INVESTIGATIONES LINGUISTICAE VOL. XLI, 2018
© INSTITUTE OF LINGUISTICS – ADAM MICKIEWICZ UNIVERSITY
AL. NIEPODLEGŁOŚCI 4, 60-874, POZNAŃ – POLAND

Testing Epistemic Injustice

Elin McCready[1]

AOYAMA GAKUIN UNIVERSITY

mccready@cl.aoyama.ac.jp

Grégoire Winterstein

UNIVERSITÉ DU QUÉBEC À MONTRÉAL

winterstein.gregoire@uqam.ca

Abstract

This work builds on the trivial observation that everyone is not trusted equally. One's gender, ethnic group, occupation etc. will affect how one's information is believed and interpreted by others. We begin by reviewing past approaches to reliability and epistemic injustice, and the factors which affect how one's reliability is evaluated by others in discourse. We then discuss recent experimental results which show that the linguistic manipulation of gender seems to affect the strategies with which the source's reliability is evaluated. We argue that masculine sources benefit from more charitable assumptions than feminine ones. To support this claim, we present the results of a fine-grained categorization task. The results of this task seem to support our claim about charity, i.e. that a masculine source can more easily claim competence about a topic categorized as feminine, whereas the converse appears less true.

1 Introduction

Should one believe the speech of others? This is a fundamental question in pragmatics, and one which has been given a wide range of answers, both

[1] The authors wish to thank Zoe Luk and Regine Lai for their help in preparing and discussing the Cantonese materials used in this work, and Heather Burnett, Henriëtte de Swart and audiences at ICAL and the University of Paris 7 for helpful advice and comments.

within philosophy and in linguistics. However, it is a truism that not all people should be trusted equally; and, while it is perhaps less of a truism, it is clear that some people are trusted less than they should be, and some are trusted too much. This point is as true for groups of people as it is for individuals. In recent years this phenomenon has been dubbed epistemic injustice in the philosophical literature, where the fact that the speech of certain kinds of people is devalued relative to the speech of others has been explored extensively as an aspect of the oppression such groups face in human societies.

This kind of epistemic injustice is prominent for certain ethnic groups, for certain kinds of minorities, and for women, who, while of course not a minority, have been systematically denied power of some kinds in a wide range of human societies. It is often claimed that the speech of women is also devalued, in that the claims of women are ignored relative to the claims of men. The main aim of this paper is to explore the truth of this claim from a linguistic perspective. Judging the reliability or trustworthiness of particular speakers depends on a range of properties of those speakers, as discussed extensively in section 2, where some details of the claims of the literature on epistemic injustice are also elaborated on; in practice, this means that, given specific speakers, it is difficult to isolate changes in epistemic authority (or lack thereof) resulting from their gender as opposed to other confounding factors. Further, the particular topic under discussion can influence the epistemic authority or reliability assigned to a testimonial source: although a group may be deemed unauthoritative in general, they still might be judged as having epistemic authority on a particular topic or set of topics (e.g., though politicians might generally be judged untrustworthy, they might be taken to be very authoritative with respect to safe locations for receiving bribes).

We therefore conducted experiments aimed at isolating shifts in epistemic authority based on gender by means of purely linguistic factors, as detailed in section 3. The main technique used was to present source-based information and ask participants to rate the convincingness of the information provided, where the source was varied for gender by using gendered pronouns or other nominals. In follow-up experiments, we examined the influence of topic bias: certain topics are stereotypically associated with women and men, and we investigated the degree to which such associations push judgements of authoritativeness upward or downward for particular genders. This set of experiments is detailed in section 4. Section 5 concludes and indicates some directions for future research.

2 Source reliability and epistemic injustice

This section provides background on notions of reliability, trustworthiness and epistemic authority in linguistics and philosophy, on how judgements of reliability are formed, and on how they can be skewed in the context of communications by groups assigned low status or actively denigrated by general society. We focus first on philosophical traditions of reliability judgement and communication, starting with a brief discussion of the work of Hume (1977) and Reid (1997), and then turn to the implementation of McCready (2015), which gives a formal treatment of reliability judgements in the context of the theory of repeated games and probability. We will extract some relevant lessons from these theories, which show themselves in a different form in the cases of epistemic injustice we then present.

2.1 Reliability and trust

Without trust, communication is impossible. Since hearers would never believe the content of what is said, there would be no reason for speakers to produce it, and the whole practice would quickly die out; this can be viewed as one motivation for the Gricean Maxim of Quality, which dictates that speakers be truthful to the best of their ability (Grice, 1975). The question of how hearers should behave is somewhat more fraught. One possibility would be to have a comparable principle which we might call the assumption of Quality: assume your interlocutor is being truthful, and believe the content of what they say. But this is a dangerous move from the perspective of the hearer, as it's often the case that speakers intend to deceive or are problematic in other ways which are less malicious but still make it inadvisable to trust their words: for instance, they might just be wrong in what they say. Considerations like these led Sperber et al. (2010) to propose their principle of Epistemic Vigilance, which instructs hearers to be careful about who they take as authoritative in their speech and worthy of trust.

But how exactly is this project to be carried out? There are basically two possible views of the issue. The first, exemplified by Hume (1977), advises a kind of principle of charity, on which one ought to trust one's interlocutor in the absence of reasons not to do so; the second, due to Reid (1997), takes distrust to be the base position and advises us to trust only if we can find reasons to do so (cf. van Cleve 2006). Both of these positions have their supporters in the current literature on the philosophy of testimony.[2] The exact way in which they differ empirically will depend entirely on what we take possible reasons to trust, or not to trust, to consist in.

Let us consider one exemplar theory in some detail, that of McCready (2015), which uses elements of both views. On McCready's view, judgements of reliability are based on information coming from two sources: histories of interaction and initial judgements.

Histories of interaction are, as one would expect from the name, records of the informational exchanges in which the speaker and hearer were both involved. The records contain information about what was communicated, the topic of communication, and whether that information was correct, for each action; the records themselves are constructed on the basis of repeated information exchange games. Taking the proportion of

[2] See e.g Coady (1992), Lackey and Sosa (2006), Lackey (2008) for some summary and discussion.

information exchanges in which the communicated information was correct yields a real number which can in turn be viewed as the probability of the speaker's communicating accurate information in the next round of exchange. That probability can then be used to determine, for a given speaker, receiver, and situation, whether it is advisable to trust the speaker on the next exchange, or not. The theory also allows restriction to particular topics due to the presence of topic-related information in the records, a point we will return to in the next subsection.

Plainly, this method will not suffice for all cases. In the first interaction between agents, there will be no history which can be used to gauge the reliability of the agents' speech. At this point, decisions must be made. One option is the Humean one: trust the other agent in the first interaction, to give the benefit of the doubt. This is in fact an optimal strategy in many situations, as choosing not to trust initially can often lead to a pattern of doubt, which in turn has the capacity to destroy the interaction completely (cf. work on the iterated Prisoner's Dilemma by e.g. Nowak 2006). However, it is also extreme to simply allow trust at any given initial point: surely there are cases in which judging someone reliable is neither descriptively (empirically) or prescriptively (normatively) correct, as when the guy on the street corner offers to sell you the Brooklyn Bridge (for cheap!). The theory should allow for red flags.

The way in which this is handled is to allow initial judgements about the probability that an interlocutor is reliable, based on observations of them and their properties, and on background assumptions. Observing that an individual is standing on the corner making implausible claims about what he's got to sell will (likely) lead to a low initial probability of reliability, while the reputable accountant professionally dressed in her office with degrees mounted on the walls will (likely) be judged more reliable in the initial stage. Such judgements are Reidian, but on both the Reidian and Humean views later interactions are able to alter the judgement about reliability that is made. Still, it's clear that in this theory, as in others, the various observable properties of the individual speaker play a very large role in determining whether that individual is to be trusted. This factor is very common in modern theories of testimony, but comes with certain moral problems, as will be detailed in the next subsection.

2.2 Epistemic injustice

What sorts of properties have an influence on judgements of testimonial epistemic reliability? The particulars depend on the individual theory (and are rarely spelled out in detail), but there are a variety of properties which could be relevant. For instance, in the context of a legal trial, being a witness could lead to heightened degrees of reliability, given that one swears to tell the truth, or one might take adults to be in general more reliable than children, or to have more epistemic authority. Such judgements seem relatively unpernicious, as they are founded on factors that may genuinely increase the probability that a piece of testimony is truth-tracking: in the first case, the threat of external punishment, which has been shown to be effective in increasing cooperativity in game-theoretic contexts (e.g. Gintis 2009), and personal experience in the second case, which is probability-increasing on the not unreasonable assumption that adults are better at forming judgements than children are. Thus, observation of relevant properties is a good strategy for making decisions about testimonial reliability (Fricker, 1995, McCready, 2015).

But not all relevant properties are so benign. The kinds of properties that agents use in judgements of epistemic reliability are of many different types. Some are relevant to epistemic authority and reliability in ways that have clear relations to truth-tracking, as with the cases in the previous paragraph. In other cases, though, the kinds of properties at issue are deemed relevant for less justifiable reasons, or aren't relevant at all but are mere bleedover from other areas of social life. A particularly problematic domain involves judgements about testimonial reliability stemming from the race or gender of the speaker. As Fricker (2007) discusses extensively in a philosophical context, considerations of race and gender and the stereotypes they come with can have a large influence on the degree of credence given to individual testimony.

Within human societies, one often finds biases relating to ethnic groups and gender. For example, in contemporary US society, men are privileged over women and whites/Caucasians are privileged over other ethnic groups; this privilege manifests itself in various ways which prove to be relevant for attributions of testimonial reliability. For example, one often finds ethnic groups or genders associated with stereotypes relating to the (lack of) epistemic reliability: for instance, Asian people might be deemed highly reliable with respect to scientific or mathematical information, giving an upgrade on the credence attached to their testimony on such matters, or women judged to be generally emotional rather than logical, giving a general downgrade on their testimony. A key difference between these two cases should be noted, namely that, in the first case, the upgrade in testimony relates to a particular topic while in the second case the downgrade is applied across the board. A second kind of case involves general distrust for the whole group, leading to the ignoring or general devaluation of their testimony, as one often finds with the testimony of oppressed groups regarding their oppression. We will see this distinction again in the following sections. Fricker (2007) places this whole set of phenomena under the rubric *testimonial injustice*, itself a subspecies of *epistemic injustice*.

This phenomenon is one which is both philosophically interesting and societally important; despite that, it has received little attention from an experimental perspective, much less a linguistic one. Our earlier work was aimed at filling this gap. The next section summarizes some of this earlier research and discusses the continuation of it that is the topic of the present paper.

3 Reliability in argumentation, Bayesian perspectives

In addition to the philosophical work introduced above, the question of the reliability of a speaker has been addressed in argumentation studies. These studies seek to determine either the normative or rational principles which condition the quality of an argument, i.e. how convincing it is (or should) be to an audience (see van Eemeren et al. 2014 for a near comprehensive overview of argumentation studies). A large part of the concerns of argumentative studies is to determine what are the factors which make for a successful argumentation. These include for example the argument scheme being used by the speaker (e.g. argument from authority, or *reductio ad absurdum*) and its conditions of use (Walton et al., 2008), or notions of argument validity either in a strict logical framework or an informal logic one (Johnson and Blair, 2002, Johnson, 2006, Blair, 2011). In addition to these, the question of the reliability of the source of an argument is also considered by some.

In most instances, it seems that, from a normative point of view, the reliability of the source of an argument should not play a role in its evaluation. Traditional approaches of argumentation consider that a good argument should stand on its own independently of who offered it. However, in many instances of real life argumentation, it is obvious that the identity of the source of an argument bears on its acceptability. A judgment coming from an expert in the field will be more trusted than coming from a random person, and many important societal decisions rest on the judgment of experts (e.g. the decision to allow a drug to be distributed). Briñol and Petty (2009) offer a comprehensive overview of the variety of factors pertaining to the source of an argument that might affect how an argument is processed.

One way to account for source reliability is to classify arguments which hinge on the reliability of the source. This is the approach of (Walton et al., 2008) who consider a class of arguments from source which for example contains the argument from authority/position to know, and the *ad hominem* argument (which involves a direct attack on the reliability of a source). They then delineate the conditions under which such arguments are appropriate.

Another way to approach source reliability is found in the Bayesian approach to argumentation fostered in (Oaksford and Hahn, 2004, Hahn and Oaksford, 2006). It offers a rich flexible framework in which to model various aspects related to argumentation, including the effects of source reliability. The basic tenet of that approach is that argumentation aims at raising the belief of an audience about a proposition. To that effect, the speaker asserts a constellation of arguments (or premises) aiming at raising that belief. In traditional Bayesian fashion, beliefs are identified with probabilities, and thus, a piece of evidence e will argue in favor of a conclusion C iff $P(C|e) > P(C)$, i.e. if and only if the belief in C knowing that e is true (the *posterior* belief) is higher than the *prior* belief in C. Using Bayes' rule, one can show that the posterior belief is higher than the prior if and only if the *likelihood ratio* $\frac{P(e \mid C)}{P(e \mid \neg C)}$ is higher than 1, i.e. iff $P(e \mid C) > P(e \mid \neg C)$. The likelihood ratio is thus taken as a measure of the *diagnosticity* of evidence e, i.e. its ability to provide information about the probability of C.

In the Bayesian approach to argumentation, the question of the reliability of the speaker has been investigated by Hahn et al. (2009) and Oaksford and Hahn (2013). The upshot of their approach is that it allows them to take into account the effects and interaction of a variety of factors, including the speaker's reliability, the strength associated with the content of the argument, and the prior belief of the judge of the argument in the conclusion.

Hahn et al. (2009) show how a less than perfectly reliable source of evidence affects the posterior belief of a judge after receiving a piece of evidence. The reliability is simply factored into the likelihood component via marginalization as in (1) (where R stands for the speaker/source being reliable).

1. $P(e \mid C) = P(e \mid c, R) \times P(R) + P(e \mid c, \neg R) \times P(\neg R)$

In the case of a fully reliable speaker the reliability component can be ignored, but whenever the source is not perfectly reliable, then equation 3 (and its counterpart for $P(e \mid \neg C)$) offer a way to directly predict how the posterior belief will be affected by variations in speaker reliability. (Hahn et al., 2009) provide experimental evidence which supports the predictions by this model.

Oaksford and Hahn (2013) use the same framework and approach to source reliability to investigate the case of *ad hominem* arguments, as in (2).

2. *A:* After listening to him, I think it might be possible that Ford cars simply drive better.
   B: Actually, you should be certain that they don't drive better.
   A: Why do you think that?
   B: Because how would he know? He doesn't know the first thing about cars.

Oaksford and Hahn (2013) show that the effectiveness of the *ad hominem* argument notably depends on the prior belief of the speaker in the conclusion. The stronger the belief in C, the less impact the *ad hominem* argument will have. They model this by assuming that in a dialog like (2), agents assume by default that the source of the argument is reliable. This comes as a consequence of a more general *principle of charity* which prompts hearers to make the most of whatever is uttered by a speaker (Wilson, 1959, Davidson, 1974), and which dovetails with the Humean position already mentioned. Again, the authors present experimental evidence which supports the predictions of the Bayesian, notably that pertaining to the importance of the prior belief in the conclusion at stake in the *ad hominem* attack.

3.1 Source reliability and gender

3.1.1 Experimental results

In (McCready and Winterstein, 2017) we introduce experimental evidence which supports the idea that the gender of the source affects its perceived reliability.

Two series of experiments were run among different linguistic communities: American English speakers based in the USA and Cantonese speakers based in Hong Kong. In both cases the participants were presented with scenarios such as (3) which instantiate an argument from authority, i.e. which use the source's reliability as the reason for accepting a claim.[3]

3. A and B are friends. A wants to buy a power drill and is thinking about which one to buy. A wants a high performance drill to perform heavy duty work.
   A: I wonder if this one is a good choice.
   B: I have a friend who says HE knows a lot about power tools, and HE says this model is really powerful.

Participants were asked to judge how convinced they thought *A* was, given what *B* told him. Two conditions were manipulated in the experiment.

The first factor was the gender of the referent. This was manipulated by changing the pronoun in small capitals in (3) to either *he*, *she* in English, or gendered terms for younger cousins in Cantonese.[4]

We chose to manipulate gendered pronouns in order to vary gender judgements for two reasons, both relating to the manner in which information about gender is introduced by pronominals. Two major types of conventional content can be separated out within natural language: so-called at-issue and not-at-issue content. At-issue content is the content of the main claims of the sentence in terms of truth-conditional semantics (see e.g. Heim and Kratzer 1998 for a basic introduction to the notion). Not-at-issue content is just that content which is not at issue: for example, it could be content which is taken for granted, as with presuppositions (Beaver, 1997), or irrelevant to truth completely, as with expressive content (Potts, 2007, McCready, 2010).[5] In pronominals, gender is part of not-at-issue content, though it is currently a matter of debate whether it is presuppositional(e.g. Sudo, 2012) or expressive (e.g. McCready, 2014). Still, the fact that it is not at issue means that it is highly suitable for an experiment of this kind.

[3] The English experiment also involved instances of the ad hominem argument which we will not discuss here.

[4] Several native speakers confirmed that younger cousins do not hold a particular authority in Hong Kong society, that is, they are neither thought to be reliable like elders or looked down upon like younger siblings.

[5] The precise boundaries of these areas are under debate. For a recent view, see Tonhauser et al. (2013).

There are several reasons for this. The first reason involves the experiment itself. Making a claim about the gender of the source which is explicit and possibly unnatural in context makes it possible that, for some participants, the question of gender becomes salient and their attitudes toward this issue could lead to bias in their answers; thus, explicit introduction can be a confounder in this setting. The fact that the content is not at issue means that it is in some sense not open to question, or unchallengeable (cf. Potts, 2005); so it has a good chance of both getting accepted by the experimental participant in a way that is 'under the radar,' and unnoticed, avoiding the worry about confounds. Second, the introduction of a piece of content as not-at-issue also can help to avoid possible biases resulting from participant judgements about the possible (un)reliability of the speaker herself.

The other factor we manipulated was the gender bias of the topic being discussed. This is because some topics are intuitively felt to be more feminine (e.g. *sewing*) or masculine (e.g. *power tools*), given existing social biases. [6] To assess these biases, we ran two preliminary categorization tasks prior to the main experiments. In those categorization tasks, participants had to select the gender they thought was the most linked to a variety of topics. The topics which were the most biased were selected for the main experiment.

Among the results of the experiment we found that there was an interaction between the gender of the source on the convincingness of the argument and the topics' biases, which we interpret as an effect of the gender of the source on its perceived reliability. Unlike what could have been expected, the results do not show a symmetric pattern where masculine sources are trusted for masculine topics and vice-versa for feminine sources. Rather, we observed that masculine sources seemed to be trusted in roughly the same manner for all topics, whereas female sources were distrusted for masculine topics.

3.1.2 Modeling

The experimental results presented above show that when additional evidence about the source of authority is available, participants seemed to discard the principle of charity, i.e. the default assumption of reliability of the source. More precisely, the results suggest that participants kept being charitable with male sources, but modified their impression of reliability with female sources depending on the topic being discussed.

In Fig. 1 we use a Bayesian Belief Network representation (Pearl, 2009) to show the causal links between the different factors involved when evaluating the presentation of a piece of evidence by a source in order to target a conclusion. This representation is a refined version of that proposed by (Hahn et al., 2009) who do not consider factors that might affect reliability. We consider the following variables:

[6] Hereafter we will use the terms *masculine topic* and *feminine topic* for these categories; this terminology should not be taken as an endorsement, but as shorthand for 'topics socially categorized as masculine/feminine.'

- HYP is the hypothesis being discussed, i.e. the goal or conclusion of the argument.
- EVI is the piece of evidence being presented by the source of information.
- REL is the (perceived) reliability of the source of information.
- FTR are the observable features of the source such as its gender, grooming, occupation etc.
- TOP is the general topic being discussed. This variable is also observed when dealing with a particular example of argumentation, i.e. its value is known to the participants.

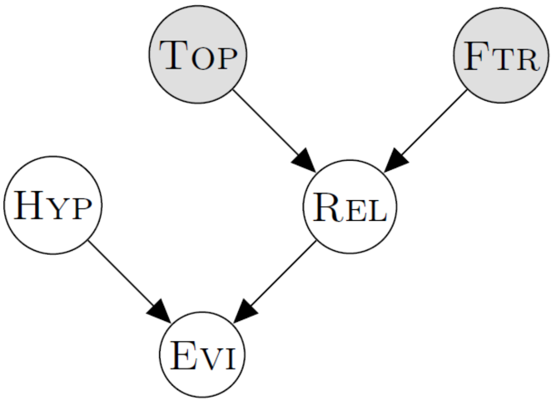


Figure 1: Causal model for evidence reports with factors affecting the source's reliability. Observed variables are greyed out.

The directed arrows indicate a causal link between variables. Thus, the truth of an hypothesis (e.g. having contracted a disease) will affect the probability of generating a given piece of evidence about it (e.g. showing a given symptom) rather than the other way around. Similarly, we assume that the perceived reliability of the source is affected by its features and by the topic being discussed, e.g. some topics are supposed to be known by all or very few etc. Note that the joint probability distribution represented in that network corresponds to the beliefs of the judge, i.e. the causal links represent how they think variables affect each other.

Observe that in Fig. 1, the variable REL d-separates the variables TOP and FTR from the evidence report EVI. This means that the evidence report is independent of the topic and features of the source conditional on the relevance: $(Evi \perp\!\!\!\perp \{Top, Ftr\} \mid Rel)$ (Pearl, 2009). Therefore, our proposal is indeed a supplement to the view of reliability proposed by Hahn and colleagues. If a speaker is charitably assumed to be reliable, i.e. once REL is fixed, then FTR and TOP have no influence anymore. The framework thus offers the necessary flexibility to account for our data: it takes into account the effect of the variations of the speaker's reliability, but also factors which affect that perceived reliability.

To evaluate the cogency of a piece of evidence e for a conclusion H, we consider the probability $P(H \mid e, g_i, t_i)$, with t_i the topic being discussed and g_i the observed gender of the source. The way this probability behaves regarding the prior belief in H is mediated by the likelihood ratio, $\frac{P(e \mid H, g_i, t_i)}{P(e \mid \neg H, g_i, t_i)}$. To integrate the reliability of the source along with the other factors in Fig.1, we can rewrite likelihoods as in (4).

4. $P(e \mid H, g_i, t_i) = P(e \mid H, R, g_i, t_i) P(R \mid g_i, t_i) + P(e \mid H, \neg R, g_i, t_i) P(\neg R \mid g_i, t_i)$

3.1.3 Open issues

The formalization introduced above provides a way to account for some of the experimental results we mentioned. However, there are still some loose ends to fully validate the model.

One issue concerns the question of the categorization of topics as masculine and feminine ones. The categorization tasks we used in our initial experiments were rather coarse: a given topic was set firmly in a gender category, and there was no finer measurement of the affinity of a gender with each topic. This was a simplification: rather than dealing in such absolute terms, it is more appropriate to think about the relation between topics and gender as a probabilistic one, i.e. by asking questions such as “given that the source of information is male, how probable is it that he knows about topic X”, or “given that we’re talking about X, how likely is it that a knowledgeable person about X is male/female?”.

Furthermore, the results suggest that men are judged reliable overall, even for feminine topics. We evoked the possibility that this was due to a charity assumption, but it could also be the case that the strength of the association between gender and topics was weaker in the feminine case. This is something which the categorization task cannot reflect since participants were forced to select a unique category for each topic rather than giving nuanced judgments. Concretely, we need a way to evaluate the quantity $P(R \mid g_i, t_i)$ that appears in 1 in a more subtle way which would be a better representation of the probabilistic relationship we assume between source reliability, source gender and topic. This will be the topic of Section 4.

Another issue has to do with the fact that in 3.1.1, the source of the information is claimed to be knowledgeable about the topic being discussed. We assumed that being knowledgeable about a topic entails being a reliable source about it, but this is disputable. We leave that debate for further work.

4 A finer bias categorization

In this section we present the results of an experiment which aimed at a better understanding and categorization of the explicit gender biases of the topics used in our experiments.

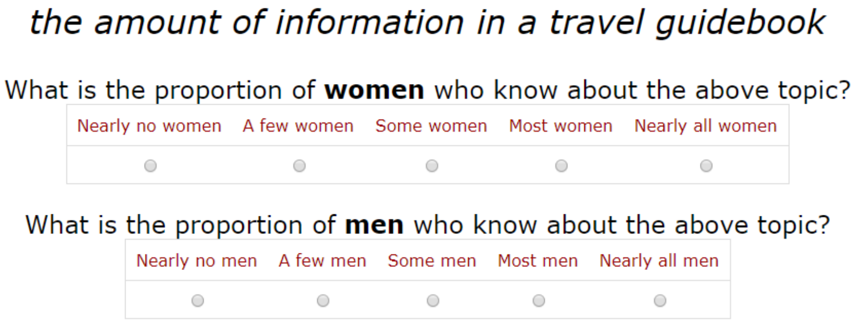
We make the hypothesis that the strength of the causal links between the variables in Fig.1 can be measured, and then used to fit empirical data. Specifically, for a given topic we can obtain an evaluation of the proportions of men and women which are deemed to be knowledgeable, and thus reliable, about this particular topic. Formally, this means that for a topic t_i and gender g_i, we have access to $P(R|t_i,g_i)$: the probability that a source of information is reliable knowing that we're dealing with topic t_i and the source is of gender g_i. Assuming the simplified hypothesis that we are restricted to two genders, measuring the proportion of people of each gender which is knowledgeable about a topic gives us access to the joint probability distribution regulating how reliability is affected by gender and topic.

Beyond this, the experimental results will also be used to compare gender biases about comparable topics between different linguistic communities: English speakers from the USA and Japanese speakers from Japan.[7]

4.1 Materials and Method

Two experiments were run, with a similar structure and methodology, one in Japanese and one in (American) English. 28 voluntary native Japanese speakers participants from Japan (10 female, 18 male) were recruited by snowball sampling for the Japanese experiment. 48 native English speakers participants from USA (23 female, 24 male, 1 other) were recruited through the *Prolific Academic* platform, and paid the equivalent of one British pound for their participation.

The experiment consisted in two on-line questionnaires (one per language, both hosted on the *IbexFarm* platform). After indicating their consent and their gender, participants were presented with 39 topics for which they had to indicate the proportion of women and that of men who they thought is knowledgeable about that particular topic. Figure 2 shows one item from the English experiment.



the amount of information in a travel guidebook

What is the proportion of **women** who know about the above topic?

Nearly no women	A few women	Some women	Most women	Nearly all women
○	○	○	○	○

What is the proportion of **men** who know about the above topic?

Nearly no men	A few men	Some men	Most men	Nearly all men
○	○	○	○	○

Figure 2: Screen capture of one item in the English categorization task

[7] A Cantonese experiment is underway, but results are not available yet.

The 39 topics were chosen using several intuitive criteria. On one hand we stuck to topics with low stakes for the participants in the conversation. A topic that might have life and death consequences (such as the use of a drug or the safety of a car) were left out because such implications are bound to affect judgments of arguments which involve them (typically, the acceptability threshold of the arguments are set much higher). Second, we tried to pick a balanced set of topics in terms of gender biases, i.e. we selected an equal number of topics which we intuitively felt would be judged to be more masculine, feminine or neutral (i.e. known by an equal proportion of men and women). Another point of attention in the selection of the topics was that they had to be relevant to both American and Japanese participants (e.g. baseball coaching), or be of comparable importance (e.g. Disney and Sanrio characters). Finally, the 18 topics used in the experiments discussed by (McCready and Winterstein, 2017) were nearly all part of the materials here with the exception of three of them which either involved high stakes or did not translate well in Japanese.

4.2 Results

4.2.1 Observation of the data

On Figure 3 we plot the average scores for the proportion of men and women who were judged to be knowledgeable about each topic. Each dot is colored according to the author's prior intuition about the topic's bias (`BPrior` = neutral prior, `FPrior` = feminine prior and `MPrior` = masculine prior).

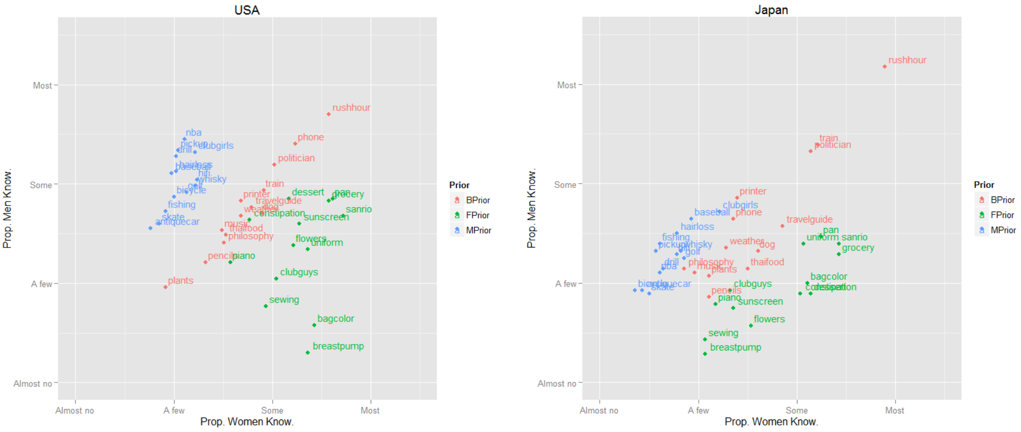


Figure 3: Categorization: main results and comparison with authors' intuitions

A visual inspection of the data shows that, overall, the results fit with the authors' predicted bias of each topic, especially for the USA data. Neutral topics (in red) are along the diagonal of the plot. The blue masculine topics occupy the upper part of the quadrant, corresponding to a higher proportion of men who were thought to be knowledgeable about the topic, and vice-versa for the green feminine ones.

The observation of the results also shows that the feminine and masculine clouds of topics have different shapes, both within and between the two experiments. Generally, the masculine clouds appear more compact, whereas the feminine topics are more spread out, suggesting a higher variability. Another difference seems to be that the Japanese results are more skewed toward the bottom left corner of the graph, meaning that overall participants were more reluctant to attribute knowledge to people in general. We are unsure of the reason for this difference; one explanation is that it has to do with the particular way the questions were posed in Japanese, or that it has to do with the semantics of knowledge attributions in that language. This issue merits further investigation.

4.2.2 Clustering of the data

To better assess the way topics form coherent sets, we used unsupervised clustering techniques on the averaged data sets. The Hopkins statistic (Japan: 0.22 , USA: 0.44) and visual assessment of cluster tendency suggest that there is a moderate tendency of the data to aggregate in clusters. To determine the number of clusters we used elbowing by examining the plot of within groups sums of squares by number of clusters. In combination with the expectation that the number of clusters should be between 3 (one cluster per main gender bias) to 6 (adding one subdivision within each group), the optimal number of clusters for USA was determined to be 6 and 5 for Japan. We then used K-means clustering (MacQueen, 1967) and hierarchical clustering on both data sets. The clusters were mostly stable across methods. Figure 4 shows the results of the hierarchical clustering.

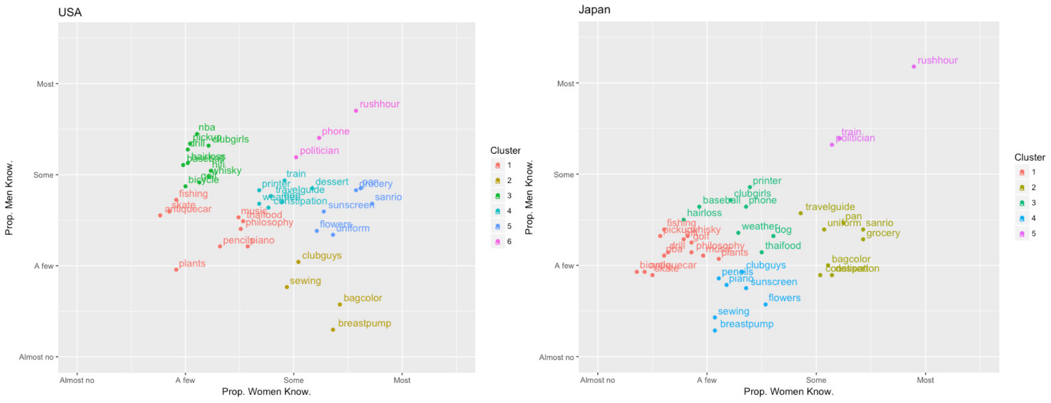


Figure 4: Hierarchical Clustering on the USA and Japan categorization results

The clusters suggest similar conclusions as those above: for each linguistic community the feminine topics are spread out and divided in two clusters. Roughly, one cluster contains topics known by most women and some women (*sunscreen*, *Sanrio* (=*Disney* for USA) etc.) and another known by some women but few men (*sewing*, *breast pumps*)

This bi-partition is not reflected in the masculine topics. For each linguistic community, one clear cluster of masculine topics stands out, with another cluster mixing neutral topics (USA/Japan: *Thai food*) and somewhat masculine oriented topics (USA: *antique cars*/Japan: *baseball*). The shape of this cluster is the only that changes significantly depending on the clustering method being used.

4.2.3 Fitting with previous data

In addition to the grouping of the topics in clusters, we also evaluated the effect of the measured quantities on the convincingness scores measured in the experiment reported by (McCready and Winterstein, 2017). In that experiment, participants indicated how convincing they found an argument based on a claim of being knowledgeable about the topic at hand. Originally, we considered a ternary predictor for convincingness which corresponded to the topic bias (masc./fem./neutral). Here we use instead the values measured in the experiment presented above. Specifically, we consider three independent factors as predictors of the values of convincingness:

- *Gender*: the gender g of the source of information
- *Reliability*: the average proportion of individuals of gender g deemed to be knowledgeable about the topic being discussed (measured on a scale of 1 to 5, obtained in the presently reported experiment).
- *Opp.Reliability*: the average proportion of individuals of gender $\neg g$ deemed to be knowledgeable about the topic being discussed (measured like *Reliability*).[8]

There were thus two data points per topic: one for masculine sources, one for feminine sources, for a total of 28 data points (only the topics present in the convincingness experiments could be used out of the 39 involved in the present study).

We fitted linear mixed models to the data with optimized random structures, using model comparison via likelihood ratio tests to assess the significance of the factors and their interactions.

Given the paucity of the data, we found few significant effects. The variable *Opp.Reliability* does not seem to have any effect, even nearly significant, meaning that participants did not consider the association of the topic with the opposite gender of that of the source when evaluating arguments. However, we observed a marginally significant interaction between *Gender* and *Reliability* ($\chi^2 = 3.05, p = 0.08$). More specifically it appears that for female sources, there is a positive correlation between Reliability and the convincingness of the argument: the more women were judged to be competent on a topic, the more the argument was judged persuasive.

[8] Again this assumes the oversimplifying hypothesis that there are only two mutually exclusive genders.

4.3 Discussion

The results of the refined categorization experiment also confirm that there is a gender asymmetry afoot when dealing with arguments from authority.

A first remark is that the strength with which topics are associated with a gender does not seem to differ between masculine and feminine topics. Therefore, the results discussed by (McCready and Winterstein, 2017) cannot be explained by treating them as artifacts of a difference of that order. This strengthens the hypothesis that the differences are rather due to the gender of the source.

While the biases do not differ in strength, they do however differ in the way they are spread. This asymmetry of the topics' gender biases can be accounted for in at least two ways. On one hand, it can simply come from a badly balanced choice of topics to begin with, i.e. it could well be that with a choice of different masculine topics a bi-partition comparable to the one observed for feminine topics would be found. On the other hand, it could also reflect a genuine difference in how topics are gender biased. This would mean that masculine topics are more homogeneously identified as masculine and known by men in general, whereas feminine topics would not necessarily entail that a feminine source is knowledgeable about them (but rather suggest that if a person is knowledgeable about them, that person is most likely female).

Beyond that observation, we also found that, on our limited data set, the use of refined scores of reliability rather than simple categorical variables tends to support our initial conclusions. The scores we collected are supposed to reflect the proportion of women/men who are knowledgeable about a given topic. As such they represent $P(R \mid g_i, t_i)$ the probability of the source being reliable knowing that the source is of gender g_i and we are dealing with topic t_i. The analysis of the results shows that this quantity only seems to be taken into account with female sources when evaluating the convincingness of the argument.

We already mentioned that one way to account for this is to assume that when the source is masculine the charity principle applies, whereas it does not for a feminine source. This can be slightly refined. The test items of the convincingness experiment involved a claim of competence by the source itself. If that claim is accepted by the participant, then the question of the reliability is settled and its gender and the fit of the gender with the topic have no influence anymore. The differences between masculine and feminine sources can thus be interpreted as differences in how ready we are to believe a source who claims to be competent. Being charitable then means accepting a claim of competence, even about a topic for which prior knowledge tells us that the gender of the source does not make it likely the source is knowledgeable.

5 Conclusion

This paper has discussed certain sources of bias in attributions of epistemic authority and reliability, and presented experimental results designed to investigate these biases and their robustness across cultural contexts. The results open various avenues for future research.

From the experimental point of view, we plan on running more complete experiments with more statistical power, that cover all topics, in the three languages that we are considering (Japanese, Hong Kong Cantonese, English). Hopefully, this should strengthen our hypothesis that different strategies are at play when evaluating arguments coming from differently gendered sources.

Another direction for future work is to consider topics for which *implicit* biases are known (Greenwald et al., 2009). In this work, we measured explicit biases. While this allows us to get a fine grained characterization of the biases by gender, those results are not perfect images of the actual implicit biases which most likely influence participants' judgments. Our results do suggest that, in addition to our explicit biases, participants might apply different heuristics regarding charity according to the gender of the source; so a more thorough investigation of implicit biases appears warranted.

Finally, we will also investigates other source biases. Gender is but one feature that enters into consideration when interpreting an argument. Other elements include other observable features of the source such as their ethnic group, but also features of the interpreter. Most important among these is the alignment of the features of the interpreter with those of the source. Previous work has shown that such group membership alignment can improve the persuasiveness of an argument (Fleming and Petty, 2000). Yet, our preliminary results show that masculine judges might be more distrustful of masculine sources for masculine topics. We will thus closely examine this data, and propose a refined version of the strategies at play in the evaluation of source reliability.

References

Beaver, D. 1997. Presupposition. In: van Benthem, J. and ter Meulen, A. (Eds), *Handbook of Logic and Language*. Amsterdam: Elsevier, pp. 939–1008.

Blair, J. A. 2011. Informal logic and its early historical development. *Studies in Logic*, vol. 4, pp. 1–16.

Briñol, P. and Petty, R. E. 2009. Source factors in persuasion: A self-validation approach. *European Review of Social Psychology*, vol. 20; pp. 49–96.

Coady, C. 1992. *Testimony: a Philosophical Study*. Oxford University Press.

Davidson, D. 1974. On the very idea of a conceptual scheme. In Davidson, D. (Ed.), *Inquiries into truth and interpretation*. Oxford: Oxford University Press. pp. 183–198.

Fleming, M. and Petty, R. 2000. Identity and Persuasion: An Elaboration Likelihood Approach. In Terry, D. J. and Hogg, M. A. (Eds), *Attitudes,* *Behavior, And Social Context. The Role of Norms and Group Membership.* Mahwah, NJ: Lawrence Erlbaum, pp. 171–199.

Fricker, E. 1995. Telling and trusting: Reductionism and anti-reductionism in the epistemology of testimony. *Mind*, vol. 104, pp. 393–411.

Fricker, M. 2007. *Epistemic Injustice*. Oxford: Oxford University Press.

Gintis, H. 2009. *The Bounds of Reason: Game Theory and the Unification of the Behavioral Sciences*. Princeton: Princeton University Press.

Greenwald, A. G., Poehlman, A., Uhlmann, E. L., and Banaji, M. R. 2009. Understanding and using the implicit association test: III. meta-analysis of predictive validity. *Journal of personality and social psychology*, vol. 97(1), p. 17.

Grice, H. P. 1975. Logic and conversation. In Cole, P. and Morgan, J. L. (Eds), *Syntax and Semantics Volume 3: Speech Acts*, Academic Press. pp. 41–58.

Hahn, U., Harris, A. J., and Corner, A. 2009. Argument content and argument source: An exploration. *Informal Logic*, vol. 29(4), pp. 337–367.

Hahn, U. and Oaksford, M. 2006. A Bayesian approach to informal argument fallacies. *Synthese*, vol. 152, pp. 207–236.

Heim, I. and Kratzer, A. 1998. *Semantics in Generative Grammar*. Number 13 in Blackwell Textbooks in Linguistics. Blackwell, Oxford,England.

Hume, D. 1977. *An Enquiry Concerning Human Understanding*. Hackett. First published 1748.

Johnson, R. H. 2006. Making sense of informal logic. *Informal Logic*, vol. 26, pp. 231–258.

Johnson, R. H. and Blair, J. A. 2002. Informal logic and the reconfiguration of logic. In Gabbay, D., Johnson, R., Ohlbach, H., and Woods, J. (Eds), *Handbook of the logic of argument and inference: Turn towards the practical*, Amsterdam: Elsevier. pp. 340–396.

Lackey, J. 2008. *Learning from Words: Testimony as a Source of Knowledge.* Oxford: Oxford University Press.

Lackey, J. and Sosa, E., editors 2006. *The Epistemology of Testimony*. Oxford: Oxford University Press.

MacQueen, J. 1967. Some methods for classification and analysis of multivariate observations. In California Press, U. (Ed.), *Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability*, Berkeley. pp. 281–297,

McCready, E. 2010. Varieties of conventional implicature. *Semantics & Pragmatics*, vol. 3, pp. 1–57.

McCready, E. 2014. A semantics for honorifics with reference to Thai. In Aroonmanakun, W., Boonkwan, P., and Supnithi, T. (Eds), *Proceedings of PACLIC 28*, Chulalongkorn University. pp. 513–521.

McCready, E. 2015. *Reliability in Pragmatics*. Oxford: Oxford University Press.

McCready, E. and Winterstein, G. 2017. Negotiating epistemic authority. In Kurahashi S. et al (Ed.), *New Frontiers in Artificial Intelligence*. Springer, Berlin. to be published.

Nowak, M. 2006. *Evolutionary Dynamics*. Belknap Press.

Oaksford, M. and Hahn, U. 2004. A Bayesian approach to the argument from ignorance. *Canadian Journal of Experimental Psychology*, vol. 58, pp. 75–85.

Oaksford, M. and Hahn, U. 2013. Why are we convinced by the ad hominem argument?: Bayesian source reliability and pragma-dialectical discussion rules. In Zenker, F. (Ed.), *Bayesian Argumentation*, Springer, NL. pp. 39–58.

Pearl, J. 2009. *Causality: Models, Reasoning and Inference.* Cambridge University Press, New York, 2nd edition.

Potts, C. 2005. *The Logic of Conventional Implicatures*. Oxford: Oxford University Press. Revised version of 2003 UCSC dissertation.

Potts, C. 2007. The expressive dimension. *Theoretical Linguistics*, vol. 33, pp. 165-198.

Reid, T. 1997. *An Inquiry into the Human Mind on the Principles of Common Sense*. Edinburgh University Press. Originally published 1764.

Sperber, D., Clemeant, F., Heintz, C., Mascaro, O., Mercier, H., Origgi, G., and Wilson, D. 2010. Epistemic vigilance. *Mind & Language*, vol. 25, pp. 359–393.

Sudo, Y. 2012. *On the Semantics of Phi Features on Pronouns*. PhD thesis, MIT.

Tonhauser, J., Beaver, D., Roberts, C., and Simons, M. 2013. Toward a taxonomy of projective content. *Language*, vol. 89, pp. 66–109.

van Cleve, J. 2006. Reid on the credit of human testimony. In Lackey, J. and Sosa, E., editors, *The Epistemology of Testimony*, Oxford: Oxford University Press. pp. 50–74.

van Eemeren, F. H., Garssen, B., Krabbe, E. C. W., Snoeck Henkemans, A. F., Verheij, B., and Wagemans, J. H. M. 2014. *Handbook of Argumentation Theory*. Dordrecht: Springer.

Walton, D. N., Reed, C., and Macagno, F. 2008. *Argumentation Schemes.* Cambridge: Cambridge University Press.

Wilson, N. L. 1959. Substances without substrata. *The Review of Metaphysics*, vol. 12(4), pp. 521–539.