How many kinds of reasoning? Inference, probability, and natural language semantics

Daniel Lassiter, Noah D. Goodman
Stanford University

October 28, 2014

To appear in *Cognition*: please consult final version at
http://www.journals.elsevier.com/cognition/.

Abstract

The "new paradigm" unifying deductive and inductive reasoning in a Bayesian framework (Oaksford & Chater, 2007; Over, 2009) has been claimed to be falsified by results which show sharp differences between reasoning about necessity vs. plausibility (Heit & Rotello, 2010; Rips, 2001; Rotello & Heit, 2009). We provide a probabilistic model of reasoning with modal expressions such as "necessary" and "plausible" informed by recent work in formal semantics of natural language, and show that it predicts the possibility of non-linear response patterns which have been claimed to be problematic. Our model also makes a strong monotonicity prediction, while two-dimensional theories predict the possibility of reversals in argument strength depending on the modal word chosen. Predictions were tested using a novel experimental paradigm that replicates the previously-reported response patterns with a minimal manipulation, changing only one word of the stimulus between conditions. We found a spectrum of reasoning "modes" corresponding to different modal words, and strong support for our model's monotonicity prediction. This indicates that probabilistic approaches to reasoning can account in a clear and parsimonious way for data previously argued to falsify them, as well as new, more fine-grained, data. It also illustrates the importance of careful attention to the semantics of language employed in reasoning experiments.

Keywords: Reasoning, induction, deduction, probabilistic model, natural language semantics.

Suppose that you have learned a new biological fact about mammals: whales and dogs both use enzyme B-32 to digest their food. Is it now *necessary* that horses do the same? Is it *plausible*, *possible*, or *more likely than not*? Expressions of this type—known as *epistemic modals* in linguistics—have played a crucial, though neglected, role in recent work that argues for a sharp qualitative distinction between inductive and deductive modes of reasoning. In the paradigm introduced by Rips (2001) and extended by Heit and Rotello (2010); Rotello and Heit (2009), participants are divided into two conditions and are either asked to judge whether a conclusion is

"necessary" assuming that some premises are true, or whether it is "plausible". The former was identified with the deductive mode of reasoning, and the latter with the inductive mode.

These authors asked participants in both conditions to evaluate a variety of arguments, some of which were valid under classical logic and some invalid under classical logic, or contingent. An example contingent argument might be "Cows have sesamoid bones; Mice have sesamoid bones; therefore, Horses have sesamoid bones". An example valid argument might be "Mammals have sesamoid bones; therefore, horses have sesamoid bones." They found that there was a non-linear relationship between the endorsement rates of arguments depending on condition: participants in both conditions generally endorsed logically valid arguments, but participants in the deductive condition were much less likely to endorse contingent arguments than those in the inductive condition. These results were interpreted as a challenge to theories of reasoning which rely on a single dimension of argument strength and interpret deductive validity as simply the upper extreme of this dimension, guaranteeing maximal strength (Johnson-Laird, 1994; Osherson, Smith, Wilkie, Lopez, & Shafir, 1990). In particular, Rips and Heit & Rotello argued that non-linearities cannot be accounted for by probabilistic theories of reasoning, which identify the strength of an argument with the conditional probability of the conclusion given the premises (Heit, 1998; Kemp & Tenenbaum, 2009; Oaksford & Chater, 2007; Tenenbaum, Griffiths, & Kemp, 2006). On the other hand, they claimed that the results are consistent with two-process theories of reasoning (Evans & Over, 1996).

We argue that the key manipulation of previous experiments hinges not on a psychological distinction between two reasoning processes, but rather on facts about the semantics of modal words such as "necessary" and "plausible", which can be modeled using a single underlying scale of argument strength—conditional probability. We propose an account of reasoning with epistemic concepts, motivated by recent work in linguistic semantics, which predicts non-linear response patterns while retaining a single dimension of argument strength. The model predicts a spectrum of response patterns depending on the modal language used, not only the two corresponding to "necessary" and "plausible". It also makes strong monotonicity predictions that would be very surprising under two-dimensional theories of reasoning.

We tested the claim that the modal word is the crucial factor using a new paradigm that isolates its effects. Our arguments had the same form as those used in previous research, except that we placed a modal word of interest in the conclusion:

> **Premise 1:** Cows have sesamoid bones.
> **Premise 2:** Mice have sesamoid bones.
> **Conclusion:** It is {plausible/necessary/possible/likely/probable/certain} that horses have sesamoid bones.

We will refer to configurations such as "It is plausible/possible/etc. that *C*" as a **modal frame**. If varying the modal frame gives rise to the non-linear pattern of responses found in previous work, this would indicate that an explanation of these results could be derived from the meaning of these modal words.

Together, the model and experimental evidence indicate that the negative conclusions of previous work regarding one-dimensional theories of argument strength are not warranted: it is possible to explain non-linear response patterns with a probabilistic account of argument strength. Indeed,

non-linearities are to be expected given independently motivated facts about the semantics of epistemic modals. In addition, we confirm the monotonicity prediction, lending direct support to a one-dimensional, probabilistic theory.

1 Deduction and Induction

There is a long history of reflection on the relationship between inductive and deductive reasoning (Gabbay & Woods, 2009). In philosophy, it has long been assumed that the two are distinct. Deductive reasoning has generally been thought of as involving mathematically well-defined procedures for drawing out consequences which follow *with certainty* or *of necessity* from some body of evidence. Induction, on the other hand, involves going beyond the available evidence, drawing conclusions that are likely or plausible but (often) not certain to be true. The traditional characterization has thus generally taken for granted that induction and deduction are different, with deduction occupying a position of pride in mathematical and logical reasoning. Since David Hume's seminal work on the topic, induction has generally been viewed very differently, as a psychological process which is practically necessary for humans but lacking in formal justification (see Vickers, 2012).

Deductive reasoning is frequently assumed to be relevant to psychologically interesting phenomena such as conditional and syllogistic reasoning. For example, many psychological models of syllogistic reasoning have explicitly invoked some standard mathematical formulation of deductive inference, either syntactic/proof-theoretic (Rips, 1994) or semantic/model-theoretic (Johnson-Laird, 1986) in character.

With the advent of Bayesian approaches to inductive inference in philosophy (Ramsey, 1931; Earman, 1992) and psychology (Osherson et al., 1990; Oaksford & Chater, 1998; Tenenbaum, Kemp, Griffiths, & Goodman, 2011), the theoretical gap between these two types of reasoning has begun to narrow. It has often been observed that the validities of classical logic can be recovered as a limiting case of standard probabilistic inference, where attention is restricted to arguments in which the conclusion has conditional probability 1, given the premises, no matter what background beliefs a particular individual may have (e.g., Jaynes, 2003).

Oaksford and Chater (2001, 2007) have argued that there is no psychologically relevant distinction between induction and deduction: the reasoning strategies that are employed by participants in experiments designed to test their deductive reasoning skills are the same that are employed for everyday inductive reasoning tasks. Oaksford & Chater argue that both kinds of reasoning are well-modeled if we assume that reasoners maintain probabilistic models of the world around them and reasoning proceeds by conditioning on the data received—i.e., by Bayesian inference. In the simplest case on which we will focus here (where the premises do not contain conditionals, and are taken as known facts) this account treats deductive reasoning involving a set of premises P in the following way: participants evaluate potential conclusions C by finding the conditional probability of C in their (subjective) model, on the assumption that the premises are true. Given a choice between "valid' and "invalid" participants should respond "valid" if this conditional probability is greater than the relevant response threshold, in the absence of intervening factors. If the response threshold is 1, then the choice will mirror the judgments predicted by a theory based on explicit

deductive reasoning.[1] On this account, then, the distinction between inductive and deductive reasoning is not a qualitative distinction of different modes or processes, but merely a quantitative difference in response threshold.

Note that the above characterization of probabilistic approaches to reasoning is simplified in two important ways. First, everyday reasoning critically handles situations in which the inputs to reasoning, the premises, are themselves uncertain. There are various ways to generalize basic probabilistic reasoning to situations involving uncertain information, notably Jeffrey conditionalization (Jeffrey, 1965, §11) and hierarchical modeling (Pearl, 1988, §2.2.1). It remains to be seen which of these alternatives, if any, is empirically adequate: useful perspectives on this issue are provided by Hadjichristidis, Sloman, and Over (2014); Pfeifer (2013); Stevenson and Over (2001); Zhao and Osherson (2010). Second, reasoning from conditional premises (Oaksford & Chater, 2007; Pfeifer & Kleiter, 2010) is an important task at the intersection of semantics and psychology of reasoning, which may require some amendment to the simple conditional probability approach. It is controversial what semantics is appropriate for conditional sentences, and even whether they can be assigned truth-values and conditioned on. We focus in this paper on the case of simple non-conditional sentences presented as facts, with the expectation that the contributions of this paper will be synergistic with extensions to handle uncertain and conditional premises. (Indeed, conditional reasoning is one of the core cases discussed by Oaksford and Chater (2001, 2007) in the context of presenting a model along these lines.)

2 Difficulties for a Unified Model

Rips (2001) conducted an experiment designed to investigate whether there is a qualitative distinction between deductive and inductive reasoning. Participants in two groups were asked to judge arguments either according to whether the conclusion was *necessary* (assuming that the premises were true) or whether it was *plausible*. The mode of judgement was explained in each condition by extensive task instructions. Most participants in both conditions accepted valid arguments and rejected contingent arguments whose conclusion was not causally consistent with the premises, such as "Car X strikes a wall, so Car X speeds up". However, participants differed by condition in whether they rejected non-valid arguments which were causally consistent with the premises: those in the inductive condition generally accepted arguments such as "Car X strikes a wall, so Car X slows down", while those in the deductive condition did not.

Rips argued that this result falsifies theories of reasoning in which argument strength is a one-dimensional quantity such as conditional probability: "[i]f participants base all forms of argument evaluation on the position of the argument on a single psychological dimension, then induction and deduction judgments should increase or decrease together" (p.133). However, Rips' discussion

[1] Note however that not all arguments that have conditional probability 1 are classically valid; e.g., the probability that a real number selected uniformly at random from the interval $[2,3]$ is not π is 1, but this is obviously not a deductively valid conclusion (Rips, 2001). It is not clear whether this mathematical fact is relevant to everyday reasoning problems, though. If it is, then the empirical predictions of the two accounts would be slightly different when the response threshold is 1; however, the model and experiments we present below suggest that enforcing a maximal response threshold in this way may be practically impossible.

made clear that the argument against a unified model goes through only on the assumption that subjects are making judgments by comparing the strength of each argument to response thresholds which are either deterministic or stochastic with equal variance. (This assumption is important since the model we sketch below does not have this feature: see the next section.)

Heit and Rotello (2010); Rotello and Heit (2009) extended Rips' paradigm in a number of ways, providing a new method of quantitative analysis for these effects. Their core finding was that d', a standard measure of sensitivity in Signal Detection Theory (SDT), was significantly higher in the deductive condition across a variety of arguments types and manipulations. d' is defined as $z(H) - z(F)$, the difference between the z-scored hit rate H and false alarm rate F (Macmillan & Creelman, 2005). This difference means that participants in the inductive condition were less sensitive to argument validity than participants in the deductive condition (see Table 1). Differential sensitivity indicates that the difference between conditions is not simply a shift in

	Deduction	Induction
Acceptance, valid	.94	.95
Acceptance, contingent	.06	.17
Sensitivity (d')	3.31	2.56

Table 1: Acceptance rates and d' in Experiment (1a) of Rotello & Heit 2009 (three-premise arguments only).

response criterion, assuming equal-variance noise for different argument types. Thus we cannot fit a one-dimensional SDT model to such results. In accord with Rips, these authors claimed that the non-linear relationship between validity and condition is a challenge to probabilistic theories of reasoning. They argued that the results were better captured by a two-dimensional SDT model with possibly orthogonal dimensions of inductive strength and deductive validity, in which the response criterion can vary in both dimensions.

3 The Probability Threshold Model

In a brief comment on the line of work just discussed, Oaksford and Hahn (2007, p.276) suggested that the results "tell us more about the semantics/pragmatics of the words 'necessary' and 'plausible', where it has been known for some time that they don't neatly carve up some underlying probability scale". In this section we propose a model of the relationship between probability and epistemic concepts such as certainty, necessity, and plausibility which makes good on this suggestion. The model is a development of recent work in the formal semantics of natural language. It predicts the possibility of non-linear response patterns and variation in d' depending on the modal language used, and makes a novel monotonicity prediction which we will test in upcoming sections.

3.1 Model description

Our probabilistic model of reasoning with epistemic modals is intended to capture the following intuitions. A maximally strong conclusion *C* remains maximally strong whether you ask if it is *possible, plausible, likely* or *necessary*; a maximally weak conclusion remains maximally weak under the same conditions; but there is much more flexibility for uncertain conclusions. If *C* has a probability of .4, it presumably will count as *possible* and perhaps as *plausible*, but it would not seem to be *likely* and surely not *necessary* or *certain*. Thus the effect of an epistemic modal on a conditional probability should be a transformation that preserves the minimum value 0 and the maximum value 1.

As it happens, recent work on the semantics of epistemic modals contains all the components of a model which has these features: putting these components together, we call this the **Probability Threshold Model** (PTM). The crucial modals *necessary* and *plausible*—among others such as *certain*, *likely*, *probable*, and *possible*—fall into the grammatical class of **gradable adjectives**, as discussed by Lassiter (2010, to appear); Portner (2009); Yalcin (2007, 2010) among others. Like other gradable adjectives, they can be used in comparative and degree modification structures: for example, conclusion *C* might be *more plausible* than *C′* given some evidence, or *C* might be *very likely* or *almost necessary*. This corresponds, for instance, to one person being *taller* than another, and a glass of water being *very large* or *almost full*. Gradable expressions are generally treated in formal semantics as functions which map objects (individuals, verb phrase meanings, propositions, etc.) to points on a *scale* and compare them to *threshold values* (Kennedy, 2007). For example, the meaning of *tall* is a function which compares an individual to a height threshold, returning True if the individual's height exceeds the threshold: *Al is tall* is true if and only if Al's height exceeds threshold θ_{tall}. (For linguistically-oriented surveys of adjective semantics see Lassiter, in press-a; Morzycki, to appear.)

The location of the threshold is often uncertain, a fact which is closely related to the problem of vagueness (Williamson, 1994). We follow a number of recent proposals (Edgington, 1997; Frazee & Beaver, 2010; Lassiter, 2011; Lassiter & Goodman, 2013; Lawry, 2008; Schmidt, Goodman, Barner, & Tenenbaum, 2009; Verheyen, Hampton, & Storms, 2010) by assuming a distribution on thresholds $P(\theta)$. Thus, for instance, the probability that *Al is tall* is equal to the probability that the *tall*-threshold value falls below Al's height.

$$P(Al\ is\ tall) = P(height(\mathrm{Al}) > \theta_{tall}) = \int_0^{height(\mathrm{Al})} P(\theta_{tall})\, d\theta_{tall} \tag{1}$$

We do not make any assumptions here about how threshold distributions are derived, but see Lassiter & Goodman, 2013 for a proposal to derive them from pragmatic reasoning, which makes crucial reference to prior knowledge and utterance context.

Recent work suggests that a threshold semantics is appropriate for gradable epistemic modals as well, and that probability is a good choice for the underlying scale (Lassiter, 2010, in press-b, to appear; Yalcin, 2010). We can summarize this model by treating epistemic modals such as *plausible* and *necessary* as placing conditions on the probability of their sentential arguments, for instance:

$$It\ is\ plausible\ that\ q \Leftrightarrow P(q) > \theta_{plausible} \tag{2}$$

We will assume that the modals employed in our experiment below — *plausible, necessary, possible, probably, likely*, and *certain* — all have lexical entries with this schematic form.

$$\textit{It is M that } q \Leftrightarrow P(q) > \theta_M \tag{3}$$

Since these thresholds are uncertain, we assume a distribution over thresholds for each modal. Thus, for any such modal θ_M,

$$P(\textit{It is M that } q) = \int_0^{P(q)} P(\theta_M)\, d\theta_M. \tag{4}$$

We extend this analysis from modalized sentences to arguments with modalized conclusions by assuming that the modal takes scope over the underlying argument. That is, if argument a is of the form "P_1; P_2; so, it is M that C", we interpret it as "It is M that q" where q is "P_1; P_2; so, C". As described above, we take the underlying probability for such an argument to be the conditional probability of the conclusion given the premises, $P(q) = P(C|P_1, P_2)$. From Equation 4, we can then define:

$$\textbf{strength}_M(a) \quad = \int_0^{P(C|P_1,P_2)} P(\theta_M)\, d\theta_M. \tag{5}$$

One elaboration is needed: a speaker whose subjective probability of rain is .5 may not choose to assert "It's raining" or "It's not raining" with equal probability. Similarly, participants in a reasoning experiment may reject arguments of the form "P_1; P_2; so, C" if $P(C|P_1, P_2)$ is only .5, rather than accepting or rejecting at chance. That is, the strength of an unqualified argument "P_1; P_2; so, C" itself should not be treated as the conditional probability of the conclusion given the premises. Instead, we assume that there is a silent modal and a corresponding assertibility threshold θ_{assert} (Lewis, 1979; Davis, Potts, & Speas, 2007) and distribution $P(\theta_{assert})$, which determine $\textbf{strength}_{\varnothing}(a)$ as above.[2]

3.2 Predictions

We assume that participants' choices in a reasoning experiment will reflect the argument strength with some decision noise: Participants' inclination to agree with argument a_i in a forced-choice setting is governed by equation 5 together with a noise parameter ε, interpreted as the proportion of trials in which participants choose a response at random.

$$P(\text{"agree"}|a_i) = \textbf{strength}_M(a_i) \times (1 - \varepsilon) + \frac{\varepsilon}{2} \tag{6}$$

So interpreted, the PTM makes two predictions that will be important in discussion below. First, it predicts a monotonicity property that we will call **No Reversals**: for any two modals M_1 and M_2 which fit the lexical schema of equation (3), and for any two arguments a_1 and a_2, if $\textbf{strength}_{M_1}(a_1) > \textbf{strength}_{M_1}(a_2)$ then $\textbf{strength}_{M_2}(a_1) \geq \textbf{strength}_{M_2}(a_2)$. (The Appendix gives

[2] Interestingly, in the experiment reported below the pattern of responses to unmodalized arguments is most similar to the pattern in the *necessary* and *certain* conditions, indicating that $P(\theta_{assert})$ is similar in shape to $P(\theta_{certain})$ and $P(\theta_{necessary})$: see Figures 3 and 4 below.

a proof of the No Reversals property.) For instance, it should not be possible for a_1 to be more *likely* than a_2 while a_2 is more *plausible* than a_1. Notably, a two-dimensional model of argument strength like that of Rotello and Heit (2009) does not make this prediction: with two orthogonal dimensions of strength and a response criterion which can vary in both dimensions, there can easily be arguments a_1, a_2 such that a_1 is stronger by criterion c_1 but a_2 is stronger by criterion c_2. For two-dimensional theories, it would therefore be a very surprising coincidence if we find monotonic results under different "kinds of reasoning": non-monotonicity would be the usual situation.

The second important property of this model is that it predicts the possibility of non-linearities of the type described by Heit and Rotello (2010); Rips (2001); Rotello and Heit (2009) and discussed above. To illustrate, we focus on a particularly simple instantiation of the PTM in which the threshold distribution is a power law:

$$P(\theta_M) = \alpha_M \theta^{\alpha_M - 1}, \quad \alpha_M \in \mathbb{R}^+. \tag{7}$$

It is easy to show that for this power law distribution,

$$\textbf{strength}_M(a_i) = P(Conclusion(a_i)|Premises(a_i))^{\alpha_M}. \tag{8}$$

Depending on α_M we get a variety of possible curves, a few of which are sketched in figure 1a (setting $\varepsilon = .1$ for illustrative purposes). The power-law model illustrates how the PTM can account for non-linearities: α_M does not greatly influence the response probability for very high or low confidence, but there is significant variation in predicted response probabilities in the middle range depending on α_M. This feature leads to a prediction that the choice of modal will have less influence on response rates for arguments with very high strength (e.g., logically valid arguments) than on those with intermediate strength.

In particular, d' will vary depending on α_M. Suppose for illustration that the mean strength of logically contingent arguments in some sample is .5, and that the strength of logically valid (inclusion or identity) arguments is 1. The d' statistic estimated from this data should then be (on average)

$$\begin{aligned} d' &= z(1^{\alpha_M}(1-\varepsilon)+\varepsilon/2) - z(.5^{\alpha_M}(1-\varepsilon)+\varepsilon/2) \\ &= z(1-\varepsilon/2) - z(.5^{\alpha_M}(1-\varepsilon)+\varepsilon/2) \end{aligned} \tag{9}$$

If $\varepsilon = .1$, we expect the observed d' to be related to α_M as in figure 1b. The value of the d' statistic is not predicted to be constant in our probabilistic model, but should depend on the choice of M. Thus, a model with one dimension of argument strength (conditional probability) is able to predict non-linearities of the type previously claimed to be problematic for probabilistic accounts.

4 Experiment

Our experiment modified the design of Rotello and Heit (2009) in order to test directly the predictions of the Probability Threshold Model: the specific non-linear response pattern observed should be controlled by the choice of modal frame, and argument strength ordering should be preserved between modal frames (No Reversals). Previous experiments in this vein attempted to induce participants to employ different modes of reasoning using extensive task instructions

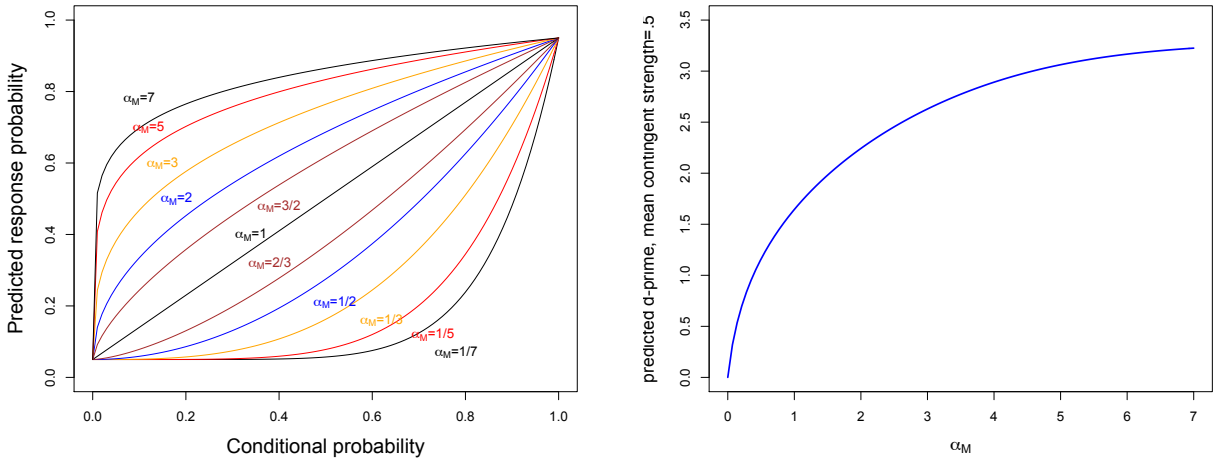

Figure 1: (**a**): Predicted response probability for various settings of α_M under the power-law version of the PTM. (**b**): Example of relation between α_M and d'.

that were manipulated between participants. This set-up makes it difficult to isolate the source of the observed non-linearities. Because we hypothesize that these effects are due to the semantics of the modal expressions employed, we also predict that it should be possible to replicate non-linearities within participants, using a very minimal manipulation within the stimuli and keeping the instructions constant: we varied only the choice of modal *M* within the stimulus sentence. This minimal manipulation allowed us to isolate the effect of the modal frame by examining acceptance rates across conditions and for individual arguments of varying strengths, regardless of condition. Because the task instructions were not manipulated it is possible to use a (more powerful) within-participants design. We also used more than two epistemic modal expressions, including *possible, likely, probable*, and *certain* in addition to *plausible* and *necessary*. The inclusion of additional modal vocabulary could allow us to observe more than the two "modes" of reasoning studied in previous work. We maintained other relevant features of the design of Rotello and Heit (2009), including the division of arguments into a "Contingent" condition and two "Valid" conditions, namely "Identity" and "Inclusion".[3]

[3]In Rotello & Heit's terminology, our Contingent arguments were "Invalid", and our Identity and Inclusion arguments were sometimes aggregated as "Valid". We avoid the term "Invalid" here because it sometimes connotes a normative judgment to the effect that such arguments ought to be rejected. Note also that we do not intend to make any claims about the ecological validity of the somewhat odd Identity arguments or theoretical claims about whether "reasoning" per se is involved in their evaluation: the goal is rather to replicate previous results using a modified paradigm which clarifies the source of the patterns observed.

Contingent:	*Identity*:	*Inclusion*:
Cows have enzyme X.	Horses have enzyme X.	Mammals have enzyme X.
Seals have enzyme X.	Cows have enzyme X.	Cows have enzyme X.
So, it is *M* that horses have enzyme X.	So, it is *M* that horses have enzyme X.	So, it is *M* that horses have enzyme X.

Figure 2: Structure of sample arguments in our three conditions.

4.1 Methods

4.1.1 Participants

507 participants were recruited using Amazon's Mechanical Turk platform, with a restriction to participants located in the United States. They were compensated for their participation.

4.1.2 Materials

Overall there were three argument types—contingent, identity, and inclusion—and seven modal frames—no modal, "possible", "plausible", "likely", "probable", "necessary", "certain". See Figure 2 for an overview of the argument structures.

The properties used in the arguments were chosen from a list of 21 unfamiliar biological and pseudo-biological properties adapted from Osherson et al. (1990), such as "use enzyme B-32 to digest food", "have sesamoid bones", "have a blood PH of 7.9", and "use acetylcholine as a neurotransmitter". No property was seen twice by any participant.

As in Osherson et al., 1990, the animal in the conclusion was always "horses". There were nine other animals that could appear in the premises: "cows", "chimps", "gorillas", "mice", "squirrels", "elephants", "seals", "rhinos", "dolphins". Premises in the contingent condition used distinct pairs of these nine animals. Premises in the inclusion condition used one of these animals and either "animals" or "mammals". Premises in the identity condition used one of these animals and "horses". The order of premises was randomized; neglecting order there were thus 63 specific premise patterns.

In each argument, the conclusion was either in the "no modal" frame—"So, horses have *X*"—or in the modal frame "So, it is *M* that horses have *X*", where *M* was one of: "possible", "plausible", "likely", "probable", "necessary", "certain".

4.1.3 Procedure

Participants saw 21 arguments, one with each target property in randomized order. Each participant saw each modal frame three times: three blocks were generated for each participant, with one modal frame occurring in each block in random order. Argument types were assigned to trials such that each participant saw 16 contingent, 2 identity, and 2 inclusion arguments, and one additional argument (to fill out the three blocks of modal frames) which was contingent 80% of the time, identity 10%, and contingent 10%. These argument types were randomly assigned to trials and

were fleshed out by choosing specific premises in a pseudo random fashion. (Specifically, the nine animals were randomly drawn as needed until exhausted, at which point they were reset to the complete 9; this assured that animals would tend to be distributed though out the trials.[4])

Participants were instructed to answer each question according to whether they agreed with the conclusion, assuming that the premises were true. For each question participants were asked to select "agree" or "disagree" and to give a confidence rating on a five-point scale. At the end of the experiment, participants were asked to state their native language, with the clarification that this is "the language that was spoken at home when you were a child".

The full experiment can be found at `http://www.stanford.edu/~danlass/experiment/animals.html`.

4.2 Results

We considered data only from the 484 participants who reported that their native language was English. We removed data from 44 participants who gave the same response to all questions, as well as 196 individual trials in which subjects responded very quickly (less than 3.5 seconds). The results reported are from 8710 remaining trials.

4.2.1 Non-linearities

Recall the crucial finding of Heit and Rotello (2010); Rotello and Heit (2009) showing that sensitivity to argument validity is greater when participants are asked to judge whether a conclusion is "necessary" than when they are asked whether it is "plausible" (Table 1). In more neutral terms, Heit & Rotello found that asking about necessity versus plausibility affects response probabilities more for Identity and Inclusion arguments than it does for Contingent arguments. We replicated this effect with our within-participants manipulation of modal frame (Table 2). If we follow Heit and Rotello (2010); Rotello and Heit (2009) in grouping together the Identity and Inclusion arguments as "'Valid", the difference in d' values is significant ($p < .05$) based on a 10,000-sample bootstrap test.

	Necessary	*Plausible*
Acceptance, Valid	.82	.94
Acceptance, Contingent	.41	.82
Sensitivity (d')	1.26	.67
Sensitivity, 95% CI	[1.06,1.47]	[.43,.97]

Table 2: Comparing 'Valid' (Identity and Inclusion) and Contingent arguments: acceptance rates and d' in our experiment (*plausible* and *necessary* only), with 95% confidence intervals.

[4]Because of the random selection of premises participants occasionally saw the same premise categories twice, though with different target properties. The median number of distinct premise patterns was 19 (of 21 trials).

Disaggregating these conditions, the pattern of results is similar, as in Tables 3 and 4. In the Inclusion condition there is a significant difference in d' values between "necessary" and "plausible" arguments, $p < .05$. In the Identity condition the difference is not significant, but the failure of significance in this condition can be attributed to reduced statistical power in the disaggregation: random sampling of conditions led to fewer data points in the Identity & "Necessary" condition than in the Inclusion & "Necessary" condition. Indeed, it is clear from Tables 3 and 4 that our participants accepted Identity and Inclusion arguments at almost identical rates. Because we found

	Necessary	*Plausible*
Acceptance, Inclusion	.84	.94
Acceptance, Contingent	.41	.82
Sensitivity (d')	1.33	.69
Sensitivity, 95% CI	[1.08,1.62]	[.36,1.18]

Table 3: Comparing Inclusion and Contingent arguments: acceptance rates and d' in our experiment (*plausible* and *necessary* only), with 95% confidence intervals.

	Necessary	*Plausible*
Acceptance, Identity	.80	.94
Acceptance, Contingent	.41	.82
Sensitivity (d')	1.19	.65
Sensitivity, 95% CI	[.94,1.48]	[.33,1.09]

Table 4: Comparing Identity and Contingent arguments: acceptance rates and d' in our experiment (*plausible* and *necessary* only), with 95% confidence intervals.

no systematic differences between the Identity and Inclusion conditions, we continue to aggregate the Identity and Inclusion conditions as "Valid" for increased statistical power in the remainder of this section.

Comparing Table 1 with Table 2, there are two clear differences between our results and those of Rotello and Heit (2009): our participants rejected Identity and Inclusion arguments with *necessary* more often than with *plausible*, and they accepted contingent arguments at much higher rates. These factors contributed to lower d' in both conditions. Both differences can be glossed as less strong responses in our experiment than previous experiments. The previous lack of difference between modal frames for Identity and Inclusion arguments is plausibly a ceiling effect that does not occur in our paradigm. The higher acceptance rate for Contingent arguments signals less certainty that they should be rejected. Two plausible sources for these differences are the minimal, within-participants manipulation in our paradigm, and the differences in the concreteness of materials. Rotello & Heit used the same predicate "have property X" in all arguments, using the variable X in the stimuli and instructing participants to treat property X as a novel biological property; following Osherson et al. (1990) we used concrete novel properties. While these differences

may be interesting in themselves, the fact that d' differed significantly for "necessary" and "plausible" suggests that our within-participants manipulation successfully captured the core features of between-participants manipulations in previous work. That is, the difference previously reported can be elicited by a difference of a single modal word, and so appears to be triggered by semantic properties of the modal words.

The PTM predicts two general patterns directly related to the issue of non-linear response patterns. First, the choice of modal should have less influence on acceptance rates with valid (Identity & Inclusion) arguments than with contingent arguments, since the former have conditional probability close to 1 (Figure 1a). This is indeed what we find: see Figure 3.

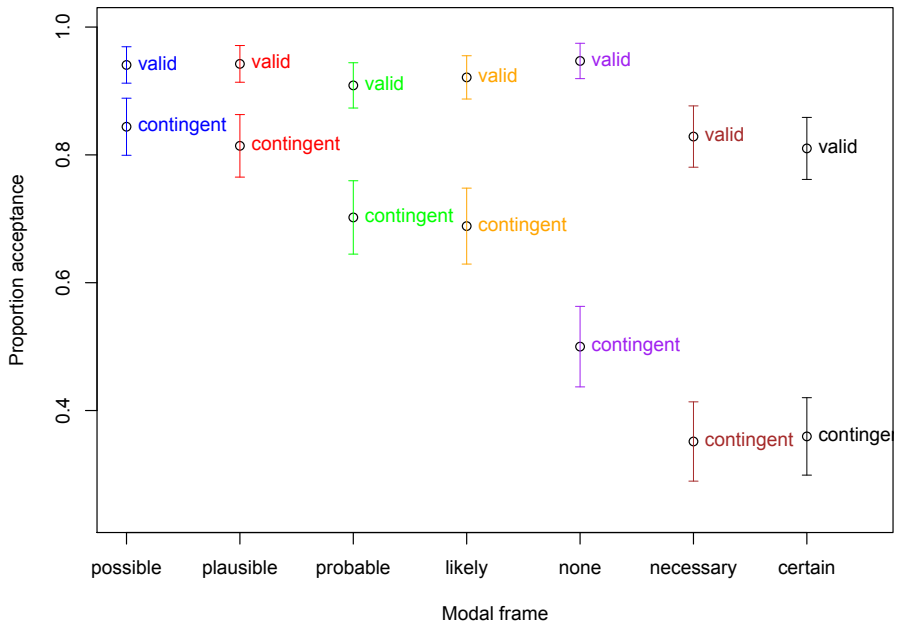

Figure 3: Valid (Identity & Inclusion) and Contingent endorsement rates by modal frame with 95% CIs.

Second, our model predicts the possibility of a continuous gradient in d' values (Figure 1b). The experimental results confirm this prediction as well: see Figure 4. We conclude that there are not simply two "modes" of human reasoning—deductive/"necessary" and inductive/"plausible"—but a spectrum that depends on the modal frame used.

4.2.2 Quantitative fit of power-law model

For a model which makes specific quantitative predictions, such as the power-law model described above, it is possible to test quantitative predictions against the acceptance rate across modal frames

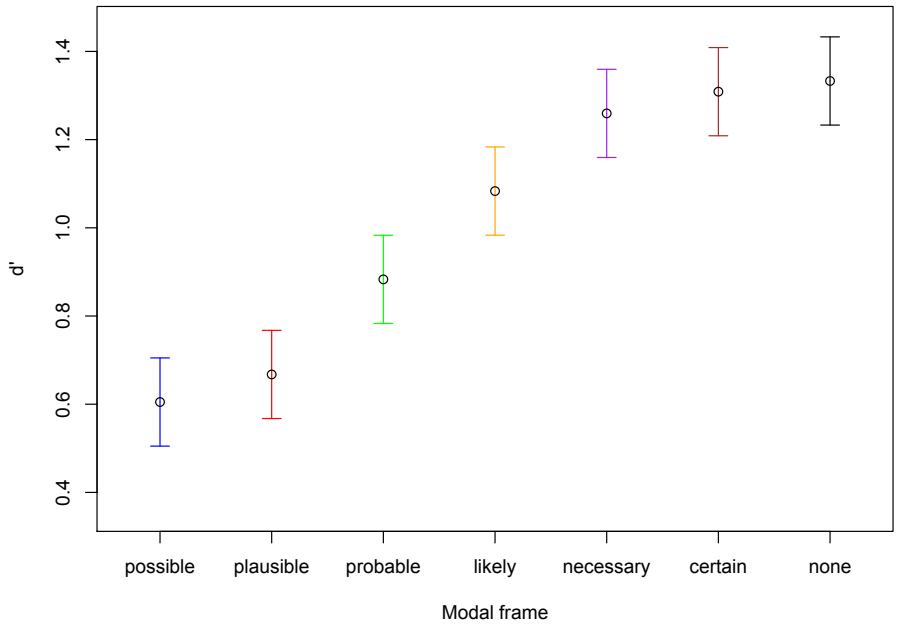

Figure 4: d' by modal frame with 95% bootstrap confidence intervals.

for each argument type. Although we have no independent estimate of the conditional probability of each argument, this model allows us to choose a baseline condition and use it to make predictions about acceptance rates in the other conditions. The form of the power-law model entails that the choice of baseline condition should not affect predictions, because of the following relationship:

$$\begin{aligned} pr(\textit{Is it } M_1 \textit{ that } C|P) &= pr(C|P)^{\alpha_{M_1}} \\ &= pr(C|P)^{(r \times \alpha_{M_2})} \\ &= pr(\textit{Is it } M_2 \textit{ that } C|P)^r \end{aligned} \tag{10}$$

In accord with this prediction we found no systematic effect on model fit by varying the choice of baseline. The primary effect of this choice, then, is that estimates of α_M for the other conditions are up to multiplication by α_0, the parameter which determines the shape of $P(\theta_M)$ for the baseline condition in this version of the model.

Here we use the no-modal condition as a baseline. To be clear, we do not believe that acceptance rates in this condition are an estimate of the true conditional probability of the unmodalized conclusion given the premises: rather, we suggested above that acceptance rates in the no-modal condition should be much more conservative than we would obtain by sampling from the conditional probability. This expectation was already confirmed by the results in figures 3 and 4, where the no-modal condition patterned most closely with the "necessary" and "certain" conditions.

Figure 5 plots the endorsement rate of each argument type in the baseline unmodalized con-

dition against the endorsement rate for the same argument type in various modal frames. We calculated the best-fit α_M for each condition. For each graph in Figure 5 the curve determined by this α_M and equation 6 superimposed, with the overall best-fit noise parameter $\varepsilon = .1$. As the R^2 values in Figure 5 and Table 2 show, this model captures much of the variance in the data, but not all of it.

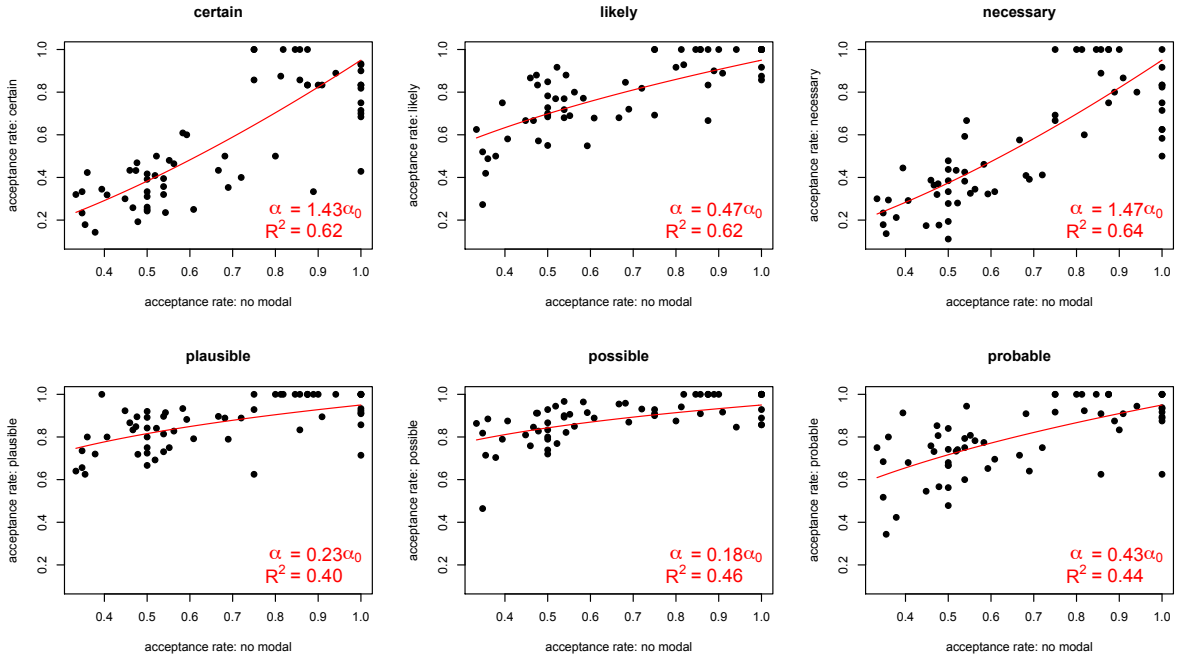

Figure 5: Endorsement rates by argument and modal frame. Curves represent model predictions using (6) and the best-fit α_M for each modal M.

In order to discern whether the remaining variance was due to systematic factors that the power-law model does not capture, we performed a split-half reliability test, randomly dividing the data into equal-sized halves 10,000 times and testing the correlation between the two halves on the same measure that was used to fit our model: acceptance rate for each argument and modal. The model correlations were overall very close to the split-half correlations (Table 5).

The power-law model does seem to capture much of the true structure in the data, though a good deal of intrinsic noise remains. These results do not support the power-law model unreservedly, though: it would be easy to devise other instantiations of the PTM with equal or perhaps slightly better quantitative fit than the power-law model. It is not obvious how to specify competing models that rival the power-law model in simplicity, however. With this caveat, we believe that the power-law model is at least a strong competitor in this realm, and that the remaining noise is plausibly captured by diversity between participants in (for instance) intuitions about the novel biological properties used in our stimuli and the causal relationships that are relevant in this domain (Kemp & Tenenbaum, 2009).

Table 5: Model correlations vs. split-half reliability results.

Modal	model R^2	mean split-half R^2
certain	.62	.70
likely	.62	.58
necessary	.64	.71
plausible	.40	.34
possible	.46	.34
probable	.44	.53

4.2.3 Monotonicity

Recall that, unlike two-dimensional models, the PTM predicts a consistent order in the endorsement rates of arguments across modal frames (No Reversals). In order to test this prediction we considered two tests, a permutation test and a model-based simulation. The average value of Spearman's rank-order correlation for all pairwise comparisons between modal frames in our experiment was .64. A 10,000-sample permutation test revealed that this correlation was highly significant, $p < .0001$.

We then tested the monotonicity predictions directly, using the power law model as the representative of the PTM. As described in section 4.2.2, we used equation 6 with ε fitted globally and α_M fitted by condition against the no-modal baseline. We simulated 50,000 data sets with the same number of observations per argument/modal frame combination that obtained in our experiment. These simulations represent the distribution of results that we would expect to encounter in an experiment of this size if the power-law model is correct. The PTM predicts that, with an infinite number of observations, any pair of conditions would have a rank-order correlation of 1; but sampling error introduces the likelihood of less-than-perfect rank-order correlations with a finite number of observations. This method of simulation thus allows us to inspect the model's predictions about rank order for any specific, finite N.

Figure 6 plots the model's predictions for all 21 pairwise rank-order correlations by condition against the empirical correlations. The model's predictions and the experimental results are highly correlated ($R = .86$, $p < .0001$), and the observed correlations fall within the simulated 95% confidence interval in all 21 pairwise comparisons between modal frames. These results suggest that the model's monotonicity predictions fit the experimental results well, providing further support for our claim that argument strength is based on a single scale which is manipulated by modal expressions.

In contrast, if the relative strength of arguments were not constant across modal frames we would expect to see lower rank-order correlations in the experimental data: the points in Figure 6 would tend to be in the lower right-hand quadrant, instead of falling along the $x = y$ line. This result is not directly incompatible with a two-dimensional model, since such a model could replicate our model's monotonicity predictions by placing no or very little weight on one of the dimensions. However, it would be a surprising coincidence if the seven conditions tested here simply happened to agree on the relative strength of arguments, if the correct model of argument strength is two-

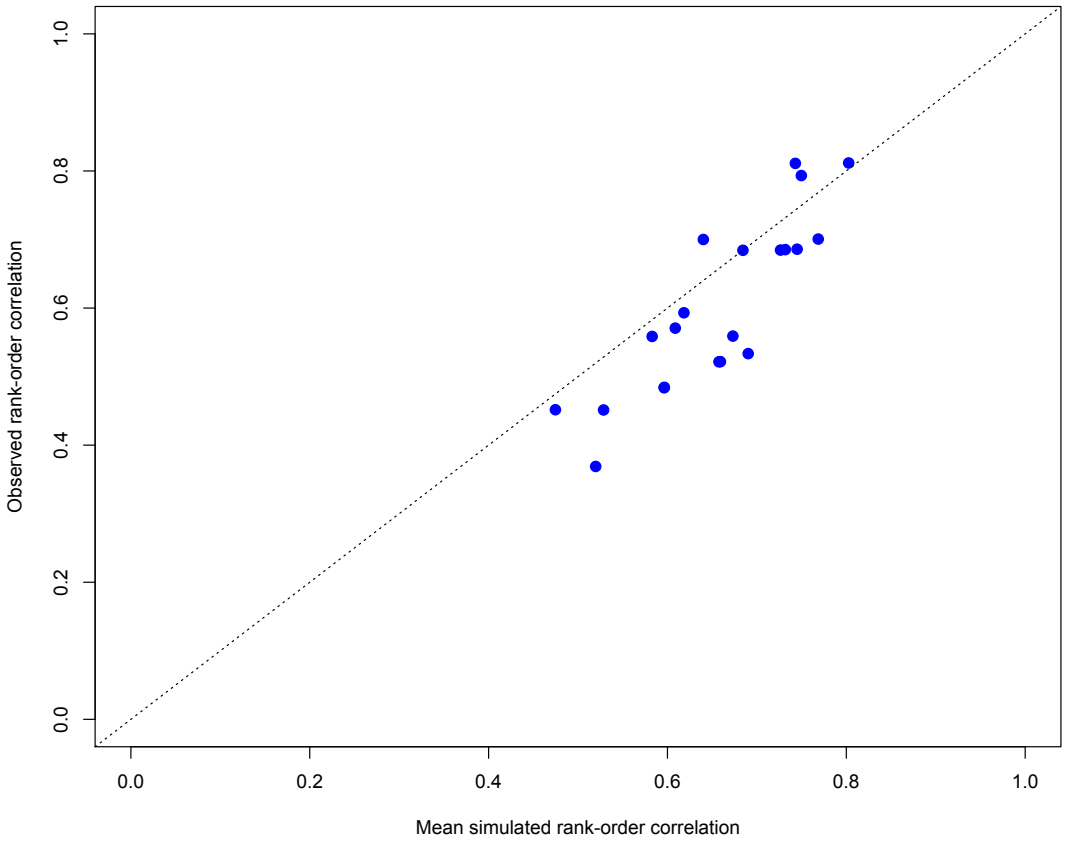

Figure 6: Model-generated vs. empirical rank-order correlations between argument acceptance rates in all 21 pairwise comparisons among modal frames.

dimensional and does not enforce this prediction.

5 Discussion and Conclusion

We have shown that non-linearities in responses to Valid and Contingent arguments can be explained by a simple, one-dimensional probabilistic model, the Probability Threshold Model. Non-linearities are thus are not evidence against a probabilistic account of reasoning. In addition, we uncovered evidence which supports the PTM over a two-dimensional theory: a spectrum of d' values depending on modal frame and monotonicity of argument strength across modal frames. The modal language used in the experiments was the crucial factor in creating different response patterns, and attention to the formal semantics of epistemic modals led directly to the PTM model.

This conclusion does not rule out the possibility that there is a qualitative distinction between

inductive and deductive reasoning among those with logical or mathematical training. Neither does it answer the question of whether deductive and inductive reasoning have different teleological or normative profiles (see for example Evans & Over, 2013). However, our results do call into question previous efforts to show that such a distinction is necessary to account for everyday human reasoning, suggesting instead that a single process may underlie both.

Several additional phenomena have been used to support a two-criterion account of reasoning; we briefly consider the most important of these, leaving deeper exploration to future work. Argument length tends to affect plausibility judgments more than necessity judgments (Rotello & Heit, 2009). This is not unexpected in a probabilistic theory, because we expect that adding premises tends to increase the probability of the conclusion given the premises (Heit, 1998). The non-linearities predicted by equation 6 lead us to expect that the same change in probability will have different effects depending on the modal used. However, Rotello and Heit (2009) show that including additional premises can sometimes *decrease* the endorsement rate of necessity conclusions. This is unexpected under the simple Bayesian model of reasoning that we have employed in this paper. We speculate that this decrease in endorsement is a pragmatic effect—perhaps reflecting the perceived confidence or reliability of the speaker—that could be incorporated into a more sophisticated probabilistic model.

Contrary to our monotonicity results, Rips (2001) found a crossover effect in which participants in the "necessary" condition were slightly more likely to endorse logically valid arguments that were inconsistent with causal knowledge than participants in the "plausible" condition, but the reverse was true for arguments consistent with causal knowledge. Heit & Rotello's work did not find any analogous effect, nor did we in the experiment described here. However, it is possible that the effect is real and attributable to the many differences in materials and methods between these experiments. In particular, Rips used content knowledge in a central way in his experiments, manipulating the causal-consistency of conclusions as well as the mode of reasoning. Here too we believe the most plausible explanation of this effect will be found in the pragmatics of causality-violating premises, but we must leave this issue to future work.

Our probabilistic theory is a computational-level description (Marr, 1982), and hence leaves open the psychological process by which people evaluate arguments. One possibility is that people are only able to *sample* possible worlds in accord with the distribution implied by the premises (Vul, Goodman, Griffiths, & Tenenbaum, 2014), and evaluate the truth of the conclusion in these sampled worlds. If people take several samples and respond "yes" when the conclusion is true in each sampled world, we recover a power law. If the average number of samples depends on the modal, we recover the specific case of the PTM model described above. For instance, we would posit that people tend to take more samples to evaluate "necessary" conclusions than "plausible" conclusions. This process-level implementation predicts that under time pressure, when people can take fewer samples, "necessary" would begin to look more like "plausible". Indeed, this is exactly the finding of Heit and Rotello (2010). This interpretation also suggests points of connection with a probabilistic theory based on mental models (Johnson-Laird, 1994).

In order to focus on the effect of epistemic modal conclusions, and their implication for modes of reasoning, we have adopted a simple approach to the underlying evaluation of arguments: conditional probability given simple, certain premises. This must eventually be elaborated to account for

more complex forms of reasoning, for instance those involving uncertain and conditional premises. Some of these extensions may follow from careful treatment of the semantics of words used to form premises and conclusions, as we have argued for here, while others may require revision to the underlying model of argument evaluation. As an example of the former, our linguistically-oriented perspective suggests addressing conditional premises by drawing from the large literature on the semantics and pragmatics of indicative, subjunctive, and other types of conditionals. Unfortunately, there is little consensus, even on such basic questions as whether conditionals have truth-conditions and (relatedly) whether the Equation between probabilities of conditionals and conditional probabilities is satisfied (Edgington, 1995; Égré & Cozic, to appear; Bennett, 2003). The psychological literature on conditionals is equally rich (Douven & Verbrugge, 2010; Evans, Handley, & Over, 2003; Oaksford & Chater, 2007, among many others). We believe there is synergy between the linguistic and psychological progress on conditionals, which could be incorporated into the current approach, but this complex topic will require further research.

Premises that convey, or are interpreted as, uncertain evidence provide an additional challenge. One approach would be to generalize probabilistic conditioning to directly addresses uncertain evidence, for instance via Jeffrey conditionalization (Jeffrey, 1965); this proposal has been subject to conceptual objections (Pearl, 1988, §2.3.3) and its empirical status remains unclear (Hadjichristidis et al., 2014; Pfeifer, 2013; Stevenson & Over, 2001; Zhao & Osherson, 2010). A different approach is to condition not on the facts *in* the premise but on the fact *of receiving* the premises, using a hierarchical Bayesian (Pearl, 1988; Tenenbaum et al., 2011) or probabilistic pragmatics (Frank & Goodman, 2012; Goodman & Stuhlmüller, 2013) approach. Crucially, the core of our results (such as monotonicity) would apply to any theory that derives a single dimension of strength for the underlying argument that is then operated on by the epistemic modal frame.

We have illustrated an approach to reasoning based on an overall probabilistic view of inference, together with careful attention to natural language semantics. The immediate result of this approach is a model that explains modes of reasoning in a new way, and a novel experimental methodology designed to test this model. We have provided evidence that there are many gradations between "necessary" and "plausible" when judging conclusions, and that a monotonicity pattern emerges which supports the Probability Threshold Model over two-dimensional models. In addition, we found a reasonably good quantitative fit to a simple power-law variant of the PTM. We believe that the approach of this research—careful attention to language for development of both theory and experimental methodology—will prove fruitful in clarifying a variety of phenomena related to human reasoning.

6 Acknowledgements

We thank three anonymous *Cognition* reviewers, Evan Heit, and four CogSci 2012 reviewers for helpful discussion of earlier versions. This work was supported by a John S. McDonnell Foundation Scholar Award and ONR grant N00014-09-1-0124 to Noah D. Goodman.

References

Bennett, J. F. (2003). *A Philosophical Guide to Conditionals*. Oxford University Press.

Davis, C., Potts, C., & Speas, M. (2007). The pragmatic values of evidential sentences. In M. Gibson & T. Friedman (Eds.), *Proceedings of Semantics and Linguistic Theory 17* (pp. 71–88). Ithaca, NY: CLC Publications.

Douven, I., & Verbrugge, S. (2010). The Adams Family. *Cognition*, *117*(3), 302–318.

Earman, J. (1992). *Bayes or bust? a critical examination of bayesian confirmation theory.* Cambridge, MA: MIT Press.

Edgington, D. (1995). On conditionals. *Mind*, *104*(414), 235. doi: 10.1093/mind/104.414.235

Edgington, D. (1997). Vagueness by degrees. In R. Keefe & P. Smith (Eds.), *Vagueness: A reader* (pp. 294–316). MIT Press.

Égré, P., & Cozic, M. (to appear). Conditionals. In M. Aloni & P. Dekker (Eds.), *Cambridge handbook of semantics.* Cambridge University Press.

Evans, J., Handley, S. J., & Over, D. E. (2003). Conditionals and conditional probability. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *29*(2), 321.

Evans, J., & Over, D. E. (1996). *Rationality and reasoning*. Psychology Press.

Evans, J., & Over, D. E. (2013). Reasoning to and from belief: Deduction and induction are still distinct. *Thinking & Reasoning*, *19*(3-4), 267–283.

Frank, M., & Goodman, N. (2012). Predicting pragmatic reasoning in language games. *Science*, *336*(6084), 998–998.

Frazee, J., & Beaver, D. (2010). Vagueness is rational under uncertainty. *Proceedings of the 17th Amsterdam Colloquium*.

Gabbay, D. M., & Woods, J. H. (2009). *Handbook of the history of logic: Inductive logic* (Vol. 10). Elsevier.

Goodman, N. D., & Stuhlmüller, A. (2013). Knowledge and implicature: Modeling language understanding as social cognition. *Topics in cognitive science*, *5*(1), 173–184.

Hadjichristidis, C., Sloman, S. A., & Over, D. E. (2014). Categorical induction from uncertain premises: Jeffrey's doesn't completely rule. *Thinking & Reasoning*.

Heit, E. (1998). A Bayesian analysis of some forms of inductive reasoning. *Rational models of cognition*, 248–274.

Heit, E., & Rotello, C. (2010). Relations between inductive reasoning and deductive reasoning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *36*(3), 805.

Jaynes, E. (2003). *Probability theory: The logic of science*. Cambridge University Press.

Jeffrey, R. C. (1965). *The logic of decision*. University of Chicago Press.

Johnson-Laird, P. N. (1986). *Mental models: Towards a cognitive science of language, inference, and consciousness*. Harvard University Press.

Johnson-Laird, P. N. (1994). Mental models and probabilistic thinking. *Cognition*, *50*(1-3), 189–209.

Kemp, C., & Tenenbaum, J. (2009). Structured statistical models of inductive reasoning. *Psychological review*, *116*(1), 20.

Kennedy, C. (2007). Vagueness and grammar: The semantics of relative and absolute gradable adjectives. *Linguistics and Philosophy*, *30*(1), 1–45.

Lassiter, D. (2010). Gradable epistemic modals, probability, and scale structure. In N. Li & D. Lutz (Eds.), *Semantics & Linguistic Theory (SALT) 20* (p. 197-215). CLC Publications.

Lassiter, D. (2011). Vagueness as probabilistic linguistic knowledge. In R. Nouwen, R. van Rooij, U. Sauerland, & H.-C. Schmitz (Eds.), *Vagueness in communication* (p. 127-150). Springer.

Lassiter, D. (in press-a). Adjectival modification and gradation. In S. Lappin & C. Fox (Eds.), *Handbook of semantics.* Blackwell.

Lassiter, D. (in press-b). Epistemic comparison, models of uncertainty, and the disjunction puzzle. *Journal of Semantics*.

Lassiter, D. (to appear). *Measurement and Modality: The Scalar Basis of Modal Semantics*. Oxford University Press.

Lassiter, D., & Goodman, N. D. (2013). Context, scale structure, and statistics in the interpretation of positive-form adjectives. In T. Snider (Ed.), *Semantics & Linguistic Theory (SALT) 23* (p. 587-610). CLC Publications.

Lawry, J. (2008). Appropriateness measures: an uncertainty model for vague concepts. *Synthese*, *161*(2), 255–269.

Lewis, D. (1979). Scorekeeping in a language game. *Journal of Philosophical Logic*, *8*(1), 339–359. doi: 10.1007/BF00258436

Macmillan, N., & Creelman, C. (2005). *Detection theory: A user's guide*. Lawrence Erlbaum.

Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information*. Henry Holt and Co.

Morzycki, M. (to appear). *Modification*. Cambridge University Press.

Oaksford, M., & Chater, N. (1998). *Rationality in an uncertain world: Essays on the cognitive science of human reasoning*. Psychology Press.

Oaksford, M., & Chater, N. (2001). The probabilistic approach to human reasoning. *Trends in Cognitive Sciences*, *5*(8), 349–357.

Oaksford, M., & Chater, N. (2007). *Bayesian rationality: The probabilistic approach to human reasoning*. Oxford University Press.

Oaksford, M., & Hahn, U. (2007). Induction, deduction, and argument strength in human reasoning and argumentation. In A. Feeney & E. Heit (Eds.), *Inductive reasoning: Experimental, developmental, and computational approaches* (pp. 269–301). Cambridge University Press.

Osherson, D., Smith, E., Wilkie, O., Lopez, A., & Shafir, E. (1990). Category-based induction. *Psychological review*, *97*(2), 185.

Over, D. E. (2009). New paradigm psychology of reasoning. *Thinking and Reasoning*, *15*(4).

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann.

Pfeifer, N. (2013). The new psychology of reasoning: A mental probability logical perspective. *Thinking & Reasoning*, *19*(3-4), 329–345.

Pfeifer, N., & Kleiter, G. D. (2010). The conditional in mental probability logic. *Cognition and conditionals: Probability and logic in human thought*, 153–173.

Portner, P. (2009). *Modality*. Oxford University Press.

Ramsey, F. (1931). Truth and probability. In *The foundations of mathematics and other logical essays* (pp. 156–198). Routledge and Kegan Paul, Ltd.

Rips, L. (1994). *The psychology of proof: Deductive reasoning in human thinking*. MIT Press.
Rips, L. (2001). Two kinds of reasoning. *Psychological Science*, *12*(2), 129.
Rotello, C., & Heit, E. (2009). Modeling the effects of argument length and validity on inductive and deductive reasoning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*(5), 1317–1330.
Schmidt, L. A., Goodman, N. D., Barner, D., & Tenenbaum, J. B. (2009). How tall is *Tall*? Compositionality, statistics, and gradable adjectives. In *Proceedings of the 31st annual conference of the cognitive science society.*
Stevenson, R. J., & Over, D. E. (2001). Reasoning from uncertain premises: Effects of expertise and conversational context. *Thinking & Reasoning*, *7*(4), 367–390.
Tenenbaum, J., Griffiths, T., & Kemp, C. (2006). Theory-based bayesian models of inductive learning and reasoning. *Trends in cognitive sciences*, *10*(7), 309–318.
Tenenbaum, J., Kemp, C., Griffiths, T., & Goodman, N. (2011). How to grow a mind: Statistics, structure, and abstraction. *Science*, *331*(6022), 1279.
Verheyen, S., Hampton, J., & Storms, G. (2010). A probabilistic threshold model: Analyzing semantic categorization data with the rasch model. *Acta psychologica*, *135*(2), 216–225.
Vickers, J. (2012). The problem of induction. In E. Zalta (Ed.), *The stanford encyclopedia of philosophy.* Center for the Study of Language and Information. Retrieved from `http://plato.stanford.edu/entries/induction-problem/`
Vul, E., Goodman, N., Griffiths, T. L., & Tenenbaum, J. B. (2014). One and done? Optimal decisions from very few samples. *Cognitive Science*.
Williamson, T. (1994). *Vagueness*. Routledge.
Yalcin, S. (2007). Epistemic modals. *Mind*, *116*(464), 983–1026. doi: 10.1093/mind/fzm983
Yalcin, S. (2010). Probability Operators. *Philosophy Compass*, *5*(11), 916–937.
Zhao, J., & Osherson, D. (2010). Updating beliefs in light of uncertain evidence: Descriptive assessment of Jeffrey's rule. *Thinking & Reasoning*, *16*(4), 288–307.

7 Appendix: Proof of monotonicity

Theorem 1 (No reversals) *For any epistemic modals M_1 and M_2 with meanings as in equation 3 and any arguments a_1 and a_2, if* $\textbf{strength}_{M_1}(a_1) > \textbf{strength}_{M_1}(a_2)$ *then* $\textbf{strength}_{M_2}(a_1) \geq \textbf{strength}_{M_2}(a_2)$.

Proof: Let p_i denote $P(Conclusion(a_i)|Premises(a_i))$. By assumption and the definition in equation 5:

$$\textbf{strength}_{M_1}(a_1) - \textbf{strength}_{M_1}(a_2) = \int_{p_2}^{p_1} P(\theta_{M_1})\, d\theta_{M_1} > 0. \tag{11}$$

For a probability density $P(t)$ over $[0,1]$, $\int_x^y P(t)dt \geq 0$ if and only if $y \geq x$. Thus $p_1 \geq p_2$. From this we can conclude that

$$\textbf{strength}_{M_2}(a_1) - \textbf{strength}_{M_2}(a_2) = \int_{p_2}^{p_1} P(\theta_{M_2})\, d\theta_{M_2} \geq 0. \tag{12}$$

This theorem applies to pairs of 'upward' epistemic modals fitting the schema in equation 3, with a > sign in their lexical entries. The proof extends in an obvious way to pairs of 'downward' modals such as *unlikely*, *implausible*, *impossible*, and *doubtful*. Pairs of one upward and one downward modal will precisely

reverse the monotonicity pattern: if a_1 is more certain than a_2 then a_2 is more improbable than a_1. Finally, there are modals such as *It is between 40% and 60% likely that* which are strictly non-monotonic, for clear semantic reasons.