preparing food for a large group of people. His aim is the greatest enjoyment of the greatest number of people. There is excess demand for the food he makes, and so he's trying to make as much as possible—as well as making it tasty. Scott can hire more chefs, leading to better recipes and more expert food preparation. He can also buy more cooking equipment, enabling him to output the same food but in larger quantities.\(^{15}\) Let \(C\) be the number of chefs Scott has and \(E\) be the number of units of cooking equipment. We can imagine that for small values of \(C\) and \(E\), there is a multiplicative relation between \(C\), \(E\) and \(U(C,E)\). For example, for a given quantity for \(E\), if Scott doubles \(C\) then we could imagine a doubling in utility. (If quality of recipes and of food preparation doubles, then every dish served will double in quality and so the total utility doubles). Likewise, doubling the amount of cooking equipment doubles the number of dishes produced (without changing the quality of the dishes) and so doubles the number of people who get food, hence doubling the utility. Here is a utility function with this property (where ‘\(CE\)’ means ‘\(C \times E\)’):

\[
(6) \quad U(C,E) = CE
\]

This utility function seems likely to have limited application. We know that adding more chefs can only improve the recipes so much (and in practice we expect sharply

\(^{15}\) Assume that all other inputs to Scott's food production don't cost him anything—he already has ample free labor to oversee the food, but it's not skilled.
diminishing returns to more chefs). Moreover, we may expect Scott's utility to be non-linear in the number of people satisfied. (A doubling of the number of satisfied customers will not double his utility).

Nevertheless, the example of Scott illustrates a way in which goods can interact that differs from independent goods, substitutes, and complements. For independent goods, changing the quantity of one good has no impact on the utility gained by a marginal increase in the other good. (Whether you have one or two pairs of speakers makes no difference to the utility gained by having a printer). For substitutes, marginal increases in one good can change the marginal utility of increases in the other good, but only by making the utility lower (due to diminishing marginal utility). Compare this to the food example. If Scott has lots of equipment and so can produce lots of food, then he gets more benefit from having more chefs than if he produces only a small amount of food. Something like this is also true for complement goods. However, for complements, there are no gains in utility without increases in both goods in the appropriate ratio. (If you have no print cartridges, you get no benefit from adding more printers). Whereas in the food example, having more chefs will always improve utility (even if only by a small amount) and the same is true for increases in productive capacity.

Equation (6) above is an example of a Cobb-Douglas Utility function. This utility function (named for the inventors of an analogous production function in economics) is for goods that have a multiplicative effect on utility. (Put differently: for cases where one good is a limiting factor on utility relative to the other goods). The general functional form is the following:

\[
U(X,Y) = X^a Y^b
\]

Here, the constants \( a \) and \( b \) control the marginal utility of each the goods, i.e. the rate of increase of \( U \) as one of \( X \) and \( Y \) increases. Utility increases fastest when the two goods are combined in a ratio that's a function of \( a/b \). The indifference curves for this function are below:
Diagram 3 (left)
Indifference curves for \( U(X,Y) \) above, with quantity of \( X \) on the x-axis and \( Y \) on the y-axis and with \( U(X,Y) = X^a Y^b \). Colored lines are bundles with utility 1 (navy), 2 (green), 3 (brown) and 4 (cyan) respectively. Note that if we hold fixed the value of (say) \( A \) at \( X=1 \), and then vary \( Y \), there is no change in the marginal value of increasing \( X \).

The example given above of a Cobb-Douglas utility function (the food example with Scott) was more plausible as a case of a utility function for a company, or some other kind of organization, than as a utility function for an individual. In general it is harder to come up with goods that have multiplicative utility for individual. There are cases similar to the food example. It could be that someone’s scientific output doubles in value as a multiplicative function of the number of students in the lab and the amount of equipment and computing power in the lab. We could also consider the quality of different eating experiences. If we had a small slither of tiramisu, we can improve the taste by adding a small amount of rum. If we now have a full serving of tiramisu, the same amount of rum might now have even more effect on the overall experience, as it will add some rum flavor to the whole serving. However, it’s clear that the multiplicative effect will be weaker, because there is a “complements” interaction with the rum and tiramisu that is also important to the utility. Another possible example is the utility gained from relaxing activity (e.g. lazing in the pool) after a workout. Marginal increases in the length of the relaxation activity have more impact on utility if the workout was longer or harder.
1.3.2. Preference for variety

The examples discussed above in connection to the Cobb-Douglas utility function are related to a broad class of important examples involving what I’ll call “preference for variety”. Imagine Dave goes to the fridge and chooses a strawberry yoghurt over a vanilla yoghurt. If the next day Dave chooses vanilla (when faced with an identical choice and assuming Dave has ample past experience with each flavor) we would probably not attribute to Dave a change of preference. We would probably think that Dave just likes some variety in his yoghurt flavors. (The same goes for someone who eats Indian food one night and Chinese food the following night). Preferences for variety exist for a very wide range of goods or options: from very particular and insignificant options (e.g. flavor of yoghurt, which song from a CD to play, which route to take home) to life choices of great importance (e.g. jobs, significant hobbies, living location).

Preference for variety is closely related to diminishing marginal utility, and I want to briefly explore that relation. First, take the diminishing marginal utility of money. Assume that I get a yearly salary and I’m not allowed to save any of my money. Then my utility increases more if my yearly income moves from $10,000 to $30,000, than if it does from $150,000 to $170,000. We get evidence for this from my risk-averse choices. The intuitive underlying reason for the diminishing utility of money is that the $10,000-30,000 raise enables me to purchase things that make a bigger impact on my quality of life (e.g. a better apartment, a car) than the $150,000-170,000 raise. Note that if I can save money, then the situation is more complicated. The extra money in the $150,000-170,000 case might have a big impact if I am later out of work. The utility gained from the extra money will therefore depend on the likelihood of my being alive in future years and in need of using it.

Related to the diminishing utility of wealth is the diminishing utility of various goods. Suppose there’s a cake that needs to be eaten today. Then I get more value in going from 0 to 1 slice than from 1 to 2 slices or 4 to 5 slices. (After four slices, I may get no value at all from extra slices). Yet as with money, things are more complicated if my consumption is spread out across time. If I eat only one slice per month, then going from 4 to 5 slices

\[16\text{Another factor here is whether I discount my future welfare relative to near-term welfare.}\]
could be just as good as going from 0 to 1 slice. So a good can have sharply diminishing marginal utility for an agent when consumed at one time, without the agent having any preference for variety even over short time spans. For example, someone might be able to only manage one slice of cake at a time, but still prefer to eat the same cake every day with minimal loss of utility over time.

We could try to assimilate these two phenomena of diminishing utility and preference for variety. After eating a slice of cake the utility of an extra slice immediately goes down. This utility will rise over time but the growth rate can vary between individuals. For some it might take hardly any time at all, for others it might be a day or week. The diminishing utility of cake slices over a given time period will depend on this time-utility function. The problem with this assimilation is that it fails to capture important differences in the underlying phenomena. Moving from 4 to 5 slices of cake will be bad for most people due to satiation. This is nothing particular to the kind of cake in question. On the other hand, a preference for having different kinds of dessert on consecutive days seems to be based on a preference for different flavors and taste experiences.

Assimilating diminishing utility and preference for variety is not straightforward. For the rest of this chapter, I will use the terms in the way suggested by the cake example above. That is, “diminishing utility” will refer to phenomena like the decline in utility of subsequent slices of cake at a given time. “Preference for variety” will be used for phenomena like the decline in utility (independent of satiation) of a given kind of cake that results from having consumed that kind of cake recently.

There is lots of work in economics and other social sciences studying diminishing marginal utility of wealth in humans\(^\text{17}\). In contrast, I know of very little work studying the preference for variety within a utility-based framework\(^\text{18}\). There are two kinds of choice situations that one could try to model and predict that depend on preferences for variety. First, there is the choice between single items at a particular time. For example, if Dave goes to the fridge for yoghurt after every meal, we can try to predict what flavor he

---

\(^{17}\) The phenomenon is more complex than I space to discuss here. It seems that some of the value of wealth is contextually sensitive to one’s peer group. There’s also some psychology/experimental philosophy looking at human intuitions about this notion: Greene and Baron 2001.  
\(^{18}\) “Novelty, Preferences and Fashion” (Bianchi 2001) explicitly addresses the dearth of work in this area and discusses some economic work that at least touches on these issues.
chooses. Here we need some function that assigns flavors decreasing utility as a function of when they were last sampled. Second, we could model the choice of bundles of goods that are chosen with preference for variety in mind. For example, suppose Dave slightly prefers vanilla to strawberry, in the sense that if he had no yoghurt for months and could only choose one, he’d take vanilla first. Given his preference for variety, Dave might still buy 6 vanilla and 6 strawberry yogurts. Thus we would need a utility function over bundles of vanilla and strawberry that assigns more utility to the bundle (6 vanilla, 6 strawberry) than to any other bundle of 12 yogurts total. The utilities we get for these bundles would not necessarily be very robust: If Dave switches to eating yogurt less frequently, he may switch to a bundle with a higher proportion of vanilla.

Here we have the same problem noted above concerning what the utility function is defined over. The entity with more intrinsic value is the sum of the yogurt-consumption experiences, rather than the bundles containing different numbers of vanilla and strawberry yogurts. To better match the actual structure of the agent’s values, we need a model which takes into account when the yogurt will be eaten and how much preference for variety Dave has.

It seems likely that preference for variety as manifest in choices over bundles will combine aspects of the “independent” and “substitutes” Utility functions. Having had lots of vanilla yoghurt, we switch to strawberry because vanilla has got boring and strawberry is “newer” or “fresher”. As with substitutes, there is declining marginal utility in additional units of any kind of yoghurt. But there is separate, independent diminishing utility in particular flavors. It’s also worth considering that there is some Cobb-Douglas-type non-linearity with some goods. Maybe two different art forms or different holiday locations can interact in a positive and non-linear way.

---

19 It’s worth noting that the choices people make when purchasing a bundle may not match the choices they make if they are making each choice separately. On the one hand, it may be hard to predict when you’ll get bored of something. And on the other hand, people choosing bundles might think more carefully about the need for variety because they have the whole “schedule” laid out before them. There is empirical work studying inconsistencies between this kind of decision: see Loewenstein and collaborators on the so-called “Diversification Bias”.

100
1.4. Temporally extended preferences

In the previous section, I introduced a third kind of utility function (the Cobb-Douglas function) and gave some textbook-style applications for this function in modeling human preferences over goods that interact multiplicatively. I speculated that this kind of interaction is less common than substitute or complement interactions. I introduced preference for variety and discussed its relation to diminishing marginal utility and how it could be modeled using bundle utility functions. The preference for variety is a very common feature of human preferences that any comprehensive modeling effort will have to reckon with. Moreover, it suggests the importance of temporal features of consumption in determining value. The value of a bundle of different flavors of yogurt will be contingent on my consuming the yogurt in the right sequence and over the right duration. What has value is a particular world history (the one in which I alternate between flavors and space out my consumption) and having the mixed bundle makes this world history possible.

There are other examples where the entities bearing value are even more particular. Consider trying to solve a hard problem in mathematics or physics. (The problem could be an unsolved open problem, or it could be a recreational or textbook problem). If the problem is interesting and theoretically important, then after working on it for a while I will probably be extremely curious to know the answer. However, if you offered to tell me the answer, I would say no. I would even incur costs to avoid hearing the answer. However, once I’m almost sure that I’ve got the answer myself, I’ll be very happy to hear you relay the right answer.

Here's a similar example. When you finish a videogame you are generally treated to an ending sequence that presents the climax of the narrative of the game, often with interesting visuals. Having played much of the way through (and got caught up in the narrative), you might be very curious to see the ending, but you wouldn’t want to see it unless you had finished the game yourself. You value seeing the ending sequence of the game and seeing the closure of the narrative, but part of the value of the seeing it is conditional on having “earned it” by succeeding at the game and following the narrative all the way through to the end. Similarly, a novelist may crave praise and prizes from
critics but might not accept them she hadn’t produced work that she felt merited these accolades. (Imagine the novelist gets praise for a work by a mysterious author of the same name. Or that the novelist writes a very mediocre work that is feted because people have some independent reason for flattering the novelist).

One way of interpreting these examples is to say that the person involved just values a particular activity (problem solving, videogame playing, novel writing). But these cases seem more subtle. The goal of these activities (e.g. solving the math problem, finishing the game, producing a good novel) seems psychologically important. Failure to attain the goal could be much more disappointing than losing a friendly and informal game of soccer. On the other hand, it also seems clear that if solving the math problem could be had in an instant (by a click of the fingers) then it would lose its pull as a goal. Both the temporally extended process and the goal seem to be important in these cases.

The question of how exactly to characterize the source of value in these examples is not my task here. There may be models of human preference that fail to do this but still are able to make good predictions in most cases. The problem here is that it seems that models which define utility functions over world states rather than world histories might fail to make good predictions. In the videogame case, if the source of utility is seeing the ending sequence, then you’d predict that the player would skip to that sequence without finishing the game. If the source is the fun of playing the game itself, then you’d expect the player to keep playing if he learned that a bug in his version of the game would make it freeze three-quarters of the way through (removing the possibility of seeing the closing sequence). If utility comes from a particular causal sequence of events – the player successfully playing through the whole game and then seeing the ending sequence – then you need a utility function defined in terms of world histories. Switching to world histories is a challenge because representing sequences of events (rather than world states or bundles) leads to an exponential blow-up in the space of objects that utility can be defined over. Moreover, we need to have some rich formal language in order to correctly define the temporally extended complex occurrences that constitute important sources of value.

In conclusion: There are particular aspects of human preferences that present a challenge for existing microeconomic models. Some of these preferences (e.g. to produce
significant works of art or science or to bring about other difficult, time-consuming projects like starting a school, or some other organization) seem closely related to the idea of a meaningful human life. However, it is important to point out that the ethical significance of these preferences may not be matched by their practical significance. It may be that models that are not capable of representing this kind of preference can do well in the task of predicting human choices across many situations.

Section 2. An experiment testing the utility framework

2.1. Introduction: Psychological and Folk-Psychological Preference Learning

In Section 1, I introduced a formal framework for modeling preferences over bundles of goods that have some kind of dependency or interaction. I pointed out various everyday properties of human preference and the challenges they pose for the microeconomic formal framework. In Section 2, I describe ways of testing the formal framework empirically.

There are two main empirical applications of formal models of preference. The first application is to modeling and predicting people's actual choices. The second application is to modeling how people predict the choices of others (i.e. folk psychology or Theory of Mind). I will briefly sketch each application area.

**Modeling and predicting choices (first-order preference learning)**
Predicting actual choices is important for economics, where we'd like to predict how people's spending responds to changes in their budget or to changes in the cost of goods.
But predicting choices is also important outside of the economic transactions involving money that have traditionally been the focus in economics. For psychology and for the social sciences, we may be interested in the choices of a single individual or a very small group (rather than the large populations of individuals that have great significance in economics). We may also be interested in predicting the choices of children, and of people in communities where money plays a minor role. A topic of particular psychological and philosophical interest is the way people make choices between different kinds of goods. There is an extensive literature in psychology that looks at how altruistic people are in performing certain game-theory inspired tasks (Camerer 2003). There is also work suggesting that there are some goods that subjects consider “sacred” and treat differently from most goods (Tetlock 2003). Much of this work is informal and may be illuminated by a formal model that can represent a range of different ways that goods can be related.

Folk-psychology or Theory of Mind (second-order preference inference)

The second application of formal models of preference is as a model of human folk-psychology or ‘theory of mind’. Here we are interested not in predicting what humans will do, but in modeling and predicting the predictions that humans make about other humans. I will refer to the first application as the psychological or first-order project, and the second as the folk-psychological, theory of mind or second-order project. The study of folk-psychology is a significant subfield within cognitive science. This work ranges from qualitative studies of the development of concepts like false-belief in children (Gopnik and Meltzoff 1994) to more recent work on the neuroscience of folk-psychology (Saxe 2005) and quantitative modeling of some aspects of theory of mind within a Bayesian framework (Goodman et al. 2006, Lucas et al. 2008, Baker et al. 2010). Folk-psychology is an important topic for cognitive science because understanding the intentions of the people around us is so important. Children, in particular, are dependent on others for help and teaching in many areas of life and so need to quickly develop an ability to read the mental states of others.

This chapter focuses on one particular Theory of Mind task: modeling and inferring preferences in order to predict choices. This inference is important in quotidian contexts.
In social situations we often have to predict what kinds of things other people will like. Consider the task of preparing a dinner party, or selecting entertainment for an evening, or choosing fun activities for a child. It can also be socially important to work out someone's preferences (e.g. political or aesthetic preferences) from minimal observations of the person's actions. In business, companies try to work out what kinds of goods consumers would like to buy based on previous information about their buying habits. In politics, bargains can be made when each party is able to correctly identify something that the other party values. Due to the importance of this theory-of-mind task in daily life, there is reason to think that human abilities are quite developed in this area. This suggests that (a) humans may be able to make inferences that can only be matched by quite sophisticated formal models and (b) that studying human inference may be useful in coming up with richer formal models.

2.2. Turning utility functions into a model of preference inference

2.2.1. Bayesian inference for learning utility functions

In order to model preferences for psychology or folk-psychology, we will need more apparatus than just the bundle utility functions described in Section 1 of this chapter. If I have the utility function for an agent defined on the bundles $(A_1, \ldots, A_n)$ then I can predict his choice between any set of bundles that he could have as options. However, we want to consider situations where the utility function is not just given a priori but instead has to be learned from data. There are two different steps to learning bundle utility functions. We have to learn:

i.) The functional form that someone's utility function takes over a particular set of goods. For example, suppose Lucia is assembling a table. She purchasing screws in two different sizes: "big" and "small". It could be that the two sizes of screw are complements. For instance, if the table that Lucia is assembling requires two small screws for every one big screw, with additional screws being useless unless
they come in this ratio. On the other hand, Lucia could be building something where either small or big screws can be used and so the screw sizes are *substitutes*. We want to be able to learn from observing Lucia’s choices which utility function she has over these goods. Likewise, Bob might purchase coffee and milk and only be interested in producing milky coffee with a precise but unknown ratio of milk to coffee. Alternatively, Bob might only be interested in drinking milk and coffee on their own (as substitute drinks).

ii.) The *parameter settings* for the utility function. If the screws are complements, what is the mixing ratio $a$ for big vs. small screws? If coffee and milk are complements, what is the mixing ratio? (Is Bob making a latte, cappuccino, flat white or espresso macchiato?) If the screws are substitutes, how much better are the big screws than the small? (Operationally, how much more would you pay for a big screw than for a small screw, assuming that big screws are stronger?)

My approach is to use Bayesian inference to do this learning. We start with a prior distribution over the different forms the function could take (e.g. a prior over the three functional forms given by the complements, substitutes and Cobb-Douglas utility functions). We then learn an agent’s utility function by updating these distributions on observing choices. Both in the first- and second-order settings, it can be important to predict choices while only having limited information about the agent’s utility functions. This is easily achieved in a Bayesian framework. For instance, if I have a 0.7 probability that Lucia has a complements function over screws with parameter $a=2$, and a 0.3 probability that she has a substitutes function with parameter $a=1$, then I predict her choices over different bundles by averaging the predictions generated by each utility function.

### 2.2.2. Learning for individuals and for populations

The approach sketched so far allows us to learn utility functions for individuals from observing their choices. How well this model works for prediction will depend on whether any of the utility functions in our hypothesis space fits the individual’s actual
pattern of choice and on whether we observe choices that are informative about the individual’s general pattern.

So far, I have only discussed models for learning the choices of individuals. A very important and closely related inference problem is to learn about the distribution of preferences within populations. It is clear that there are regularities in preference across individuals. Within my formal framework, this can be on the level of functional form of a utility function, e.g. most people would treat printers and print cartridges as complements because they have no use for the parts separately but would derive value from a working printer. It can also be at the level of parameter settings of utility functions. For example, it could be that almost everyone finds two very similar brands of orange juice to be perfect substitutes. But in other areas we find polarization, with people strongly preferring one or the other item (e.g. rival sports teams).

We would like to learn population-level regularities from observing multiple individuals drawn from the population. Those regularities can then be exploited when making predictions about new individuals taken from the same population. For instance, if for some substitute goods there are polarized preferences, then a single observation of a new person should be enough to yield strong predictions about future choices. For example, if I see someone on a match day wearing the colors of one team, I can immediately make the inference that this person has a strong preference for that team.

2.2.3. Dealing with unreliability

A further important element of a practically useful model is some formal way of dealing with human error or the “random noise” associated with human actions. The bundle utility functions are a model for \textit{homo economicus}. They predict that even if two bundles differ by only a tiny amount in utility, the agent will still take the better one every time. Humans don’t seem to work like this. When two options are very closely matched for quality (e.g. two meals in a restaurant) we are often not sure which to pick and may pick different options at different times (i.e. pick inconsistently) and also recognize after the fact that our choice was slightly suboptimal.
In my modeling in this chapter, I use a simple model of noise in human judgments that has been developed by psychologists and decision theorists (Luce 1980). The model assumes that people are more likely to take suboptimal options when the difference in utility is small and become exponentially less likely as the utility difference decreases. This model of choice is called the “Luce choice rule”. Agents who choose according to this rule do not always take the option with maximum utility, but they are very unlikely to deviate when the difference between options is big. Hence we call them “soft maximizers” of utility. The idea of humans as soft-maximizers makes intuitive sense. Someone may choose his second-best option at a restaurant if the two are very closely matched. However, it’s very unlikely that someone with a serious nut allergy would pick something with nuts in it, assuming that they had complete knowledge of the ingredients of the dish while choosing.

People do take options that seem to be disastrously bad in retrospect. However, these often involve false beliefs or significant uncertainty on the part of the chooser. When they don’t, they may involve weakness of will or other near-term biases. The cases I will discuss in this section of the chapter don’t involve these complications.

2.2.4. Generalizing the Model

The framework described so far for learning utility functions from individuals and from populations is quite general. In Section 1, I described three different functional forms for bundle utility functions and described how these functions could be combined (e.g. decomposing one big utility into many independent functions of complements, substitutes, etc.). We could consider other kinds of functional forms for modeling how goods interact to produce utility. A general concern is the tractability of the functions. Some functions would make utilities time-consuming to compute. Other functions might have maximums or derivatives that are difficult to find. In any case, the Bayesian
framework can in principle handle a much larger class of functions than I have considered and learning and prediction will take place in exactly the same way\textsuperscript{20}.

\subsection*{2.2.5. Summary}

In this section I gave a brief description of a formal model for inference about preferences that combines microeconomic utility functions, the Luce choice rule for noise in human choice, and Bayesian inference for learning individual utility functions and population distributions over utility functions. I will refer to this model (complete with the Bayesian learning component) as the \textit{Utility Model} of human preference. This model was introduced in joint work with Leon Bergen (Bergen, Evans, Tenenbaum 2010). I will now describe some experiments we performed that test how well this model fits with human folk-psychological inferences about preference.

\section*{2.3. Empirical testing of the Utility Model for folk psychology}

\subsection*{2.3.1. Overview of our experiment}

In joint work with Leon Bergen, I tested the Utility Model using a simple experimental paradigm. Subjects were given a vignette in which an agent makes a choice between two different bundles. Subjects' task was then to make an inference about the agent's preferences based on their choice, and then to predict the agent's choice between a different pair of bundles. We ran the Utility Model on the same data (building in by hand

\footnote{\textsuperscript{20}There may need to be differences in how learning is implementing to deal with the computational overhead of a larger hypothesis space. But at the conceptual or mathematical level, there is no problem with adding more functional forms to our hypothesis space.}
relevant background information communicated to subjects by the vignettes). We then compared how the Utility Model’s predictions compared to subject predictions.

We only tested some aspects of the Utility Model. The model can learn utility functions over more than two goods and model choices over more than two options, but we only looked at two-variable functions and choices between two options. We also did not test whether the human ability to learn the form of utility functions is well captured by the Utility Model. In the experiment, we only tested subject’s ability to learn the mixing ratio $\alpha$ for utility functions they were assumed to already have. What the experiment did test was (1) whether people distinguish between different kinds of goods as the economic notions of complements and substitute would predict, and (2) whether the Bayesian learning model with soft-maximization matches human predictions. We also compared the fit of the Bayesian Utility Model to the fit of some simple heuristics. These heuristics are described below.

2.3.2. Stimuli

Our subjects were given 11 vignettes, in which agents were described choosing between bundles of goods. The vignettes were inspired by textbook examples from economics. Sample vignettes are shown below. Note that subjects were told that all bundles were identical in price. Hence none of the choices described involved trading off between cost and desirability.

**Complements Vignette [annotations in red]**

Patricia has just started a new restaurant and is decorating it. She is putting up large blinds on the windows, and she thinks she needs to use different length screws for different parts of the blinds. She went to the hardware store to buy screws for putting up the first blinds, and was given two offers on boxes of short screws and long screws.

(A) A box containing 4 long screws and 2 short screws

(B) A box containing 3 long screws and 3 short screws
Patricia chose option (A) and thought that was the best choice given her options.  
[Note: I'll call this OBSERVED CHOICE]

Patricia is putting up blinds on the rest of the windows, and needs to buy more screws. She goes back to the hardware store, and is given the following offers on boxes of screws. Both boxes (C) and (D) have the same price:
(C) A box containing 7 long screws and 3 short screws
(D) A box containing 2 long screws and 8 short screws?
Which option do you think Patricia should choose from (C) and (D)?  
[Note: I'll call this PREDICTED CHOICE]

Substitutes with Preference for Variety Vignette [annotations in red]

Last year Ted bought a small box of chocolates from a specialty chocolate store. The store makes boxes with two kinds of chocolates: chocolates with nuts and chocolates with caramel. The store had two small boxes available, which both cost the same amount:

(A) A box containing 3 chocolates with nuts and 3 chocolates with caramel
(B) A box containing 5 chocolates with nuts and 1 chocolates with caramel

[I won't include the PREDICTED CHOICE here, but it exactly parallels the above scenario, with Ted choosing between larger boxes of chocolates which in some cases have quite different ratios than in the observed choice.]

Substitutes without Preference for Variety Vignette [annotations in red]

Mary and her husband often have to go out at night. When they do, they hire a babysitter to look after their four-year old son. Mary has to schedule babysitters for the next month. Mary wants to make sure that her son gets along with the babysitter, and in general, she thinks that some babysitters do a better job than others. There are two babysitters who are usually available, Jill and Sarah. Neither can work all the nights that Mary needs.
Last month, Mary chose between two different hiring schedules. Both schedules would have cost her the same amount:

(A) Sarah for 4 nights and Jill for 2 nights
(B) Sarah for 3 nights and Jill for 3 nights.

[as above we leave out PREDICTED CHOICE].

For each of these 11 scenarios, we considered 16 different *numerical conditions*. That is, conditions under which we varied the numbers appearing in the observed choice and predicted choice. Thus the ratios between the goods and the absolute quantities were varied. This is relevant to discriminating between the different utility functions in our model's hypothesis space, which make similar predictions for certain bundles.

As is clear from the sample vignettes above, the goods that the agents are choosing would be familiar to our subjects. We assume that subjects would be able to use their background knowledge to work out how the goods would interact (e.g. whether they are complements or substitutes). Hence we gave our model the prior information about the utility function type but not about the mixing ratio for the function.

### 2.3.3. Alternative folk-psychological Theories

The performance of the Utility model was tested against non-Bayesian, non-utility based heuristics. We invented a number of candidate heuristics, inspired by the work of Gigerenzer and collaborators (1996). I will describe the two heuristics that performed best in terms of fitting with subject judgments:

- The *Crude Ratio Heuristic*, like the Complements utility function, assumes that there is an ideal ratio between the two goods $A$ and $B$, i.e. a ratio in which the goods most efficiently combine to produce utility. However, it is also “greedily” assumes that the ratio of goods in the bundle that the agent was observed choosing...
is the ideal ratio of goods, and that the agent will always choose the bundle that comes closest to this ratio. (In other words, the heuristic treats the observed choice as perfectly reliable evidence of the ideal ratio of the goods for the agent).

- The Max Heuristic (like the Substitutes utility function) assumes that one of the two goods is favored over the other and that the agent will choose whichever bundle has most units of this favored good.

Both heuristics make absurd predictions when the two options differ a lot in overall quantity (e.g. the bundles (A=4, B=2) vs. (A=15, B=16)). This is because they avoid doing a weighted sum or product and instead just work with features that can be extracted very quickly by humans. However, we did not include conditions in this experiment that tested the heuristics against the model on this apparent vulnerability.

2.3.4. Results

I will not discuss the results of the experiment in detail. The Utility Model generally produced good fits to human judgments. Fits were better when the Utility Model was assuming the intuitively correct functional form (e.g. complements or substitutes, depending on the vignette in question). Thus we got some evidence that subjects distinguish between goods in the way textbook microeconomics predicts. We got some weak evidence for a distinction between the “substitutes without preference for variety” and “substitutes with preference for variety” but plan to further examine whether people make this kind of distinction in future work.

The fit of the model supports the view that humans’ ability to infer preference is sensitive to size of bundles and the kind of interaction between goods (as the Utility Model would predict). It also shows that humans are sensitive to base rates and noise in the way that Bayesian inference would predict. However, more data is still needed to give full confirmation that the Utility Model does better than the crude heuristics. In terms of pure correlation, the two best performing heuristics did as well as the Utility Model. Thus the
current experimental evidence is compatible with the view that human judgments are based on very crude heuristics in this kind of situation. However, due to the absurd predictions of the heuristics when bundles differ in overall size, we expect that further experiments will show that the heuristics are not a plausible general mechanism for making these inferences.  

\[21\] It goes without saying that more research is needed.
Section 3. The Utility Model and a Computational Account of the Knobe Effect

3.0. Introduction

I have described some early experimental support for the Utility Model as an account of folk-psychological inferences about preferences in choice situations involving multiple goods. I now want to consider how this model could be applied as the core of a computational account of the Knobe Effect (Knobe 2003, 2004, 2006). I will assume familiarity with the classic Knobe vignettes in what follows (see Knobe references).

In the classic Knobe vignettes, subjects are asked to make a judgment about the mental states of an agent who has just been described making a choice that involves a trade-off between two different "goods" (e.g. helping/harming environment and profit). Thus there is some reason to think the Utility Model might be able to model some aspects of subject judgments. However, an immediate limitation is that the Utility Model makes inferences about preferences of an agent from observed choices, rather than about their intentions (or whether they brought about a result intentionally). Still, it seems possible that part of what drives the judgments about whether the agent intended a particular outcome is whether the agent had a strong preference for that outcome. One challenge in trying to apply the Utility Model to the Knobe effect is spelling out formally some possible relations between preferences and intention judgments (i.e. the judgments subjects make in Knobe vignettes as to whether an outcome was brought about intentionally). I will discuss this below.

22 Knobe has developed related experiments that focus on attribution of causality and other relations that are prima facie not intentional [2006]. I focus on the experiments that probe mental state attribution.
Another immediate challenge is how to construe the classic Knobe vignettes in terms of choices between bundles of goods. In the Chairman vignette (and other Knobe vignettes) the agent is presented as deciding whether to implement a new plan. Thus the choice is between “new plan” and “status quo”, where the status quo is not explicitly described at all. Moreover, the “new plan” is described only in the most vague and underspecified terms, without the precise numbers that the utility functions of the Utility Model require.

Consider one of the outcomes in the Chairman vignette, namely, the outcome of “harming the environment”. We could correctly describe a wide range of activities as harming the environment, from a family’s littering a national park to a logging company cutting down an enormous area of Amazon rainforest. Standard quantitative measures of preferences, such as surveys of people’s “willingness to pay” to prevent the damage, will assign values to these outcomes that vary substantially. In the context of the Chairman vignette, there are pragmatic reasons for readers to infer that the environmental harm is significant (otherwise it wouldn’t be worthy of presentation to the chairman). So one thing we need to be careful about in analyzing the original Knobe vignette is how these vague elements might be interpreted by human subjects.

3.1. Modeling the Knobe effect using the Utility Model

This section outlines my proposal for modeling the Knobe effect. My first observation is that plans or policies that have multiple, highly predictable effects can be seen as bundles of the outcomes. (If the effects are less predictable, we will need to switch to probability distributions over bundles, but I’ll stick to the highly predictable case). Choosing between a new plan and the status quo is thus a choice between two bundles. Preferences that generate choices over bundles can be modeled using bundle utility functions. By “observing” a choice between bundles (i.e. having a particular choice on behalf of the agent narrated to us), we can use Bayesian inference in the Utility Model to update a prior distribution on the mixing ratio \( \alpha \) that the agent’s utility function has for the two goods.
Thus, the observed choice of the agent allows us to make an inference about the agent’s preference between the goods.

The second part of the proposal concerns the link between preference inference and the judgment about one of the outcomes being intentional. Suppose that the two outcomes of a choice are $X$ and $Y$. The claim is that subjects will judge an outcome $X$ as more intentional if the choice that produced $X$ has given the subject information that the agent has a preference for $X$ relative to $Y$. I will discuss below various ways to make this claim more precise.

My account of the Knobe effect is very similar to the recent account given by Uttich and Lombrozo (2010) and also resembles other accounts in the literature (Machery 2008, Sripada 2010). Another similar account, which also involves a computational model, is found in recent work by Sloman et al (2010). The main thing that’s new in my account is a quantitative, utility-based model of preference which is integrated with a Bayesian model of inference. This allows for precise formulation of hypotheses in terms what kind of information the vignette gives subjects in Knobe experiments and how the information they get is used to generate attributions of intention.

My account is also more minimalist than the related theories in the literature. It is minimalist because while it invokes notions of preference and of the informativeness of different choices, it does not make use of norms or norm violations, harms and helps, or side-effects vs. main-effects. In contrast, on Uttich and Lombrozo’s view “harm” cases are judged more intentional because they involve violations of norms. Violations of norms, whether moral norms or conventional norms, provide more information about the violator than instances of conformity to norms. On my account, there is no special role for norms. The intention judgment depends on an inference about the agent’s preferences, and choices can be very informative about preferences without constituting a norm violation. Only a very thin or weak notion of norm plays any role in my account. For me, the strength of the intention judgment will depend on how much the agent’s choice shows his preferences to differ from the subject’s expectations. If the agent’s choice shows him to be very different from the population average, then this will probably lead to a strong

---

23 Holton (2010) also invokes an asymmetry between violating and conforming to a norm to explain the Knobe effect. However, his account centers on a normative asymmetry, rather than an asymmetry in the evidence these kinds of acts provide.
intention judgment. But merely differing a lot from the population involves violating a norm only in a very weak sense. (If someone loves a particular shade of green much more than most people in the population, she doesn’t thereby violate a norm in any strong sense). So my account predicts that vignettes that involve no norm violation (moral or conventional) should still be able to produce a Knobe effect.

Similarly, my proposal should not require any actions that constitute “harms” or “helps” to generate a Knobe effect. After all, there are other kinds of actions that provide information about preferences.

Finally, my account doesn’t make a distinction between the different elements or components of the bundle that are chosen, and so we should be able to get a Knobe effect without there being an outcome that is naturally described as a side-effect. To illustrate the role of an outcome being a side-effect, consider the Chairman vignette. In the vignette, environmental harm is a side effect of the profit-generating plan because it is not part of the chairman’s main set of goals and he would still be considering the plan even if it generated profit and had no environment effect.

There are some vignettes in the existing literature that generate Knobe effects and that can be interpreted (I won’t argue for that interpretation here) as differing from classic vignettes such as the CEO case by being minimal in the way I just described. There is the Smoothie case (Machery 2008) and an unpublished case mentioned in Sloman (2010). These cases are minimal in involving no moral or conventional norms, no harm/help and not necessarily any side-effect. Uttich and Lombrozo present vignettes with a Knobe effect that involve non-statistical, explicitly verbalized social conventions and side effects.

The vignettes I introduce below have two main aims. First, they provide a stringent qualitative test for the minimalist account, by leaving out norms, side-effects, and helps/harms. Second, they allow for something that has not been explored much in the literature. Since my model is computational and involves continuous parameters for probabilities and mixing ratios, the model can generate quantitative predictions. Up to this point, I’ve stated my account in a qualitative way. I’ve said that that an outcome X will be judged intentional when its being chosen gives information about preferences

---

24 Sripada et al. 2010 and Sloman et al. 2010 are recent exceptions.
towards X. However, there are various ways in which we could model how the strength of the intention judgment depends on the strength of the preference that the choice informs us of. Here's a simple quantitative model:

As the choice inference places more posterior probability on mixing ratios with stronger biases towards X, the strength of the intention judgment for the X outcome (as measured by the standard 1-7 scale) will increase.

I will discuss below some slightly different ways of connecting the information gained from the observation with the intention judgment. I have constructed two vignettes that aim to test the predictions of the Utility Model. I will first present the vignettes and then provide more detail on the different ways of modeling subject judgments using the Utility Model.

3.2. Experimental Stimuli

I designed two vignettes in order to test the qualitative and quantitative predictions of the Utility Model as applied to the Knobe effect. My account predicts that intention judgments will vary quantitatively with how informative the agent’s choice is about his preferences. So in my vignettes, agents make choices which give information about their preferences between the goods being traded off. The bundles in the vignettes are varied quantitatively to allow for quantitative variation in the amount of information choices provide. In contrast to most existing Knobe vignettes, the agents do not violate any strong social or conventional norms (such as harming the environment). Moreover, in my first vignette (the “Investment Vignette” below) the effects of choices are not described (or easily describable) as helps and harms and they are not naturally seen as side-effects of those choices.

I tested the vignettes as follows. In each vignette, an agent chooses between two bundles, with the bundles described using precise numbers. For instance, in the Investment Vignette, an agent chooses between investment strategies that have different expected short and long-term payouts. To test quantitative predictions, I varied one of the quantities in one of the bundles. In the Investment Vignette, I varied the long-term payout.
of the first investment strategy and held everything else fixed. This generates a series of
vignettes that vary only along this one dimension, enabling me to test precise quantitative
predictions.

1. Investment Vignette [annotations in red]

John is investing some of his savings. He is choosing between two eight-year investment
strategies. These strategies both pay out twice: first after four years and then again after eight
years. John's financial advisor predicts Strategy A will pay out $70,000 after four years and will
pay out another $100,000 after eight years. The advisor predicts that Strategy B will pay out
$90,000 after four years and another $80,000 after eight years. This is summarized in the table
below:

<table>
<thead>
<tr>
<th>Strategy</th>
<th>Payout after 4 years</th>
<th>Payout after 8 years</th>
<th>Total payout</th>
</tr>
</thead>
<tbody>
<tr>
<td>Strategy A</td>
<td>$70,000</td>
<td>X=$100,000 (90K, 95K, 100K, ... 160K)</td>
<td>$170,000</td>
</tr>
<tr>
<td>Strategy B</td>
<td>$90,000</td>
<td>$80,000</td>
<td>$170,000</td>
</tr>
</tbody>
</table>

John chooses Strategy B. Just as the advisor predicted, John makes $90,000 after four years and
makes another $80,000 after eight years.

John's payout decreased from $90,000 after four years to $80,000 after eight years. On a scale of
1-7, how much do you agree/disagree with the following statement:
"John intentionally caused his payout to decrease."

<table>
<thead>
<tr>
<th>Strongly disagree</th>
<th>Neutral</th>
<th>Strongly agree</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>4</td>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td>7</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

2. North-South Vignette [annotations in red]

Mr. Jones is the CEO of a major company. One day, his vice-president presents two different
plans for restructuring the company. The main difference between the plans is how they affect the two regional divisions of the company. The plans are predicted to impact profits as follows:

**Plan A**
+$10$ million rise in profit for Northeast Division of the company
-$1$ million drop in profit for South Division of the company

**Plan B**
+$6$ million rise in profit for Northeast Division of the company
-$0.5$ million drop in profit for South Division of the company

[variants: +$2m rise NE vs. -$2m drop S, +$4m rise NE vs. -$4m drop S, ...]

Mr. Jones decides to implement Plan A. The vice-president's prediction was correct: Plan A resulted in a $10$ million rise in profit for the Northeast Division and a $1$ million drop for the South Division.

After implementing Plan A, profits in the South Division dropped by $1$ milllion. On a scale of 1-7, how much do you agree/disagree with the following statement:
* Mr. Jones intentionally harmed the profits of the South Division. *

<table>
<thead>
<tr>
<th>Strongly disagree</th>
<th>Neutral</th>
<th>Strongly agree</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>$&lt;_{C}$</td>
<td>$&lt;_{C}$</td>
<td>$&lt;_{C}$</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td>$&lt;_{C}$</td>
<td>$&lt;_{C}$</td>
</tr>
</tbody>
</table>
3.3. Experimental Results

Results for the Investment Vignette are displayed below. When \( X \), the size of the second payout from Strategy \( A \), is less than $100K, the mean judgment is below 4.0 (labeled “neutral” on the questionnaire) and so people on average don’t see the outcome as being done intentionally. For \( X \) greater than $100K, the outcome is seen as (on average) intentional. Thus, as the vignettes vary quantitatively in this one variable, we get a gradual Knobe-effect switch from “unintentionally” to “intentionally” judgments. We get this same basic result from the North-South Vignette, although the average intention judgments are higher for the lowest values (almost at “neutral”).

Investment Vignette Results and Graph

<table>
<thead>
<tr>
<th>X in $1000s</th>
<th>90</th>
<th>95</th>
<th>100</th>
<th>110</th>
<th>130</th>
<th>140</th>
<th>160</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>2.35</td>
<td>3.41</td>
<td>4.00</td>
<td>4.50</td>
<td>4.69</td>
<td>4.88</td>
<td>5.00</td>
</tr>
<tr>
<td>SD</td>
<td>2.03</td>
<td>2.21</td>
<td>2.00</td>
<td>2.00</td>
<td>1.94</td>
<td>1.45</td>
<td>2.00</td>
</tr>
<tr>
<td># subjects</td>
<td>16</td>
<td>28</td>
<td>24</td>
<td>21</td>
<td>31</td>
<td>25</td>
<td>29</td>
</tr>
</tbody>
</table>
North-South Vignette Results

(North, South) profits in $m for Plan B

<table>
<thead>
<tr>
<th></th>
<th>(3, -3)</th>
<th>(6, -0.5)</th>
<th>(2, 2)</th>
<th>(4, 4)</th>
<th>(5, 5)</th>
<th>(8, 8)</th>
<th>(10, 10)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>3.80</td>
<td>3.89</td>
<td>5.04</td>
<td>5.37</td>
<td>5.46</td>
<td>5.78</td>
<td>5.44</td>
</tr>
<tr>
<td>SD</td>
<td>2.22</td>
<td>2.12</td>
<td>1.60</td>
<td>1.99</td>
<td>2.12</td>
<td>1.10</td>
<td>1.54</td>
</tr>
<tr>
<td>n</td>
<td>40</td>
<td>37</td>
<td>28</td>
<td>29</td>
<td>25</td>
<td>31</td>
<td>33</td>
</tr>
</tbody>
</table>

3.4. Modeling the Knobe Effect

The results discussed above provide some confirmation for the qualitative predictions of the Utility Model account of the Knobe effect. I will now describe some different ways of modeling the effect quantitatively. The idea is simple. To predict intention judgments, we
first model the inference subjects make about the preferences of the agent based on his choice. This inference is the basis for the intention judgment.

3.4.1. Preference Inference

We use the Utility Model to model the preference inference. In doing so, we assume subjects model the agent as having a certain bundle utility function\textsuperscript{25} over the set of bundles that are available. Of the three utility functions discussed above, the Substitutes function is the obvious choice for both vignettes. For the Investment Vignette, we are asking “How much does someone value money earned in 4 years vs. in 8 years time?” and “What is the marginal value of more money in 4 years or 8 years?”. These questions are important in economics. I will sidestep these questions and start by assuming that subjects have a very simple model of the agent’s utility function\textsuperscript{26}. I will use the standard additive Substitutes function (without including a parameter for diminishing marginal utility). Thus we have:

\[ U(S,L) = S + \alpha L \]

Here \( S \) is the short-term (4-year) gain and \( L \) is the long-term (8-year) gain. The mixing ratio \( \alpha \) can be seen as a measure of how much one discounts long-term gains. This will vary between individuals. Hence we expect subjects to start with a probability distribution over the value for \( \alpha \). On observing the agent’s choice, they will update this distribution based on the evidence that the choice provides.

Here’s an illustration. We start with a symmetric prior distribution on \( \alpha \) (in blue on the graph below), with mean \( \alpha=0.5 \) and small variance. Thus we have a prior expectation that the agent will either assign the same weight to short- and long-term gains, or have a

\textsuperscript{25} It would be easy to model subjects as having a probability distribution over utility functions the agent might have.

\textsuperscript{26} We are interested in folk-psychology or folk-economics here, rather than economics. Still, it’s plausible that subjects have a more sophisticated model of the utility function than the one we present here. As discussed above, the general framework presented here could incorporate more sophisticated models.
small bias in one direction. We observe the agent choosing the ($90K, $80K) Strategy B over the ($70K, $120K) Strategy A. This gives strong evidence that the agent is discounting long-term payouts by a substantial amount, and so the posterior (in black on the graph) shifts over towards 1.

3.4.2. Relating Intention to Preference

There are various ways that subjects' inference about the agent's preference (say for short- or long-term gains) could be related to their intention judgment. The basic idea is that outcomes will be judged more intentional if they are brought about by (or flow from) a (strong) preference. This can be made precise in different ways:

1. **Model 1: Absolute preference (with either graded or binary preference)**

   The outcome is more intentional if the posterior expectation for $\alpha$ (i.e. the mixing ratio which measures the magnitude of the preference) is higher (graded account).

   Or, on the alternative binary account, if $P(\alpha > 0.5)$, the subjective probability that

---

27 It may be odd to consider someone who values long over short-term gains. But it's not obvious that folk-psychology would rule this out a priori in the way economists might. (One might prefer long-term gains to make sure that money is available long-term even if short-term gains are squandered.)
the agent is biased\textsuperscript{28} in a particular direction, is high. For both of these accounts, what matters is the absolute preference of the agent and not how the agent differs from a statistical norm.

2. Model 2: Relative preference

The outcome is more intentional if the posterior for $\alpha$ differs more from the subjective prior or from the norm. Here what matters is whether weighting $\alpha$ is unusually far from the prior or norm, i.e. whether the preference for one good over the other is unusually strong. Obviously this measure will depend on what the prior/norm is. If we are dealing with a CEO, is the norm/prior relative to all adults or relative to all business people?

I will discuss further modeling questions below. Before doing so, I want to illustrate the kind of predictions we can get from the models discussed so far. I generated predictions for Model 2 (above) using various weak assumptions about the relevant prior probabilities and noise parameters\textsuperscript{29} against the subject data from the Investment Vignette. I include a linear and a log version of Model 2, both of which are graphed below.

\textsuperscript{28} The bias could be arbitrarily small. What matters on this account is the strength of the belief that the agent has some bias.

\textsuperscript{29} There is not space to discuss these assumptions in detail. One thing to note is that how well the model fits the human data is not dependent on the precise values I ended up using.
Diagram 5 (above)

Scatterplot showing model predictions for Model 2 vs. aggregated subject judgments for the Investment Vignette. Model predictions (which are a measure of the change in estimated preferences for short vs. long-term gains) on the x-axis and subject intention judgments on y-axis. Linear correlations are 0.87 and 0.98 for linear and log models respectively.
Let's return to the question of how to model intention judgments using the Utility model. The experiments based on the two vignettes shown above do not allow us to easily differentiate between the two basic models (absolute and relative) described above. We could distinguish by measuring people's priors over the preferences of the agents involved in various vignettes before letting them "observe" the agent making a choice. We could also design new vignettes in which background knowledge about preferences for the agents involved is controlled.

The distinction between absolute and relative preference is just one kind of refinement we need for a complete model. Here are some further questions that a model needs to decide:

- Does the choice itself have to provide information about the deviation from the prior in order to be judged intentional? (For example, would the agent in the Investment Vignette's heavy discounting of long-term gains be judged intentional even if you already knew he heavily discounted long-term gains?).

- Does the choice itself have to be counterfactually dependent on the preference? (For example, if we know an agent heavily discounts long-term gains, does it look intentional when he takes an option with trivial long-term gains when his alternative option also had equally trivial long-term gains?).

A different set of questions concerns the broadness of applicability of the Utility Model as a model of choice and preference inference. One reason the classic Knobe results are interesting is that they suggest a surprising connection between moral judgment and intention judgment. My account of the Knobe Effect does not give any special role to moral or conventional norms. Thus, on a formal level, it would treat a preference for (say) protecting the environment as the same kind of thing as a preference for long-term vs. short-term wealth. That is, one's preference for helping the environment should be commensurable with preferences for satisfying occupational norms (e.g. making profit) and "egocentric" preferences (e.g. eating nice food or socializing). For instance, even a staunch environmentalist would be willing to accept (at least as revealed by behavior) some harm to the environment for huge benefits in other areas. This is a psychological (or first-order) question about human preference, rather than a question of folk-psychology.
But if predicting human choices requires treating moral attitudes as a distinct kind of preference, then this psychological question could be important in giving a full account of folk-psychological phenomena like the Knobe effect.
References for Chapters 1 and 2


Botvinick M, Toussaint M. 2012. “Planning as Inference”.


