Representing Relevance

Robert Hartzell
The University of Texas at Austin
rhartzell15@gmail.com

January 29, 2025

Abstract

I begin with a gap in the literature on conversational relevance, wherein utterances that shift probability distributions included in the common ground do not count as relevant if they do not rule out one or more answers to the question under discussion. In order to provide a satisfying account of probabilistic conversational relevance, I introduce a relevance measure, $R(\cdot)$. I motivate six axioms for such a function, and show that they uniquely characterize the symmetrized Kullback-Leibler divergence. I then show how we can incorporate this result into an expanded definition of conversational relevance. **Keywords: Language, Common Ground, Relevance, Probability, Formal Epistemology.**

1 Introduction

Some utterances are relevant, and others are not. This is a trivial statement, but to properly categorize an utterance as relevant or irrelevant is a nontrivial task. In this paper, I present one way to define the relevance of certain kinds of conversational moves. I do so by motivating the need for a graded measure of probabilistic relevance and proving that one measure follows uniquely from the desiderata that I motivate.

Ever since Grice introduced his well-known maxims (1975), philosophers and linguists have sought to provide a satisfying account of relevance. My aim is not to arbitrate between these various accounts. Instead, I work specifically from Roberts' influential model (2012), looking to usefully expand upon the definition

of relevance that she gives. However, the issues that I identify will arise under any model of conversational relevance, and the definition of relevance that I give should be suitable for any sufficiently strong model.[1]

Roberts defines relevance in the context of interlocutors who are engaged in the process of answering a question (the **question under discussion**) - in her model, an utterance is relevant when it eliminates at least one of the possible answers to the **question under discussion**. In this paper I expand upon this definition in order to capture the relevance of a specific kind of utterance: those that do not completely rule out any answers to the **question under discussion**, but that shift probability distributions defined over its possible answers. An agent might learn evidence that substantially changes the likelihood that one answer to the question obtains, without actually excluding any possibility; we would like a model that makes this evidence relevant to the discourse.[2]

I will argue for it further in §2, but the motivation for this turn is fairly straightforward and intuitive. Roberts' original model is elegant and powerful, but it leaves out a class of utterances whose ubiquity in everyday conversation cannot be overstated. If we can expand the original setup to include these probabilistic cases without disrupting the original structure, we will have made a useful contribution.

In order to provide a satisfying account of probabilistic relevance, I introduce a way to measure probabilistic conversational relevance, the function $R(\cdot)$.[3] To characterize $R(\cdot)$, I motivate some desiderata that (I posit) any such measure

[1] E.g., Sperber and Wilson offer an account of relevance that has been enormously influential in linguistics, psychology, and elsewhere (1986). Although I am not working within their framework, the relevance measure I introduce could be used to measure the information that an utterance conveys, which is central to their account.

[2] Of note, this notion is not novel. Beaver, Roberts, Simons, and Tonhauser mention the probabilistic limitations of the original definition in *What Projects and Why* (2010).

[3] In Roberts' model, relevance is a relation between a move and a question, whereas $R(\cdot)$ is a function of probability distributions. I am not claiming that $R(\cdot)$ is the relevance-relation, and the definition of relevance that I arrive at maintains the relation between move and question. $R(\cdot)$ is a way of backing into the relevance of a move to a question by looking at probability distributions before and after the fact. I discuss this further at the end of §3.4.

should have. I then prove that these desiderata, which I present as axioms, characterize a single divergence measure: the symmetrized Kullback-Leibler (KL) divergence.[4] This result is interesting in its own right. The fact that the desiderata that matter to us in gauging the relevance of an utterance characterize the KL divergence is worth noting, for philosophers of language and formal epistemologists alike.

My focus will be on measuring the relevance of an utterance *to the agent*, as opposed to its *objective* relevance. This aim is reflected in the basic structure of $R(\cdot)$ itself. Many similar measures (like those of confirmation, or explanatory goodness) rely on some combination of an agent's subjective priors and Bayes' theorem.[5] $R(\cdot)$, on the other hand, takes as its argument only two probability distributions; how the agent-in-question got from one to the other is irrelevant. In this way our measure is not normative - it does not tell us how a rational agent ought to adjust her priors. Instead, we deal only with her probabilistic attitudes before and after updating. While I am advocating for $R(\cdot)$ in the context of conversational relevance, it is worth considering how divergence measures with this kind of quality might be useful in other areas.

The paper will proceed as follows. In §2, I elaborate more on Roberts' original model, and motivate the need for a probabilistic expansion of her definition of relevance. The bulk of our work is done in §3; here I begin by describing the characteristics we might want a measure of relevance to have, and discuss the shortfalls of existing measures. I then motivate each of our axioms and prove the representation theorem for $R(\cdot)$. Upon doing so, I will be able to offer a

[4]Following convention, I refer to the KL divergence as $D_{KL}(X, Y)$ (also called *relative entropy*), and the symmetrized version as $D_{KL}(X, Y)+D_{KL}(Y, X)$. Kullback himself referred to the symmetrized measure itself as the divergence between two distributions (1959). Harold Jeffreys was, in fact, the first to introduce the measure (1939).

[5]There are several measures of concepts like confirmation, explanatory goodness, and relevance that have been developed and studied in the literature, and bear some relation to that which we will arrive at here. Some of these include, but are not limited to, those analyzed by Carnap (1945), Good (1968), McGrew (2003), and Glass and Schupbach (2023). I touch on this further in Section 3.

new definition of relevance. Because $R(\cdot)$ is a graded quantity, and relevance in Roberts' model is on-off, I suggest a stake-sensitive cut-off point (that is: an utterance is relevant when $R(\cdot)$ is greater than some k, where k varies given the stakes of the context). Finally, in §4, I discuss some likely objections.

I should offer here a quick note on my methodology, which is relatively standard. I use linguistic data to motivate the need for a probabilistically-oriented measure itself, by noting the ubiquity of ordinary-language discourse with probabilistic content. I similarly motivate the axioms from which the measure is derived, by taking note of the kinds of qualities that our ordinary usage of relevance and measure has.

2 Relevance in Structured Discourse

In this section I proceed in two parts. First, I review Roberts' model of discourse, and the definition of relevance that we will be working with. Then, I show how this definition excludes utterances that entail shifts in probability distributions defined over the possible answers to the **question under discussion**. I claim that many such utterances are obviously relevant, and that we should expand our model to include them.

2.1 Relevance without probabilities

My first goal is to briefly reiterate the structure of Roberts' original model, in order to provide better context for my own definition of relevance. I will attempt to keep this portion short and simple; the majority of the definitions she gives will be unnecessary for my own explication.

Roberts offers a precise way to model the **information structure** of a discourse (2012). A discourse is, simply, an act of conversation between two or more interlocutors who are engaged in the process of communal inquiry. The

interlocutors share between them a **common ground**, the set of propositions for which the interlocutors each behave as if they each take them to be true, and a **context set**, the set of possible worlds in which all of the propositions held in the **common ground** are true.[6] For Roberts, and thinkers working in a similar paradigm,[7] inquiry is aimed at narrowing the **context set** down to a single world: in effect, answering the super-question, what world are we in? Interlocutors do so by introducing new sub-questions and answers that help to narrow down the possibilities.

The **information structure** of a discourse is structured similarly. In a given discourse, we begin with the **question under discussion**, the question that the interlocutors are trying to answer. A question, q, is identified by the set of its possible answers, $(a_1, \cdots, a_n)$. An interlocutor then might make a conversational move.[8] There are two kinds of moves that are acceptable in this model: questions and assertions. Assertions aim to directly answer the question at hand, by ruling out one or more of the possible answers. In the above case, a successful assertion would rule out at least one of the a_i, and an assertion that fully answered the question would rule out all but one of the a_i.

Questions, on the other hand, open up new avenues of inquiry. Acceptable questions are those whose resolutions might draw the interlocutors closer to answering the previous question. We say that questions that are acceptable in this sort of way are part of a *strategy*.

We can begin here to construct an informal definition of relevance. Assertions should count as relevant when they answer the question (or one of the sub-questions) we are working on, and questions should count as relevant when their answers will bring us closer to answering the previous question. This

[6] I ignore the rich debate about what the **common ground** really is, what it means for an agent to act as if a proposition is true, and so-on.

[7] E.g., Stalnaker (1978).

[8] I use 'move' and 'utterance' interchangeably.

approximates the definition that we will be working with:

Definition 1. A move m is **Relevant**$_1$ to the question under discussion q *iff* m either introduces a partial or full answer to q (m is an assertion) or is part of a strategy to answer q (m is a question) (Roberts, 2012).

In sum, a relevant move in a discourse is one that either introduces a new question, whose answer will either partially or fully answer the **question under discussion**, or an assertion, that introduces a partial or full answer to the question under discussion.

2.2 Relevance with probabilities

To better motivate the need for an expanded definition of relevance, I introduce a simple example. The case I present here is straightforward, and might come across as trivial. This obviousness is meant to be a feature: as we will see, cases like this are not fringe scenarios meant to be explained away by appeals to vagueness and the like. Instead, they are paradigmatic examples of discourse, and, as such, ought to be captured by our models.

In what follows I use the term 'updating,' but I remain agnostic as to what that means for the agents at hand. More specifically, I do not require that our agents are updating their priors by Bayesian conditionalization. All I require is that the agents change the probabilities that they assign to the possible answers.

First, I need to make an additional assumption: that probabilities can be held in the **common ground**. I could spend a great deal of space here discussing the matter, but I will encourage curious readers to refer to Yalcin's influential paper on the subject (2012). What this really means, in the context of my discussion, is that interlocutors can, and often do, agree upon the probability distributions defined over the possible answers to the questions that they are discussing, on the ways that they will update those distributions upon

learning new evidence, and that this agreement is common knowledge for the agents.[9] I will discuss drawbacks to this in §4 but, for now, I will be referring to shared probability distributions.

I will also refer to agents who update their priors on evidence that is *entailed* by an utterance. Here I am adding something additional to Roberts' original model: evidence as a mediating concept. I do so to leave room for those utterances that are not such that they can be immediately updated upon. I do not claim that the evidence at hand is the *strongest* evidence entailed by the utterance. Instead, all I require is that each interlocutor involved deduces the same evidence as do the others. This again highlights the subjective nature of the task at hand: I care about the evidence that the agents are actually able to glean from proffered utterances, as opposed to the totality of the evidence that those utterances entail (especially when the entailed evidence might not be available or clear to the agents involved).[10]

Case 1. Consider two doctors, discussing whether or not a patient, Mo, might be a suitable candidate for a new medication. Under our model, this is the **question under discussion**. Call it Q_1 = (is Mo a suitable candidate for medication?). Now, imagine that having a certain genetic trait, G, would make Mo a good candidate. So, the doctors introduce a sub-question: does Mo have genetic trait G? Call this sub-question q_1. This is a relevant conversational move, because the answering of the sub-question answers the **question under discussion** above it.

Next, say that the instance of G in the general public is random, and that one out of every two people have it. So, our doctors, being well-educated on

[9] I treat the agents involved as reflectively aware of the relevant probability distributions that they hold. I do not see any difficulty arising in cases in which they are unaware of their own credences, provided they are able to update those credences on new evidence, and eventually reach conclusions about the answers over which the probabilities are held. However, this sort of situation would require a different conception of **common ground**, with a weaker common knowledge condition.

[10] In this way our agents are not required to be logically omniscient.

such matters, each hold prior probability over the likelihood of Mo's having G equal to .5. Call the shared probability distribution over q_1 X_1, and say: $X_1 = (x_1, x_2) = (.5, .5)$, with $x_1 = Pr$(Mo has G), and $x_2 = Pr$(Mo doesn't have G).

Our good doctors order a blood test. The results are informative, but not perfectly conclusive. Call the results of the blood test E_1, and say that one of the doctors offers E_1 as the next conversational move. They both update their priors on this new evidence, and arrive at the posterior distribution $Y_1 = (.9, .1)$, with $y_1 = Pr$(Mo has G) and $y_2 = Pr$(Mo doesn't have G).[11]

In ordinary practice, uttering E_1 would be a relevant conversational move: it substantially shifted the doctors' priors. Practically, in cases where, say, the medicine at hand has minimal negative side effects, it could be enough data for them to answer the original question to the affirmative (that, yes, Mo is a suitable candidate for the medication). But, because E_1 does not formally exclude either of the answers to q_1, it is an irrelevant conversational move under **Relevance**$_1$.

These are the kinds of cases that should encourage us to adopt an expanded definition of relevance. It is not contrived: discourse like the one described above happens all of the time. The next part of the task will be to accurately model it.

3 Towards a New Definition of Relevance

In this section, I walk through the steps needed to provide a satisfactory account of relevance, one that will be equipped to handle cases like the one described above. In §3.1, I posit that a *measure* of relevance will be necessary for such

[11] 'Posterior' probability traditionally refers to probability distributions arrived at via application of Bayes' theorem, but I use it to refer to a distribution arrived at via the application of any updating method, not just Bayes'.

a definition. In §3.2, I show why most existent measures of relevance will not suffice. In §3.3, I motivate the axioms that will determine the measure, and prove that they uniquely characterize the KL divergence. Finally, in §3.4, I give a new definition of relevance that incorporates this measure.

3.1 Probabilistic relevance is graded

My argument here is straightforward. I claim that we need a way to measure the relevance of utterances that cause probabilistic shifts, and I call that measure $R(\cdot)$.

Upon first glance, one might think that any utterance that entails any shift in the probability distribution over the question at hand should count as relevant. Surely, though, there is quite a difference between utterances that induce substantial shifts, like the one above, and an utterance that (e.g.) shifts a distribution from $(.5, .5)$ to $(.5001, .4999)$. If we count each of these conversational moves as relevant, we lose much of the explanatory power that we hope to get from relevance in the first place.

So, our concept of relevance ought to be graded. Any utterance that entails any change in the probability distribution over the answers to q has some relevance, but some such utterances have more relevance than others. Hence, I introduce a relevance measure, $R(\cdot)$. $R(\cdot)$ will measure the relevance of a move, m, to a **question under discussion**, q. We want $R(\cdot)$ to give us a numerical representation of this relevance; we cannot plug in propositional questions and assertions and expect a numerical outcome. So, we define $R(\cdot)$ as a function of X (the prior probability distribution over q) and Y (the posterior distribution, after updating on the evidence entailed by m). Intuitively, $R(\cdot)$ will work as a kind of distance measure between the two probability distributions. It ought to tell us, in essence, how different the posterior distribution is from the prior.

The greater the difference between X and Y, the more relevant m is to q.

3.2 Existing measures are inadequate

Philosophers have long debated the merits of various relevance measures, and it will be useful to consider some of them here. Most often these are meant to measure the relevance that a piece of evidence, E, has vis-à-vis a hypothesis, H, or are a measure of a hypothesis' ability to account, explanatorily, for a piece of evidence. One might believe that one of these measures will be sufficient; all we need to do now is to choose whichever best suits our purposes, and go from there. However, there are some crucial differences that we must first consider.

Let us examine Good's relevance measure (1968), defended later on by Mcgrew (2003). Of course there are others that differ substantially (e.g., Schupbach and Glass' log-measure of explanatory goodness [2023]), but they share similarities that suffice for what I say here. Consider the following:

$$r_G(e,h) = \frac{Pr(e|h)}{Pr(e)}$$

$r_G(e,h)$ takes as its arguments two propositions, e and h. Applied to our language case, h is equivalent to a possible answer to the question under discussion, and e to the move whose relevance we are trying to measure (or to the evidence that that move entails). To find the total relevance of the utterance, we could take the sum of the measure over each of the possible answers.

Here we certainly have a plausible strategy. Moreover, using a measure that takes propositions as its arguments might be a positive outcome: it simplifies our model, without requiring the agents at hand to first update their probabilities. But, herein lies the difficulty. We want to measure relevance *for the agents*; we can only do so if we allow them to first update, and then measure the distance

between the resultant probability distributions. Functions like Good's (and the others mentioned) all take propositions as their arguments, and measure a more *objective* relevance than we would like, by way of Bayesian conditionalization (as is included in the numerator of Good's measure).

To stress this point: there are reasons to allow for our agents to be non-Bayesian. One might prefer an inference-to-the-best-explanation updating rule, as Douven defends (2013). Or, the agents-in-question might update in a way that seems incoherent and irrational to the impartial observer. In actual practice, human reasoners are not perfectly rational; we would like to be able to define relevance in a way that captures actual practice. So, we leave the updating method open: the agents can move from the prior to the posterior distributions in whichever way they would like.

Given the last point, we choose to restrict our attention to measures that take probability distributions as their arguments, without supposing Bayesian conditionalization to reach the posterior distribution. There are, however, several functions we could choose from: Euclidean distance metrics, the Brier Score (see Pettigrew, 2016), the Jensen-Shannon divergence (see Lin, 1991), etc. As a result, I choose to first identify some desiderata, and use these to motivate a unique measure.

3.3 Defining $R(\cdot)$

Before we move on, let me revisit the dialectic up to this point. I first showed that Roberts-style relevance is unequipped to deal with probabilistic discourse: utterances that shift probability distributions do not count as relevant conversational moves, even when they should. I then argued that an expanded definition of relevance will require a way to measure the relevance of an utterance. Because there is such difference between the contributions of utterances that entail

large probability shifts and those that entail small ones, we need a way to be able to formally rank them. I call the function that we will use to measure the relevance of these shifts $R(\cdot)$. In the previous section, I showed that many existing relevance measures are unsatisfying for our purposes here; we want to allow our interlocutors to choose their own updating methods, in order to best measure relevance *for the agent*.

Hence, my purpose in this section is to motivate some of the qualities that our measure ought to have. These desiderata will, in turn, become axioms, and I will then prove that these axioms uniquely characterize the symmetrized KL divergence: $R(X,Y) = D_{KL}(X,Y) + D_{KL}(Y,X)$.

Take X and Y to be two probability distributions, with $X = (x_1, \cdots, x_n)$ and $Y = (y_1, \cdots, y_n)$. Call X the prior distribution and Y the posterior, obtained after our interlocutors updated on some new evidence, E, entailed by a conversational move, m. Our task is to find out how relevant m was to the conversation, by way of the size of the change that E induced.

We will start by finding the informational gain from X to Y. To do so, we will focus on the difference between the individual members of X and Y - that is, for each i, the gain from x_i to y_i. This strategy is pragmatic, allowing us to first focus on discrete units of probability, as opposed to whole distributions. It will help, too, to remember that each of the individual probabilities is defined over a possible answer to the **question under discussion**. By focusing our attention here, we are finding the relevance of the utterance in-regards-to that particular possible answer. For ease, say that the informational gain between any x_i and y_i is $d(x_i, y_i)$.

Now our job is to define some desiderata for $d(\cdot)$. To start with, consider a simple case: two distributions, X_1 and Y_1, such that $x_i = y_i$. I say that, thus, $d(x_i, y_i) = 0$. Regarding the specific possible answer over which the probabil-

ities x_i and y_i are both defined, no information has been lost or gained, and $d(\cdot)$ should reflect this. Considering the other direction next, imagine any case where we know that $d(x_j, y_j) = 0$. Then, we should know that $x_j = y_j$; any difference between the two, no matter how small, should be captured by some difference in the value of the measure. Thus, we have our first axiom:

Axiom 1. *$d(x, y) = 0$ iff $x = y$.*

Next, we want to know how our measure ought to behave when its arguments are scaled: that is, how will $d(x, y)$ compare to $d(hx, hy)$, with h an arbitrary constant? One might assume that the measurements should obviously be different. However, we should be careful here: certain shifts in probability, although small in a purely Euclidean sense, represent massive informational gain. For the sake of argument, consider three shifts in probability: one from .05 to .1 (call this shift S_1), one from .4 to .8 (S_2), and another from .2 to .3 (S_3). The salient question here is as to how $d(.05, .1)$, $d(.4, .8)$, and $d(.2, .3)$ should differ. I show below that we should have $d(.05, .1) = d(.4, .8) > d(.2, .3)$.

To demonstrate this point, I will make use of Bayes' theorem. The goal is to show that the strength of the information necessary to induce S_1 and S_2 is the same; the information required for S_3 will be weaker. Call the possible answers over which the agents hold these probabilities A_1, A_2, and A_3, respectively, and call the evidence entailed by the three utterance E_1, E_2, and E_3. Starting with S_1, by Bayes' theorem, we have it that $Pr(A_1|E_1) = \frac{Pr(A_1)Pr(E_1|A_1)}{Pr(E_1)} \rightarrow .1 = \frac{(.05)Pr(E_1|A_1)}{Pr(E_1)} \rightarrow \frac{Pr(E_1|A_1)}{Pr(E_1)} = 2$. By the same reasoning, we get $\frac{Pr(E_2|A_2)}{Pr(E_2)} = 2$, and $\frac{Pr(E_3|A_3)}{Pr(E_3)} = 1.5$.

The ratios here, $\frac{Pr(E|A)}{Pr(E)}$, tell us how likely the evidence must be given the possible answer (the *likelihood*, $Pr(E|A)$), and how unlikely the evidence must be ($Pr(E)$), in order to induce the shifts.[12] When this ratio is larger, the shift

[12]Notice that this ratio is Good's relevance measure, as discussed in the previous section.

has a more stringent informational requirement: an agent must observe evidence that is less likely in general, but more likely given the hypothesis. We see, now, that shifts that differ by scalar multiplication will always have the same likelihood-to-evidence ratio. We also see that the shift S_3, although a larger Euclidean distance, has a smaller informational requirement than does S_1. The above phenomenon is made even more acute by shifts of several orders of magnitude (e.g., imagine a shift from .00001 to .1): the informational requirements for these types of shifts are very large.

The informational gain from one probability to the next, as represented by $d(\cdot)$, should reflect these differences and similarities. Because $\frac{Pr(E|A)}{Pr(E)}$ is the same for shifts that differ by scalar multiplication, we assert that $d(hx, hy) = d(x, y)$. This, in turn, will be our second axiom, with the caveat that $h > 0$:

Axiom 2. *For any* $h \in \mathbb{R}^+$, $d(hx, hy) = d(x, y)$.

For our next axiom, I will start by introducing another case:

Case 2. Consider the doctors and patient from **Case 1**. Say that the doctors again start with shared prior $(.5, .5)$ over q_1, whether or not Mo has genetic trait G. This time, however, they order two blood tests. The evidence from the first test, E_1, moves the probability distribution to $(.3, .7)$; given this one test, they both now have decently high credence that Mo does *not* have the trait in question. To be sure, though, they order another, more accurate test. Its results are conclusive in the other direction. Following the utterance of this new data, E_2, the doctors hold probabilities $(.8, .2)$ over the answers to q_1.

Now say that the doctors hear the same evidence, but in a different order. First, E_2, the more conclusive blood test, moves the probabilities to $(.9, .1)$ over q_1. To be safe, they order the other test, too. Upon learning its result, E_1, their probabilities shift to $(.8, .2)$.

In both cases the doctors arrived at the same probability distribution. Should

$d(.5, .3) + d(.3, .8) = d(.5, .9) + d(.9, .8)$? I claim: yes. In both cases, the doctors learned the same total evidence, which affected their final probability values in the same way. Net informational gain should be the same, regardless of the path required to get there. I thus introduce our third axiom, *path-independence*:

Axiom 3. *For any* $q, q' \in (0, 1)$, $d(x, q) + d(q, y) = d(x, q') + d(q', y)$.

Next, I claim that $d(\cdot)$ should be continuous and differentiable. At this level of abstraction the point might seem weak, but it becomes clearer when we begin to more carefully consider the work that $d(\cdot)$ and $R(\cdot)$ will be doing.[13] If $d(\cdot)$ behaves badly in the realm of continuity, we end up with properties like the following: changes in probabilities resulting in changes in $d(\cdot)$ that are consistent with the size of the input, and then a small change in probability followed by a jump the size of in $d(\cdot)$. In-regards-to differentiability, without it we would get angles, sharp and drastic changes in the behavior of $d(\cdot)$, again after only small changes in the probabilistic inputs. We need our measure to make consistent claims about the relevance of various utterances; a discontinuous measure will result in discontinuous judgments of relevance.

Hence, we make our function smooth:

Axiom 4. *$d(\cdot)$ is a continuous and at least once-differentiable function ($d(\cdot)$ is a C^1 function).*

Now we turn our attention to $D(\cdot)$. Recall that $d(\cdot)$ is a measure of the informational gain between the individual probabilities; $D(\cdot)$ will be a way of summing the individual $d(\cdot)$'s. We could merely take the sum of the $d(x_i, y_i)$ for each i in X and Y, but we would run into a problem. $d(\cdot)$, on its own, is a measure of the informational gain between two individual probabilities. As such, and as ensured by **Axiom 2**, a lot of weight will be given to those small

[13] Given the additive and multiplicative properties of continuous and differentiable functions, a continuous and differentiable $d(\cdot)$ will generate a continuous and differentiable $R(\cdot)$, provided that we build $R(\cdot)$ in the right kind of way.

changes that require a great deal of information (e.g., a shift from .001 to .01). This result is necessary when we are discussing informational gain in-and-of itself, but becomes problematic when we turn our attention to relevance.

As a brief example to stress this point, consider a classic case of faulty statistics: a friend of mine, Joe, is getting tested for a very rare disease. Its likelihood, a shared prior between him and his doctor, is .0001. He tests positive for it, but the test is not perfectly accurate, and so the posterior likelihood that he has the disease is now .001. He could take this shift as hugely relevant, and spend the rest of the week in mortal terror, or he could understand that the probability of his having it is still quite low. In this case the informational gain from .0001 to .001 is high, but the *relevance* of the result is still low.

So, we walk a middle path. **Axiom 2** makes sure that small shifts are accounted for, but **Axiom 5** will make sure that those shifts do not have undue impact on the measure. We will do so by taking the *ex-post average* of $d(\cdot)$ taken over two distributions, a *weighted average* of the $d(\cdot)$ values. That is, we will multiply each $d(x_i, y_i)$ by y_i, thus weighting the measure by the size of the posterior probability value. Formally:

Axiom 5. $D(X, Y) = \sum_{i=1}^{n} y_i * d(x_i, y_i)$.

Finally, now we consider the relationship between $D(\cdot)$ and $R(\cdot)$ itself. Recall that $D(\cdot)$ is calculated by way of information *gain*. As such, it is not symmetric: $D(X, Y) \neq D(Y, X)$. I argue, though, that any measure of relevance ought to be symmetric. I care equally about evidence that brings my prior of .5 to .1, and evidence that brings a prior of .1 to .5. Our measure should not favor only gain. For our next axiom, then, we simply symmetrize $D(\cdot)$:

Axiom 6. $R(X, Y) = D(X, Y) + D(Y, X)$.

We now have enough to prove that a unique $R(\cdot)$ follows from our desiderata. First, we see that axioms 1-4 characterize a unique $d(\cdot)$:

Proof. Take $d(x,q) + d(q,y) = d(x,q') + d(q',y)$, and differentiate both sides with respect to y. We can do so by Axiom 4. We get $\frac{\partial d}{\partial y}(q,y) = \frac{\partial d}{\partial y}(q',y)$. Say $\frac{\partial d}{\partial y}(q,y) = f(y)$ (we can do so because the choices of q, q' are arbitrary and so the partial derivatives are functions only of y). Integrate both sides, getting $d(q,y) = F(y) + G(q)$, with $F = \int f$ and G a function. By Axiom 1, $d(x,x) = F(x) + G(x) = 0$. Then, $F(x) + G(x) = F(y) - F(y)$ and so $G(x) + F(y) = F(y) - F(x)$. Hence, $d(x,y) = F(y) + G(x) = F(y) - F(x)$.

Next, by Axiom 2, we have $F(ky) - F(kx) = F(y) - F(x)$. We differentiate with respect to y, yielding $k * F'(ky) = F'(y)$. Multiply both sides by y: $ky * F'(ky) = y * F'(y)$. Substitute $a = ky$: $a * F'(a) = y * F'(y)$. Then, we know that $a * F'(a)$ is a constant; no matter what we choose for a, it will still be equal to $y * F'(y)$, without change in $y * F'(y)$ (because the choice of k is arbitrary). So, say that $a * F'(a) = h$, and then $F'(a) = \frac{h}{a}$. We integrate both sides, getting $F(a) = h * log(a) + C$. Thus, $d(x,y) = h * log(ky) - h * log(kx) = h * log(\frac{y}{x})$. We normalize by setting $h = 1$, getting $d(x,y) = log(\frac{y}{x})$. □

Now that we've determined the function $d(\cdot)$, we can find $D(\cdot)$, which is just the ex-post average of the values of $d(x_i, y_i)$, for all x_i, y_i that make up the probability distributions X and Y. Hence, $D(X,Y) = \sum_{i=1}^{n} y_i * log(\frac{y_i}{x_i})$.[14] Finally, we have our relevance function: By Axiom 6, $R(X,Y) = D(X,Y) + D(Y,X) = \sum_{i=1}^{n} x_i * log(\frac{x_i}{y_i}) + y_i * log(\frac{y_i}{x_i})$.[15][16]

We can make sure, too, that $R(X,Y)$ satisfies other properties that we find desirable. For example, we want $R(X,Y)$ to be nonnegative, with 0 as its minimum. We can prove as much by showing that each of the individual summands

[14] So, $D(X,Y) = D_{KL}(X,Y)$ ($D(\cdot)$ is equal to the KL divergence).

[15] $R(X,Y) = D_{KL}(X,Y) + D_{KL}(Y,X)$.

[16] The original definition of relevance is defined only over questions with discrete answers. What about the possibility of a *continuous* answer? It is easy to come up with an example of such: say we are firing a cannonball, and want to know the range on which it might fall (between 25 and 50 yards? 50 and 100?). Included in our common ground will be a continuous probability distribution over the possibilities. We can easily amend $R(\cdot)$ to be defined over continuous distributions, by integrating as opposed to summing.

is nonnegative:

Proof. Assume, for some $x_i \in X$ and $y_i \in Y$, that $x_i * log(\frac{x_i}{y_i}) + y_i * log(\frac{y_i}{x_i}) < 0$, for proof by contradiction. Without loss of generality, assume that $x_i \geq y_i$, and so $log(\frac{x_i}{y_i})$ is nonnegative. Now, we have $x_i * log(\frac{x_i}{y_i}) < -y_i * log(\frac{y_i}{x_i})$, which means that $x_i < y_i$. But, we had already assumed that $x_i \geq y_i$, and so this is a contradiction. Thus, $x_i * log(\frac{x_i}{y_i}) + y_i * log(\frac{y_i}{x_i}) \geq 0$. □

Next, we show that $R(X, Y) = 0$ iff $X = Y$:

Proof. Assume that $R(X, Y) = 0$. As shown above, each summand itself is nonnegative. So, if their composite sum is to be equal to 0, each summand must be equal to 0. Take any such summand: $x_j * log(\frac{x_j}{y_j}) + y_j * log(\frac{y_j}{x_j}) = 0$. We get $x_j * log(\frac{x_j}{y_j}) = -y_j * log(\frac{y_j}{x_j})$, and thus $x_j = y_j$. This is true for all $j \in (1, \ldots, n)$, and so $X = Y$. Showing that the other direction follows, that $R(X, Y) = O$ whenever $X = Y$, is trivial. □

Now I am in a position to restate precisely my primary claim. Take Q as the **question under discussion** in a discourse, with possible answers $(a_1, \cdots, a_n)$, and say that X is a probability distribution in the common ground defined over the a_i. Say that an utterance m entails evidence E and that the interlocutors in the discourse update accordingly, resulting in posterior distribution Y still defined over the a_i. The **relevance** that utterance m has to the discourse is given by $R(X, Y) = D_{KL}(X, Y) + D_{KL}(Y, X) = \sum_{i=1}^{n} x_i * log(\frac{x_i}{y_i}) + y_i * log(\frac{y_i}{x_i})$.

3.4 Redefining relevance

In this section I introduce a new definition of relevance, using the measure we defined above. It will be broad enough to capture probabilistic utterances of the sort we have been working with, without ignoring utterances of the usual

(non-probabilistic) kind. To do so I will first have to introduce a stake-sensitive cut-off point, which I will discuss below.

$R(\cdot)$ is a way of measuring relevance, but it does not tell us when an utterance simply is or is not relevant. In order to be able to sort utterances into relevant or irrelevant conversational moves, we need to be able to point to an on-off marker of relevance. This marker will be a constant, k, such that an utterance counts as relevant when $R(X, Y) \geq k$, and not when $R(X, Y) < k$.

Whether or not an utterance is relevant in this way, however, depends on context. In one scenario, the stakes of a discourse might be very low: perhaps you and I are trying to decide whether or not to postpone a birthday party because of the weather. Utterances that induce shifts such that $R(\cdot)$ is very small will not matter much to us. You might have information that shifts our belief that it will rain by .01, but the stakes are low enough that this contribution is irrelevant. On the other hand, it is not difficult to imagine cases of inquiry where small shifts matter more. If you and I are trying to decide whether or not to climb Mt. Everest tomorrow, the state of the weather matters to us a great deal - in the first scenario bad weather might ruin the birthday party, but here it is a matter of life-and-death. As such, the information that shifts our belief by .01 is still relevant.

We could formalize the stakes of the discourse in a number of ways, most intuitively by measuring the difference in utility in possible outcomes to the **question under discussion**. I will not take the time to do so here, although I think it is an interesting project worthy of more research.[17] Instead, I simply say that we can assert the existence of a constant, k, that varies with the stakes of the discourse. The higher the stakes, the lower the value of k. I then have enough to introduce our new definition:

[17]There is a great deal of worthwhile literature on the matter of stake-sensitivity. For example, see Weatherson (2005), or Buckwalter and Schaffer (2015). For literature outside of philosophy, see Etchart-Vincent (2004), or Kunreuther et al. (2002).

Definition 2. A move m is **Relevant**$_2$ to the **question under discussion** q with possible answers $(a_1, \cdots, a_n)$ *iff*: 1) m is part of a strategy to answer q, 2) m rules out at least one of the possible answers to q, or 3) m entails evidence E such that, with X the probability distribution shared in the **common ground** between the interlocutors and defined over the a_i, and Y equal to X updated by the interlocutors on E, $R(X,Y) > k_j$, with $k_j \in \mathbb{R}^+$ for some context j, and $R(X,Y) = \sum_{i=1}^{n} x_i * log(\frac{x_i}{y_i}) + y_i * log(\frac{y_i}{x_i})$.

Hence, non-probabilistic assertions, those that straightforwardly rule out possible answers to the **question under discussion**, are still relevant conversational moves. **Relevance**$_2$ simply adds a clause to include utterances with merely probabilistic import.

Importantly, **Relevance**$_2$ is still a relation between moves and the **question under discussion**, as it is in Roberts' model. That is, $R(\cdot)$ itself is not the relevance relation - it is a function between probability distributions. What we have seen, however, is that questions and moves do not themselves give us enough information to determine probabilistic relevance. Instead, we need to know where the interlocutors stand on the **question under discussion** prior to the utterance, and where they stand afterwards. Their prior and posterior distributions are those standings, and $R(\cdot)$ is a way to then measure the distance between.

4 Possible Objections

Here I address four worries that the reader might have about $R(\cdot)$ and **Relevance**$_2$. I believe that each of these can be handled in turn, without impacting the strength of the model as a whole.

4.1 Why not another measure?

To start with, one might argue that there is no good reason to think that $R(\cdot)$ is the only satisfying way to measure the probabilistic relevance of utterances. Perhaps the motivations for our axioms were unconvincing, or there is a different measurement one has in mind. For example, consider Pettigrew's characterization of the Brier Score ($\frac{1}{N}\sum_{t=1}^{N}(f_t - o_t)^2$) as a way to measure the accuracy of a credence (2016). He uses similar desiderata, including a symmetry requirement, and the measure is a function of two probability measures. Why not think that this is a suitable replacement for $R(\cdot)$?

For our purposes, **Axiom 2** precludes the Brier Score. Because it is a measure of accuracy, it fails to take into account the informational requirements levied by shifts with large multiplicative factors. By the summand of the Brier Score, a shift from .1 to .9 over a_i will be larger than one from .001 to .8 over a_i.[18] By $R(\cdot)$, the shift from .001 to .8 will be much larger, because it is a shift from near certainty in *not* a_i to high confidence *that* a_i.

I use this as a case in point that each of our axioms are well-motivated. I do not wish to claim that no other measure of relevance could ever be suitable. Different philosophers might have different expectations for what a measure of relevance should do, and might thus drop some of our axioms and add new ones. I do claim, though, that the burden of proof will be on the next writer to explain why any of the **Axioms 1-6** are ill-founded.

4.2 Minimal and maximal credences

One might object to the fact that $R(\cdot)$ is undefined over 0. As regards probability theory, this is a standard feature. If one's credence in a possible answer is 0, then no evidence can change that via Bayesian updating. Similarly, no evidence

[18] And the Brier score over the full distributions will be larger, provided that some reasonable assumptions about the distributions hold.

can bring one's credence in a proposition to 1, if it was not already there. In Roberts' original model, though, as well as in our **Relevance**$_2$, interlocutors can completely rule out possible answers. It is unclear how they might do so, if they are unable to bring their credences in those answers to 0.

If our interlocutors are rational Bayesian reasoners, we need not worry. They will never have credences of 0 or 1, unless they are considering tautologies and contradictions. Instead they might make progress in a discourse by deciding that some suitably high credence has been reached, thus counting the **question under discussion** as 'answered.'

Recall, though, that we do *not* require our agents to be rational Bayesians; they can update their priors in whichever way they choose. Again, this is not an issue. If they are engaged in probabilistic discourse, they might decide that they can 'rule out' an answer if a suitably low probability has been reached. Or, following an utterance, their credences might actually jump to 1 or 0, depending on how they are updating their priors. In a case like that, the value of $R(\cdot)$ will not matter, because the utterance will count as relevant by the original definition.

Similarly, $R(\cdot)$ lacks an upper bound, and will approach infinity as the probability of a possible answer approaches 0 or 1. This feature once more accentuates the fact that $R(\cdot)$ is useful as part of a bifurcated definition of relevance; it is not suited to compare the relevance of utterances that exclude answers with that of utterances that shift probabilities. We will do best to consider $R(\cdot)$ only in the context of discourse involving probabilities. When discourse moves towards standard exclusion of possible answers, we return to Roberts-style relevance.

4.3 Doubts about the common ground

Next, one might object to the inclusion of probabilities in the **common ground**. In our model, interlocutors share precise probability distributions defined over the possible answers to the **question under the discussion**. We also assume that, upon learning new evidence, the interlocutors update these distributions in the same way. As **Case 1** shows, there *are* paradigmatic cases of probabilistic discourse in which probability distributions are shared in the **common ground**. Moreover, the inclusion of probabilities in the **common ground** is what allows this sort of discourse to happen. Interlocutors can only make movement on such probabilistic questions when they agree on the probabilites involved.

One might respond that, yes, cases of agreement exist, but $\textbf{Relevance}_2$ is limiting: it considers only such cases. In response, I point out that $R(\cdot)$ does not require two or more interlocutors; it measures the relevance that an utterance has to a single agent. It would be simple to examine cases where an utterance has a different degree of relevance for the different agents involved. Interesting work could be done on how to formalize these kinds of cases, but there is nothing about $R(\cdot)$ and $\textbf{Relevance}_2$ stopping them from being so applied.

I want to mention here, too, the potential difficulty posed by *resiliency*, as discussed by Skyrms (1981), Joyce (2005), and others. A prior probability or credence is *resilient* to the extent that it remains unchanged in the face of new data. Two agents might have the same credence in a proposition, and be presented with the same new evidence, but one's posterior credence might change more than does the other's, as a result of the evidence that that agent had updated on in the past. In order to account for this possibility, we merely have to strengthen our assumptions, and require that our agents hold credences over the propositions in question that agree in-regards-to resiliency. As before, this move is consistent with the notion of **common ground**, because interlocutors

engaged in discourse often do have shared priors, do have shared evidence, and so on. This kind of shared background is exactly what allows certain kinds of tandem investigations to progress. And, as I point out above, $\mathbf{Relevance}_2$ and $R(\cdot)$ are still useful in scenarios without agreement.

4.4 Doubts about degrees of belief

One might also worry as to the status of probabilities in the model that I have described. That is: this is a useful description of probabilistic discourse, so long as the probabilities involved are actual degrees of belief. Imagine instead that credences are beliefs *about* probabilities, as, e.g., Buchanan and Dogramaci claim (forthcoming). Applied to **Case 1**, maybe the **question under discussion** would be: (what is, roughly, the probability that Mo has G?), with the possible answers being $(0, .1, \cdots, .9, 1)$, and relevant utterances being those that rule out the possible probabilities.

This kind of strategy just begs the question. The doctors described above will still have to take evidence into account, and will still have to update their priors accordingly, even if they are only changing the probabilities that their beliefs are about, instead of changing their degrees of their beliefs. So, we are left to wonder as to the relevance of the utterance that offered the evidence that changed those beliefs. For example, say the doctors first thought that the right answer to the above question was .5. Then they heard new evidence, which made them decide that the right answer was actually .9. $R(\cdot)$ can still measure the relevance of that evidence. Even in a world in which credences are beliefs about probabilities, and not degrees of belief, there is still room to measure the relevance of utterances that change those beliefs.

5 Concluding Remarks

I end by reiterating my argument. First, interlocutors frequently hold probabilities over answers to questions as **common ground** in conversation. Next, a non-probabilistic definition of the relevance of an utterance fails to capture when an utterance might be probabilistically relevant. To account for this, we introduce a measure of relevance, $R(\cdot)$, along with a new definition: an assertion is relevant when $R(\cdot)$ is larger than a certain amount, where that certain amount is determined by the stake of the discourse.

As a final note, $R(\cdot)$ is interesting beyond the realm of conversational relevance. As I have mentioned prior, the characterization of the KL divergence that I provide is worth noting for formal epistemologists generally. Any epistemologist who desires modeling tools that apply to non-rational or non-Bayesian agents ought to be interested in relevance measures that take probability distributions as their arguments. It is further noteworthy that relevance-considerations characterize the KL divergence and not the Brier Score. In-regards-to information theory and statistics, there is a great deal of work on axiomatic derivations of various divergence measures (see Ebrahimi, 2010, and Csiszar, 1991). Our own derivation adds to this rich literature. Finally, not only is there much stake-sensitivity adjacent work in economics, but there is also a tradition of deriving similar measures from practically-motivated axioms (see Maasoumi, 1986). The derivation of $R(\cdot)$ in this paper adds to this literature as well.

References

[1] Buchanan, R., & Dogramaci, S. (forthcoming). Belief about probability. *Journal of Philosophy.*

[2] Buckwalter, W., & Schaffer, J. (2015). Knowledge, stakes, and mistakes. *Nôus*, 49 (2), 201–234. https://doi.org/10.1111/nous.12017

[3] Carnap, R. (1945). On inductive logic. *Philosophy of Science*, 12 (2), 72–97. https://doi.org/10.1086/286851

[4] Csiszar, I. (1991). Why least squares and maximum entropy? an axiomatic approach to inference for linear inverse problems. *The Annals of Statistics*, 19 (4), 2032–2066. https://doi.org/10.1214/aos/1176348385

[5] Douven, I. (2013). Inference to the best explanation, dutch books, and in- accuracy minimisation. *Philosophical Quarterly*, 63 (252), 428–444. https://doi.org/10.1111/1467-9213.12032

[6] Ebrahimi, N. (2010). Information measures in perspective. 78 (3), 383–412. http://dx.doi.org/10.1111/j.1751-5823.2010.00105.x

[7] Etchart-Vincent, N. (2004). Is probability weighting sensitive to the magnitude of consequences? an experimental investigation on losses. *Journal of Risk and Uncertainty*, 28 (3), 217–235. http://dx.doi.org/10.1023/B:RISK.0000026096.48985.a3

[8] Good, I. J. (1968). Corroboration, explanation, evolving probability, simplicity and a sharpened razor. *British Journal for the Philosophy of Science*, 19 (2), 123–143. https://doi.org/10.1093/bjps/19.2.123

[9] Grice, H. P. (1975). Logic and conversation. In D. Davidson (Ed.), *The logic of grammar* (pp. 64–75). Dickenson Pub. Co.

[10] Jeffreys, H. (1939). *Theory of probability*. Clarendon Press.

[11] Joyce, J. (2005). How probabilities reflect evidence. *Philosophical Perspectives*, 19, 153–178. https://doi.org/10.1111/j.1520-8583.2005.00058.x

[12] Kullback, S. (1959). *Information Theory and Statistics*. Wiley.

[13] Kunreuther, H., Meyer, R., Zeckhauser, R., Slovic, P., Schwartz, B., Schade, C., Luce, M. F., Lippman, S., Krantz, D., Kahn, B., & Hogarth, R. (2002). High stakes decision making: Normative, descriptive and prescriptive considerations. *Marketing Letters*, 13 (3), 259–268. http://dx.doi.org/10.1023/A:1020287225409

[14] Lin, J. (1991). Divergence measures based on the shannon entropy. *IEEE Trans. Inf. Theory*, 37, 145–15. http://dx.doi.org/10.1109/18.61115

[15] Maasoumi, E. (1986). The measurement and decomposition of multi-dimensional inequality. *Econometrica*, 54 (4), 991–997. http://dx.doi.org/10.2307/1912849

[16] McGrew, T. (2003). Confirmation, heuristics, and explanatory reasoning. *British Journal for the Philosophy of Science*, 54 (4), 553–567. https://doi.org/10.1093/bjps/54.4.553

[17] Pettigrew, R. (2016). *Accuracy and the laws of credence*. Oxford University Press UK.

[18] Roberts, C. (2012). Information structure in discourse: Towards an integrated formal theory of pragmatics. *Semantics and Pragmatics*, 5, 1–69. http://dx.doi.org/10.3765/sp.5.6

[19] Schupbach, J., & Glass, D. (2023). Conjunctive explanation: Is the explanatory gain worth the cost? In *Conjunctive explanations: The nature, epistemology, and psychology of explanatory multiplicity*. Routledge.

[20] Simons, M., Beaver, D., Tonhauser, J., & Roberts, C. (2010). What projects and why. *Semantics and Linguistic Theory*, 20, 309–327. http://dx.doi.org/10.3765/salt.v0i20.2584

[21] Skyrms, B. (1981). Causal necessity. *Philosophy of Science*, 48 (2), 329–335. https://doi.org/10.1086/289003

[22] Sperber, D., & Wilson, D. (1986). *Relevance: Communication and cognition*. Blackwell.

[23] Stalnaker, R. (1978). Assertion. *Syntax and Semantics* (New York Academic Press), 9, 315–332.

[24] Weatherson, B. (2005). Can we do without pragmatic encroachment. *Philosophical Perspectives*, 19 (1), 417–44. https://doi.org/10.1111/j.1520-8583.2005.00068.x

[25] Yalcin, S. (2012). Context probabilism. In M. Aloni, M. Aloni, V. Kimmelmann, F. Roelofson, G. W. Sassoon, K. Schulz, & M. Westera (Eds.), *Logic, language, and meaning: 18th amsterdam colloquium, amsterdam, the netherlands, december 19-21, 2011, revised selected papers* (pp. 12–21). Berlin: Springer.