Negotiating Epistemic Authority[⋆]

Eric McCready and Grégoire Winterstein

[1] Department of English
Aoyama Gakuin University
mccready@cl.aoyama.ac.jp
[2] Department of Linguistics and Modern Language Studies
The Education University of Hong Kong
gregoire@eduhk.hk

Abstract. Why do we trust what other people say, and form beliefs on the basis of their speech? One answer: they are taken to have *epistemic authority*. Intuitively this means that the other person (or institution, or group) is taken to be authoritative in what they say, at least with respect to a particular domain. Here, we want to claim that there are (at least) two varieties of epistemic authority, one based on reliability and one on assuming (nonepistemic) authority. We claim that both are subject to linguistic negotiation. This paper begins by reviewing McCready's (2015) theory of reliability, and then turns to strategies for attempting to assume epistemic authority, focusing on those involving the use of not-at-issue content. We then show the results of two experiments which test the interaction of stereotypes about gender with epistemic authority, and how this is mediated by language use, focusing on the case of gendered pronouns. Finally, the results are explored for Bayesian views of argumentation and analyzed within McCready's Reliability Dynamic Logic.

1 Introduction

Why do we trust what other people say, and form beliefs on the basis of their speech? One answer: they are taken to have *epistemic authority*. Intuitively this means that the other person (or institution, or group) is taken to be authoritative in what they say, at least with respect to a particular domain. Here, we want to claim that there are (at least) two varieties of epistemic authority, one based on reliability and one on assuming (nonepistemic) authority. We claim that both are subject to linguistic negotiation. This paper begins by reviewing McCready's (2015) theory of reliability, and then turns to strategies for attempting to assume epistemic authority, focusing on those involving the use of not-at-issue content. We then show the results of two experiments which test the interaction of stereotypes about gender with epistemic authority, and how this is

[⋆] The authors would like to acknowledge the support of JSPS Kiban C Grant #16K02640, which partially supported this research, and to thank Heather Burnett, Christopher Davis, Michael Erlewine, Regine Lai, Zoe Luk, Makoto Kanazawa, Lukas Rieser, Henriette de Swart, Christopher Tancredi, and audiences at LENLS13, the Education University of Hong Kong, Keio University and the University of Paris 7.

mediated by language use, focusing on the case of gendered pronouns. The first experiment concerns English and the second Cantonese. Finally, the results are explored for Bayesian views of argumentation.

2 Passive assumption of authority

One way to be authoritative, in the sense of having one's speech consistently believed, is to be a speaker who is judged reliable with respect to speaking truth. If one is judged reliable, one is likely to have a kind of epistemic authority, in the sense that the things one says are likely to be believed. Here, reputation is key given that belief is a form of cooperation; it is known, for the game-theoretic case that the use of reputation in strategizing in repeated Prisoner's Dilemma [18, 19] yields extremely good results, and is therefore likely to be evolutionarily stable.

One way to model reputation with respect to reliability is given by [13], which we will briefly summarize. On this theory, reputations can be derived in part from *histories*, defined as sequences of objects $act \in A, A$ the set of possible actions for a given agent in a given (repeated) game. These objects are records of an agent's actions in past repetitions of the game. Game histories are n-tuples of sequences of records representing the history of the agent's actions at each decision point. For the case of communication, these are of course histories of speech acts. A player's reputation in a game is derived from his history in that game. A player's reputation with respect to some choice is defined as his propensity, based on past performance, to make a particular move at that point in the game. Such propensities are computed from frequencies of this or that move in the history. Specifically, the propensity of player a to play a move m in a game g at move i is: the proportion of the total number of game repetitions that the player chose the action m at choice point i.

$$F_{H_a^{g,n}}(move) = \frac{card(\{act \in H_a^{g,n} | act = move\})}{card(H_a^{g,n})}$$

Always, $0 \leq F_{H_a^{g,n}}(move) \leq 1$, so the above number can be viewed as a probability: in effect, the information that the game participants have about a's likelihood of choosing move m.

An agent's propensity to play a strategy is a real number in [0, 1]. This fact supports a scalar view of propensities, and indeed of cooperation itself: an agent has a propensity for using strategy σ iff i.e. the contextual standard for having that propensity [9]. Thus,

$$Prop(a, \sigma) \text{ iff } F_{H_a^g}\sigma > s,$$

where s is the contextual standard for propensity-having. These propensities can also be used to decide whether to assign someone epistemic authority with respect to some claim. In the context of the repeated PD, [18, 19] make use of reputations and find that there are optimal strategies, given an index of reliability (here in [0,1], but for them in the range 1–5), is to trust if, for example, a has a propensity for reliability (where this for them amounts to setting some arbitrary number above which cooperation is dictated), or if $\sum_{Coop(\sigma)} F_{H_a^g}\sigma$ is above some threshold (not necessarily s) (for the sum of frequencies of all a's cooperative strategies), or if the other agent's reliabilty index is higher than yours. Since such strategies are public, the other agent has an incentive to maintain her R-rating high: i.e. to genuinely be reliable. Any of the above seem reasonable bases to choose to accept someone's epistemic authority, or not.

The above must be combined with other information about reliability. This is so because of the need to decide whether to give someone epistemic authority even in the first communication, before any kind of history is available. This decision corresponds closely to the distinction between Humean and Reidian views on trust in testimony [13, 15]. One way to model the Reidian view, on which decisions about trust aren't made automatically but rather on the basis of some metric, is that of [4], who takes speakers to make judgements about people's epistemic authority based on stereotypical information about factors like their gender, race, occupation, and personal grooming. This seems sensible: one might be more likely to believe a clean-shaven man in a suit about his having had his wallet stolen and needing money for the train than the same statement made by a homeless woman carrying a bottle in a brown paper bag (depending of course on one's other beliefs). This heuristic gives a first guess about reliability which can then be modified by interaction.

All this can be embedded in a more general model of information change; [13] proposes a new flavor of dynamic semantics for this purpose [6]. The basic idea is to virtually always update with content acquired from any source, but only 'conditionally.' To make this work, information states σ are complex and consist of possibly many substates. Each IS is a set of worlds (simplification), ordered with a 'plausibility ranking' reflecting epistemic preferences on states. Each substate is indexed by an index $j \in \mathsf{Source} \cup \mathcal{A}$. Here Source is the set of evidence sources and $\mathcal{A}$ the set of agents, which is constrained to only hold indices which the epistemic agent has had experience with. This set is ordered by a total ordering $\leq_a$, where $i \prec j$ iff $P(Rel(i)) < P(Rel(j))$, when $P(Rel(i))$ is the probability that source i yields reliable information.

Updates are of the form $E_i\varphi$, for E_i an operator indicating source in i-type evidence. A sentence $E_i\varphi$ always induces update of state σ_i. Some cases are indeterminate cases, such as the use of direct evidentials in some languages that have them, where it may not be clear what the source is: visual, auditory, . . . In such cases, all possible substates are updated. But in the testimonial case, states indexed with agentive sources a are updated. So, at the level of substates, update with φ always takes place when φ is observed — but this is *not* the same as coming to believe φ at a global level. Global beliefs are defined on the global state σ_T resulting from unifying all substates σ_i. This unification is done via a merge operation ($\Cap$): all substate content survives when non-contradictory, but in case of conflict, information from higher-ranked sources trumps lower-ranked source-indexed information. Thus the global state almost never exhibits conflicts; it only will if two sources are precisely equally ranked, which is unlikely given the range of real values, but can be explicitly banned by enforcing a version of Lewis's Limit Assumption, here for sources rather than worlds [10].

More formally, global information states σ: consist of sets of elements (substates) of the form $\sigma_i = \langle X, \leq_a \rangle$ where $X \subseteq W$ (the set of states). The substates are plausibility frames in the sense of [1, 2]: multi-agent Kripke frames $\langle X, R_a \rangle_{a \in \mathcal{A}}$, where the accessibility relations R_a are called 'plausibility orders', written $\leq_a$, and assumed to be locally connected preorders. This simplifies a bit: sometimes the substates can be more complex, in particular in the case of testimonial agents, as the substates associated with them also have a similar structure. Total information states are written σ_T, and are of the form $\langle X, \leq_a \rangle$ for $X \in \wp(W)$. They are derived by recursively merging all plausibility relations found in $\sigma_i \in \sigma$ via a lexicographic merge operation, which respects priority ordering; so an agent's beliefs thus are derived on the basis of the most reliable source, and so on down the source hierarchy. From this, we get resolution in cases of conflicting sources.

Update in this system follows the $[.]_{\Uparrow}$ of [1, 2], defined as follows.

- $\sigma[\varphi]_{\Uparrow} = \sigma'$, where $S' = S$ and $s \leq'_a t$ iff either (i) $s \notin \varphi$ and $t \in s(a) \cap \varphi$, or (ii) $s \leq_a t$.

This definition thus leaves the set of states the same, but upgrades those states which satisfy φ above those which don't, otherwise leaving the relative plausibilities untouched. Using this operation ensures that substates will be comparable without recourse to revision.

Support and entailment are defined as follows. A total information state $\langle X, \leq_a \rangle$ is said to *support* a proposition φ, $\sigma \models \varphi$, iff $\{s \in X | s \in best_a(s(a))\} \subseteq \phi$, where $best_a\phi := \{s \in \phi | t \leq_a s \text{ for all } t \in \phi\}$.[3] The definition of entailment is the standard fixed-point dynamic one modulo the use of $[.]_{\Uparrow}$, as defined above (with ';' dynamic conjunction as usual):

$$\phi_1, \ldots, \phi_n \models_\sigma \psi \text{ iff } \sigma[\phi_1]; \ldots; [\phi_n] = \sigma[\phi_1]; \ldots; [\phi_n]; [\psi].$$

Evidential update is defined via the following clause, which ensures that only the substate corresponding to the information source is updated, and all others are left alone.

1. $\sigma[\mathsf{E}_i\varphi] = \sigma'$ where, for all $\sigma_j \in \sigma$, $\begin{cases} \sigma'_j = \sigma_j[\varphi] & \text{if } i = j \\ \sigma'_j = \sigma_j & \text{if } i \neq j \end{cases}$

For an example, suppose agent a learns φ ='It is raining' from evidence source b (agent b). Then: $\sigma' = \sigma$ except that $\sigma'_b \in \sigma' = \sigma_b[\varphi]$, by the definition of evidential update.

Thus: in all cases, the result of evidential update with φ is belief in φ. But this belief may just be belief relative to the source, i.e. within σ_i for source i. 'Genuine' belief requires global belief wrt the global state. Essentially: $B_a\varphi$ iff $\{s \in \sigma_T | s \in best_a(s(a))\} \subseteq \phi$, where $best_a\phi := \{s \in \phi | t \leq_a s \text{ for all } t \in \phi\}$. The total belief state is derived by lexicographic merge, so the content of our examples will be believed unless some higher-ranked source disagrees. What happens when a conflict arises? Consider a case of conflicting agents. Agent a claims ϕ and agent b claims $\neg\phi$. a, let's suppose, is pretty trustworthy. b is unknown; let's suppose that he looks somewhat untrustworthy. The result is that $a > b$ in the priority ordering for lexicographic merge. Thus the merge of σ_a and σ_b verifies ϕ.

So far: update of substates, substates unified via merge, merge priority determined by ordering. But what's the source of the ordering? Without a substantive theory of how

[3] Note that this is essentially identical to the definition of belief in [2].

the ordering is derived, the theory seems to have little empirical content. The claim of [13] is that the ordering is probability-based. The probabilities in question are probabilities of *reliability*. They indicate the (perceived) likelihood that information derived from the source is correct.

These probabilities arise from two factors. The first factor is experience with reliability of the source, as derived from histories; the second is the initial probabilities of reliability. These come in two types: prior beliefs about the reliability of different evidence sources, and beliefs about the reliability of the providers of testimony based on various aspects of their presentation. For an example of the first, one generally can take direct evidence to be more reliable than hearsay: if I see that it's raining outside, I am likely to discount the fact that this morning's weather report said it would be sunny. For the second, as mentioned above, judgements about the reliability of individuals are often made on the basis of stereotypical factors about their appearance and how they are categorized [4]. One might judge the kempt to be more reliable than the unkempt, the professional to be more reliable than the amateur, or someone from the same social group as you to be more reliable than someone from an outgroup. As we'll see in the next section, these kinds of judgements can be manipulated, yielding effects on the attribution of epistemic authority.

The two factors are taken to interact as follows: given an initial probability and a sequence of events of information acquisition, conditionalize on the initial probability for each new acquisition event, with respect to truth-tracking. The idea is to modify the probability that the source is reliable based on whether the new information is correct or not:

$$\frac{P_I(R \cap C)}{P_I(C)}$$

The whole notion of authoritativeness analyzed here is (in a sense) a passive one. One becomes authoritative by speaking the truth and by looking reasonably trustworthy. This is a kind of authority acquired by being a good citizen in the testimonial sense, essentially that of [5]. But is there a more active way to acquire epistemic authority by linguistic means? We think yes: by use of argumentative and other linguistic devices. Some of these will be explored in the next section.

3 Using expressive content for authority negotiation

How can one actively try to acquire epistemic authority (or deny it to others), as opposed to simply acquiring it by living a virtuous testimonial life? One way, of course, is just to assert one's authority:

(1) (You should believe me because) . . .
   a. I know all about this topic.
   b. I'm your teacher.
   c. I'm your dad.

This strategy will be effective to precisely the degree that the speaker already has epistemic authority, because in the absence of epistemic authority, either the hearer won't accept what is said ((1)a), or the speaker's external authority is already rejected ((1)b,c).

Consequently, a less direct strategy (or set of strategies) is needed. In the remainder of this section, we examine the use of expressive content [21] in the assumption of epistemic authority, considering several cases.

We are choosing to focus on expressive content for two reasons. Expressive content is often talked about as 'inflicted' on the hearer [21, 16], which means (if correct) that the content of the expressive cannot easily be contested. This is an important feature when it comes to manipulations of epistemic authority (and in argumentation in general), as it removes the need to have epistemic authority already in order to have one's claims accepted, as with ((1)) above. This feature is not universally present in not-at-issue content either; [26] notes that presuppositions for example can be challenged in discourse, meaning that their content lacks the key feature of expressives we are interested in here. The second reason is the close connection of many expressives to social meanings, which are obviously relevant for epistemic authority. This point will be detailed as we proceed.

In this section, we will briefly consider the cases of particles, honorifics, and, finally, our main concern, those expressives which serve to indicate membership in various social groups.

First, particles like the Japanese *yo* (with falling intonation) work to try to 'force' the hearer to accept the content of the sentence [11, 3]. Indeed, [17] presents an analysis of this particle in terms of epistemic authority. His idea is that *yo* indicates that the speaker has at least as much epistemic authority as anyone else present with respect to the content of the sentence. This implies that the particle can be used strategically to try to claim such epistemic authority for the speaker; use of the particle (if unchallenged) indicates that the speaker already has epistemic authority.

This view has some empirical effects. In the following example, the speaker requests belief via the claim of teacherhood.

(2) watashi-wa anata-no sensei desu yo
1P.Formal-Top 2P.Formal-Gen teacher Cop.Hon PT
'I am your teacher, don't forget.'

However, the use of strengthening *yo* implicates that the speaker doesn't have authority already, which further implies that the speaker takes his epistemic authority qua teacher to be insufficient, resulting in a failed authority grab. Compare here the observation of [25] that falling *yo* infelicitous in e.g. instructions from a commanding officer in the army, because the attempt at claiming authority represented by *yo* (in the terms of this paper) is not compatible with the presence of absolute authority.

The second case is honorifics, which, although they on a separate dimension from epistemic claims (at least according to [8, 22, 12, 14], and others), to the extent that one's social status influences her epistemic authority the use of (anti-)honorifics should count as a strategy for assuming it, or taking it from others. Notably: 'raising' the addressee could cede some epistemic authority to them. In terms of examples, while the following are both grammatical and felicitous, there is a sad mismatch between content, honorific tone and particle: it's as if the speaker is desperately trying to assert himself. This is unlikely to yield genuine epistemic authority.

(3) watashi-no itteiru koto-o shinjite kudasai yo
1P.Formal-Gen saying thing believe please.Pol PT
'Believe what I'm saying, please.'

vs. the pure authority grab:

(4) ore-no itteru koto-o shinjiro
1P.Inf-Gen saying thing believe-Imp
'Believe what I'm saying!'

Finally, many expressives tag aspects of character which can be relevant to determinations of epistemic authority via social status; we can call these *social expressives*. This strategy is less direct than the above in that it is entirely a side effect. The main method here is to ascribe other individuals membership in groups which are or are not privileged in a social sense, and use that (lack of) privilege to implicate something about their epistemic authority. The same is true for slurs: by placing the addressee or other individual in a subordinate group, explicitly or implicitly (cf. [24]), it becomes possible to emphasize one's own epistemic authority over them. It is widely noted in the feminist philosophy literature (and elsewhere on the internet etc.) that the overt or covert primary position of males in society, and their consequent authority, can lead to differences in epistemic authority as well. For instance, the claims of men are often believed over the claims of women, all else being equal. If this is true, the use of e.g. gendered 2P pronouns in situations where other options are available (cf. [23]) could lead to the changes in who is taken to have epistemic authority, meaning that the use of gendered language can be a strategy for its assumption.

Here, we are interested in testimony: the main question in ceding epistemic authority involves how one should assign probabilities of likely reliability to individuals.

As mentioned above, [4] cites one technique, which is to make use of stereotypes about groups, for example that 'women are not logical', 'Asians are well educated', and so on; she presents some compelling examples of such cases, though examples which operate at the level of at-issue claims rather than expressive implications. However, many expressives tag aspects of character which can be relevant to determinations of epistemic authority via social status. We can call these *social expressives*; they are mainly terms which categorize individuals into categories that — at least on a stereotypical or prejudicial level — are relevant to the (non)attribution of epistemic authority. The basic method is to ascribe other individuals membership in groups which are associated with some stereotype, and then use that (lack of) privilege to implicate something about their epistemic authority.

Two examples of social expressives are slurs and gendered language. By definition, slurs are negative and subordinating (cf. [24]), so can be used to emphasize one's own epistemic authority over categorized individual, given that other relevant individuals share the prejudices the slurs express. With gendered language, the situation is more subtle, because gender is not in any sense pejorative in the way of slurs. Still, the deployment of stereotypes about gender to acquire epistemic authority. It is a truism (and a common claim in feminist philosophy as well [4]) that the overt or covert primary position of males in society, and their consequent authority, can lead to differences in epistemic authority as well. For example, it is often said that the claims of men are often believed over the claims of women, all else being equal. If this is true, the use of e.g. gendered 3P pronouns could easily lead to the changes in who is taken to have epistemic authority, meaning that the use of gendered language can be a strategy for its assumption. In order to see whether this is correct, we conducted several experiments, focusing on the use of gender stereotypes in argumentation.

4 Experiments: gender in argumentation

4.1 Experiment 1: English

We ran an experiment to test the relation between gendered speech and epistemic authority in argumentation. We tested two different types of argument which involve the authority of a source: the direct, or abusive, form of the *ad hominem* argument and the argument from authority (or position to know). Schematically these arguments are as follows [27]:

- Ad-hominem:
  - Source a is a person of bad character / has bad character for veracity
  - a argues that α
  - **Conclusion**: α should not be accepted
- Argument from authority (position to know):
  - Source a is in a position to know about things in a certain subject domain S containing proposition α
  - a asserts that α is true
  - **Conclusion**: α is true

In each case, the source a is part of one of the premises of the argument.

The goal of the experiment was to test whether manipulating the gender of the source induces a difference in the convincingness of the argument. We followed a protocol similar to the one used by [7] to investigate the argument from authority.

First, a preliminary experiment was run to determine three distinct sets of topics according to their gender bias. This was done as a categorization task on Amazon Mechanical Turk. Participants were presented with a topic and asked to choose which category most closely matched that topic: `Men, Women` or `Both`. 17 topics in total were tested, out of which 15 were selected, 5 for each gender category. Each topic had an agreement of 80% or above, meaning that four participants agreed the topic was associated with the relevant category. Participants could categorize multiple topics and were paid 0.05 USD for each categorized topic.

These topics were then used to produce 15 distinct arguments, in two forms: the *ad hominem* one, and the argument from authority one. Examples of each form follow (using a male biased topic):

- Authority argument
  - 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.
- *Ad hominem*
    - 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 heard from Jamie that this model is really powerful.
    - *B:* She doesn't know anything about it.

The factors investigated in the experiment were thus the following three:

- `Arg.type`: the type of argument being used (two levels: `ad-hom./auth.`)
- `Source`: the gender of the source of the information, indicated by the use of a gendered pronoun (*he, she*) or *that friend/Jamie* to use a neutral reference (three levels: `maleSrc/femSrc/neutSrc`).
- `TopicBias`: the gender bias of the topic, based on the results of the preliminary categorization task (three levels: `femB/maleB/neutB`).

450 US-based participants were recruited on the Amazon Mechanical Turk and paid 0.2 USD for their participation. They judged the convincingness of 5 different arguments (4 fillers+1 target item) presented in pseudo-random order. Convincingness was rated on a 5 point Likert scale. Linear mixed effect models with maximal random effect structure were fitted to the data using the `lmer` package in `R`. The effects of condition and group were confirmed by likelihood-ratio tests.

Results The results are shown in Fig. 1. They show that, generally, authority arguments are judged more convincing than *ad hominem* (Fig. 1, left panel, $\chi^2 = 145.38, p < 0.01$) and that the gender of the source and the gender bias of the topic have no main effect. Further analyses showed that these variables have no effect in the case of the *ad hominem* argument (Fig. 1, middle panel). However, the results of the argument from authority show that there is a significant interaction between the gender of the source and the gender bias of the topic (Fig. 1, right panel, $\chi^2 = 11.023, p = 0.026$). It was observed that, as expected, men are generally more trusted for topics biased towards men (in the `mascB` case, the difference between the masc-source and neutral-source is significant, $W = 168.5, p = 0.005$) but that women are not more trusted than men for topics biased towards women, and that there was no significant preference for neutral topics.

Discussion To explain why authority arguments are preferred to *ad hominem* ones, we argue that when considering authority arguments the only question is how reliable the source of the argument is. The reliability of the speaker is not directly relevant. This can readily be integrated in an approach like that of [7, 20] who propose a Bayesian treatment of argumentation. In that approach, the convincingness of an argument is proportional to how much the content of the argument affects the audience's prior belief in the conclusion targeted by the argument. The reliability of the source is factored in the likelihood of using an argument a to target a conclusion C. There the speaker's reliability remains constant across possible sources and does not weigh into the evaluation of the argument.

However, in the case of the *ad hominem* argument the speaker's reliability is at odds with that of the source, which might explain why those arguments are generally dispreffered since they pit the speaker's credibility against that of the source. We make

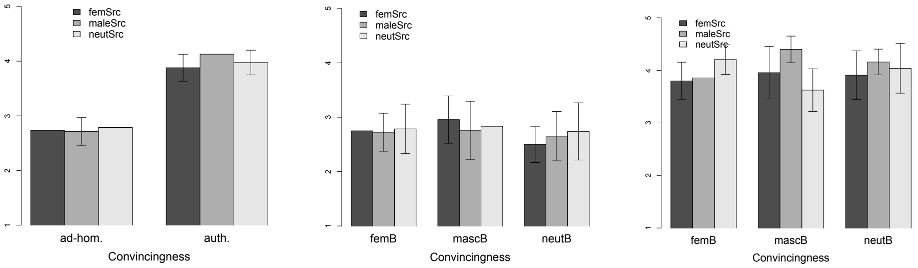

Fig. 1: Judgments of convincingness for `ArgType` vs. `Source` (left panel); `TopicBias` vs. `Source` for the *ad hominem* argument (middle panel) and authority argument (right panel).

the hypothesis that people are generally reluctant to overtly endorse arguments which directly attack other people's credibility, even though they might actually be unconsciously persuaded by them.

As stated above, gender biases can be integrated into the Bayesian approach of argumentation. This amounts to modifying the belief that the source is reliable by conditionalizing on its gender. However in the Bayesian approach the *ad hominem* and authority arguments are seen as dual to each other: one lowers the reliability of the source while the other increases it. As such, it should be expected that both forms would equally be affected by gender biases, contra the results of our experiment. One way to model that difference is to explicitly distinguish between the reliability of the speaker and that of the source of an information in the way an argument is evaluated, and account for the fact that they may potentially be at odds. The approach lends itself to such a modification, but further experimentations are needed to validate whether this move is an effective way to account for the data presented here.

4.2 Experiment 2: Cantonese

A second experiment similar to the one presented above was run using Cantonese material rather than English. This second experiment used a within-participants design, meaning that participants saw examples of each condition to be tested rather than just one single condition. The experiment aimed at reproducing the results of the English experiment with participants from a different socio-cultural background, and also attempted to overcome some of the flaws of the first experiment. First, the English experiment did not control for the stakes involved in the arguments. Some topics might have been interpreted as involving life or death situations (e.g. the safety of a car) while others were much more trivial (e.g. the authenticity of Japanese food). This was controlled for by only using topics which intuitively involved low stakes. Second, we ignored the case of neutral sources of information. This is because the value of such cases is difficult to interpret, as inaccessible participant biases, assumptions, and interpretation metrics may play roles in how the stimuli are processed. It is plausible that participants attributed a gender to the source matching the bias of the topic being discussed (e.g. male in the case of power tools), but there is no way to make sure of it. Third, since the *ad hominem* argument yielded no results and our analysis hypothesizes that it involves more complex reasoning, we only focused on the authority argument in this experiment. Fourth, the gender of the participant was also included in the analysis of the results.

As with the English experiment, a preliminary categorization task was run. Eleven voluntary participants, all native speakers of Cantonese, were recruited. They were shown 24 concepts paired with a property (e.g. the performance of a power drill) were shown to participants in Cantonese and they had to select the category which fitted the topic the best (`Men, Women, Both`). The 12 items with the highest agreement scores level were selected for the core experiment (4 in each category).

For the core experiment, we thus considered the following independent factors:

- `Source`: the gender of the source of the information (`mascSrc/femSrc`), marked by the use of gendered terms for older cousins (*biu2go1*/表哥 for male cousin, *biu2ze2*/表姐 for female cousin). Older cousins were chosen because they hold no intrinsic authority (unlike older siblings or parents who enjoy authority or younger siblings who lack it).
- `TopicBias`: the bias of the topic (`mascB/femB/neutB`), based on the categorization task.
- `GenderResp`: the self-declared gender of the respondent (`maleResp / femaleResp / otherResp`)

The experiment was run using an online questionnaire which contained 12 target items along with 24 fillers. Items and fillers were presented in a pseudo-random order with a latin-square design. 97 voluntary participants (64 `female`, 32 `male`, 1 `other`, mean age 27 years old) received a link to a questionnaire by e-mail or instant messaging. The questionnaire was hosted on the `IbexFarm` platform.

Results Overall the results confirm the results of the first experiment (Fig. 2).

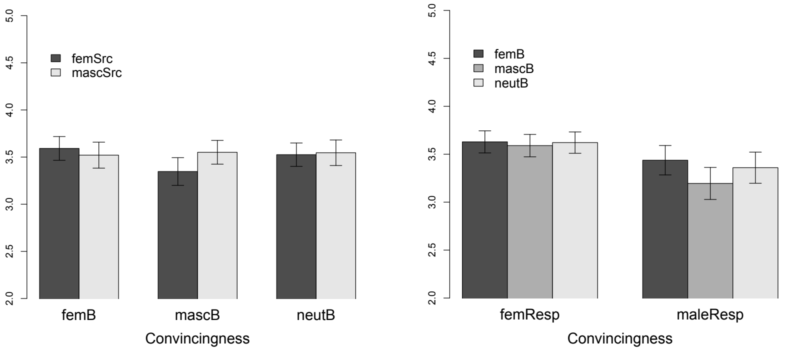

Fig. 2: Judgments of convincingness of the authority arguments: `TopicBias` vs. `Source` (left panel) and `GenderResp` vs. `TopicBias` (right panel).

There is no main effect of `Source`: masculine sources of information were overall not judged more reliable than female sources. There is a significant interaction between `Source` and `TopicBias` ($\chi^2 = 6.8$, $p = 0.048$): female sources were overall less trusted for masculine biased topics, but male sources were not less trusted for feminine biased topics.

There is a marginal effect of `GenderResp` ($\chi^2 = 5.30$, $p = 0.07$), i.e. male respondents tend to give lower scores than female participants. There is furthermore a *significant interaction* bw `GenderResp`, `Source` and `TopicBias` ($\chi^2 = 36.74$, $p = 6.27e-05$): male respondents are more skeptical about male sources in the case of male oriented topics.

Discussion The second experiment confirms the results of the first one: there is an interaction between the gender of the source of an information and the gender bias of the topic being discussed. However, there is an asymmetry between male and female sources. Male sources seem to enjoy an overall credibility, irrespective of the gender bias of the topic, whereas female sources are mostly credible on female biased topics. An effect of the gender of the respondent was also observed, with male respondents being more critical in some conditions.

To show how to handle these observations, let's consider an example like (5) under the Bayesian view on argumentation mentioned above.

(5) I have a friend who says he knows a lot about power tools, and he says this model is really powerful.

Two distinct pieces of information are given about the source in (5):

- the friend is male: $i \in T_{\text{male}}$
- the friend knows about power tools: $i \in K_{\text{powertools}}$

Let's assume that agents are given a default probability of being reliable sources of information about a domain D: $P(R_{i,D})$ (possibly high in some cases, if we follow the charity assumption postulated by [20]). When observing that i is of type T_{male} we have (via Bayes' rule):

(6) $P(R_{i,D}|i \in T_{\text{male}}) = \frac{P(i \in T_{\text{male}}|R_{i,D}) \times P(R_{i,D})}{P(i \in T_{\text{male}})}$

$P(i \in T_{\text{male}}|R_{i,D})$ is the likelihood of being of male if the agent is assumed to be reliable. This can be seen as a measure of the judge's (in our case the participant's) personal biases, which might be linked to the gender of the respondent, e.g. men have a tendency to distrust other men in general (maybe because they believe they are more competent).

The same account allows to factor in both pieces of information given about the source in (5). Equation (7) shows how to integrate two pieces of information to update a prior belief on the reliability of the speaker.

(7) $P(R_{i,D}|i \in K, i \in T) = \frac{P(i \in K|R_{i,D}, i \in T) \times P(i \in T|R_{i,D}) \times P(R_{i,D})}{P(i \in K, i \in T)}$

Equation (7) expresses the posterior probability that i is reliable in domain D, knowing that i is of type T (in (5): $T = T_{\text{male}}$) and has property A (in (5): $K = K_{\text{powertools}}$). If K is a property that is typical of type T, the quantity in (7) is very close to the one in (6), the limit case being that $T \subset K$, in which case $P(R_{i,D}|i \in K, i \in T) = P(R_{i,D}| \in T)$ (for instance, the assumption that all males are knowledgeable about power tools).

Using (7) we can see how to handle the various aspects highlighted by our experiments. The gender of the respondents affects the perceived likelihood that a reliable source is of a given gender. The quantity $P(i \in K|R_{i,D}, i \in T)$ handles the effect of the topic being discussed. For example one can assume that men are more trusted in general, no matter the topic, whereas the distribution of trust for female sources is less uniform. Another way to make use of this quantity is to reconsider the topics being discussed in the light of the exact type of gender bias involved for each topic. As the discussion above made clear, there are two situations:

1. A bias corresponding to universal competence on the part of one gender but partial competence elsewhere (e.g. all women know about cooking but only some men do)
2. A bias corresponding to nonuniversal competence on the part of one gender, but lack of competence on the part of the other gender (e.g. only some men know about power tools, but no women do)

For a topic of the first type, the information about the competence on the topic should not have further effect on the credibility of the speaker since the gender information entails it. For a topic of the second type, the two pieces of information give convergent evidence that the speaker is reliable. As of now, the categorization task we used does not allow us to distinguish between the two types of bias. In future work, we will rely on a more complex categorization of each topic which will provide that information (for example by asking participants to indicate their intuitions about the proportion of men and women who are knowledgeable about the topic). Another judgment task will then be used to check whether the topics with a bias of the second type are judged differently (e.g. more convincing) than the ones of the first type.

5 Conclusion

This paper has considered the nature of epistemic authority and two methods for acquiring and modifying it. The first passive method involves being generally perceived as reliable; for an analysis, reviewed the view of reliability of [13] — a combination of stereotype-based probability ascriptions and examination of communicative histories — and proposed it as one means of acquiring epistemic authority. The other method is more proactive: to manipulate stereotypes and other aspects of the context via the deployment of expressive content. We looked at one such instance in detail via experimental methods: the use of gendered pronouns to influence judgements about reliability, both in English and Cantonese. The results are intriguing, but still preliminary. We then proposed a Bayesian framework to account for the results.

Several directions suggest themselves for the future. The first, immediate steps involve additional experiments. We suggested above that not-at-issue content plays a different role in the manipulation of authority than at-issue content, because the efficacy of the latter already depends on the presence of epistemic authority. This difference remains to be experimentally verified, which we plan to do in the immediate future. Second, the experiments carried out so far involve subjective judgements and self-reporting tasks on the part of the experimental subjects. Avoiding potential biases, both implicit and explicit, which may be confounds for the experimental results is important and is a well-known problem in this area of research. There are several methods for addressing this problem, but the one we plan to pursue involves visual world experiments using eye-tracking; we intend to implement an experiment in this area in the near future. More generally, questions of the results on epistemic (and other) authority of the use of not-at-issue content are intriguing, especially for other sorts of expressive content such as honorification and particles; experimental approaches to these domains are also of interest, as is the examination of phenomena of other sorts such as presupposition and conversational implicature.

References

1. Baltag, A., Smets, S.: A qualitative theory of dynamic belief revision. In: Bonanno, G., van der Hoek, W., Wooldridge, M. (eds.) Logic and the Foundations of Game and Decision Theory, pp. 13–60. No. 3 in Texts in Logic and Games, Amsterdam University Press (2008)
2. Baltag, A., Smets, S.: Talking your way into agreement: Belief merge by persuasive communication. In: Baldoni, M., Baroglio, C., Bentahar, J., Boella, G., Cossentino, M., Dastani, M., Dunin-Keplicz, B., Fortino, G., Gleizes, M.P., Leite, J., Mascardi, V., Padget, J.A., Pavón, J., Polleres, A., Fallah-Seghrouchni, A.E., Torroni, P., Verbrugge, R. (eds.) MALLOW. CEUR Workshop Proceedings, vol. 494. CEUR-WS.org (2009)
3. Davis, C.: Decisions, dynamics and the Japanese particle *yo*. Journal of Semantics 26, 329–366 (2009)
4. Fricker, M.: Epistemic Injustice. Oxford University Press (2007)
5. Grice, H.: Logic and conversation. In: Cole, P., Morgan, J. (eds.) Syntax and Semantics III: Speech Acts, pp. 41–58. Academic Press, New York (1975)
6. Groenendijk, J., Stokhof, M.: Dynamic predicate logic. Linguistics and Philosophy 14, 39–100 (1991)
7. Hahn, U., Harris, A.J., Corner, A.: Argument content and argument source: An exploration. Informal Logic 29(4), 337–367 (2009)
8. Harada, S.: Honorifics. In: Shibatani, M. (ed.) Japanese generative grammar, pp. 499–561. Academic Press, New York (1976)
9. Kennedy, C.: Vagueness and gradability: The semantics of relative and absolute gradable predicates. Linguistics and Philosophy 30(1), 1–45 (2007)
10. Lewis, D.: Counterfactuals. Basil Blackwell, Oxford (1973)
11. McCready, E.: What man does. Linguistics and Philosophy 31, 671–724 (2008)
12. McCready, E.: A semantics for honorifics with reference to Thai. In: Aroonmanakun, W., Boonkwan, P., Supnithi, T. (eds.) Proceedings of PACLIC 28. pp. 513–521. Chulalongkorn University (2014)
13. McCready, E.: Reliability in Pragmatics. Oxford University Press (2015)
14. McCready, E.: The semantics and pragmatics of honorification (2015), manuscript, AGU
15. McCready, E.: Rational belief and evidence-based update. In: Hung, T.W., Lane, T.J. (eds.) Rationality: Constraints and Contexts, pp. 243–258. Elsevier (2016)
16. Murray, S.: Varieties of update. Semantics and Pragmatics 7(2), 1–53 (2015)
17. Northrup, O.: Grounds for Commitment. Ph.D. thesis, UCSC (2014)
18. Nowak, M., Sigmund, K.: The dynamics of indirect reciprocity. Journal of Theoretical Biology 194, 561–574 (1998)
19. Nowak, M., Sigmund, K.: Evolution of indirect reciprocity by image scoring. Nature 393, 573–577 (1998)

20. Oaksford, M., Hahn, U.: Why are we convinced by the ad hominem argument?: Bayesian source reliability and pragma-dialectical discussion rules. In: Zenker, F. (ed.) Bayesian Argumentation, pp. 39–58. Springer, NL (2013)
21. Potts, C.: The expressive dimension. Theoretical Linguistics 33, 165–198 (2007)
22. Potts, C., Kawahara, S.: Japanese honorifics as emotive definite descriptions. In: Proceedings of SALT XIV (2004)
23. Schlenker, P.: Maximize presupposition and Gricean reasoning. Natural Language Semantics 20, 391–429 (2012)
24. Stanley, J.: How Propaganda Works. Princeton University Press (2015)
25. Suzuki Kose, Y.: Japanese Sentence-Final Particles: A Pragmatic Principle Approach. Ph.D. thesis, University of Illinois at Urbana-Champaign (1997)
26. von Fintel, K.: Would you believe it? the King of France is back! In: Reimer, M., Bezuidenhout, A. (eds.) Descriptions and Beyond. Oxford (2004)
27. Walton, D.N., Reed, C., Macagno, F.: Argumentation Schemes. Cambridge University Press, Cambridge (2008)