Chapter 1

Introduction

Language is built upon a foundation of trust, and there is a case to be made that this trust is justified. Our modern lives depend almost entirely on testimonial knowledge (Mercier and Sperber, 2017; Levy, 2021), and language use is evolutionarily stable only because cooperativity is generally advantageous (McCready, 2015). However, listeners obviously do not always accept assertions automatically on the basis of conversational norms (Sperber et al, 2010; Farkas and Bruce, 2010), and speakers do not expect their addressees to take implausible testimony for granted (Oey et al, 2023). Even when Grice's (1975) Quality maxim is presumably in effect, interlocutors might have access to different bodies of evidence against which they evaluate facts, or else they may differ subjectively on matters of taste or opinion. This study concerns the use of reasons to negotiate the acceptance of controversial beliefs.

In argumentative scenarios, the function of a reason is to reduce or eliminate the need for trust among interlocutors by instead making the target content inferrable from the listener's accepted beliefs (Mercier, 2020). Beliefs justified by many coherent reasons, then, should be more credible– or at least, they should be no less credible– than beliefs justified by few or none. This analysis is intuitive, and it is predicted by a number of formal and probabilistic models of argument strength (e.g. Oaksford and Hahn, 2004, Godden and Zenker, 2018). The present study, however, will investigate the competing intuition that arguments are actually weakened when they are justified redundantly, that is, when (far) more justification is provided than what would be expected or required to raise an interlocutor's credence to the threshold for acceptance. Put famously by Queen Gertrude in Act III, Scene II of Hamlet: "The lady doth protest too much, methinks"

(see Figure 1.1).

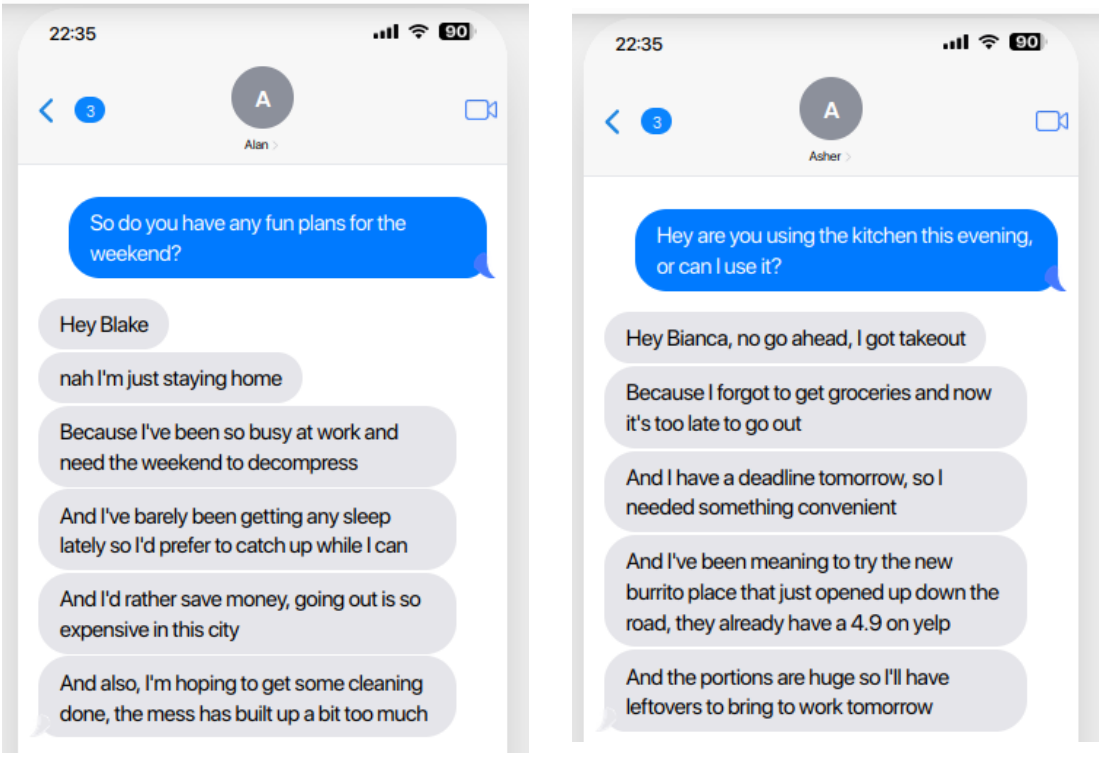


Figure 1.1: Examples of over-justification.

This effect, I propose, can be explained by the vast psycholinguistic literature on informativity. In this literature, information is defined as a measure of *surprise*, in keeping with computational theories of communication (Shannon, 1948). If speakers are to make efficient usage of their cognitive resources to maximize information exchange while avoiding misunderstandings, then longer signals should be reserved for meanings that are more surprising, or less probable. This analysis is borne out across all levels of linguistic structure, from phonemes to full utterances (Aylett and Turk, 2004; Levy and Jaeger, 2007; Mahowald et al, 2013; Asr and Demberg, 2020; Lemke et al, 2022). Likewise, comprehenders have been shown to interpret utterances with an expectation for informativity (Sedivy, 2003; Rohde and Rubio-Fernandez, 2022). They assume that efficient speakers do not mention easily inferrable content, and they derive pragmatic inferences to rationalize apparent inefficiencies. For example, assertion of default script continuations can induce atypicality inferences, e.g. *"She went to the grocery store,* ***and she paid the cashier!****"* generates an inference that the subject doesn't normally pay for her groceries (Kravtchenko and Demberg, 2022a; Ryzhova et al, 2023).

With this in mind, it seems that informal arguments might indeed come across as "defensive" rather than "persuasive" if more justification than contextually appropriate is nevertheless provided. A detailed explanation citing many reasons to make the conclusion more easily inferrable is only worth the production cost when the target proposition is particularly controversial, implausible, or otherwise difficult to integrate with the interlocutor's other beliefs. So if a rational listener were to recognize that the speaker anticipated disbelief or disagreement, he might even begin to doubt propositions he would have otherwise found innocuous. As a result, reasons that *extralinguistically* count in favor of the conclusion might *pragmatically* count against it.

The experiment reported in the subsequent chapters tests the predictions of these two competing hypotheses. In a behavioral study, participants read a series of text-message conversations, in which a character justifies an action with zero or four reasons. In half the conversations, the justifications are provided out of the blue, and in the other half, they are elicited discoursally to suppress the informativity inference. For each conversation, participants rated whether the speaker thought the listener would respond favorably or unfavorably, and whether the listener would actually respond favorably. The results of the study generally support the prediction of Queen Gertrude: participants inferred that the speakers had provided reasons out of the blue because they expected disagreement, and the arguments were ultimately weaker in the reasons-unexpected condition than in the reasons-expected condition when reasons were provided. These findings contribute new data that can help illuminate the relationship between trust and language processing.

Chapter 2

Background

2.1 Chapter roadmap

This chapter introduces two notions of rationality which, I propose, make different predictions with respect to the evaluation of explanations. The first, a psycholinguistic account (section 2.2), predicts that rational language users construct their utterances efficiently, and that comprehenders interpret utterances with that production preference in mind. As such, comprehenders are willing to adjust their background beliefs to make sense of utterances which might otherwise seem inefficient. The second account (section 2.3) draws from the psychology of reasoning, and predicts that rational agents should update their belief in a claim C just in accordance with the weight of evidence provided in support of C. Intuitively, the former account describes a pragmatic, interpretive process (recovering a speaker's meaning), and the latter an extralinguistic, evaluative process (deciding how to update one's beliefs in response to evidence).

I then review prior pragmatic accounts of explanation (section 2.4). Comprehenders are shown to anticipate an EXPLANATION continuation when the previous discourse segment is surprising or controversial; as such, a "good" or rational explanation should reduce the surprisal of the primary claim within the addressee's information state. While these accounts tend to deal with factual explanations, where the conclusion is already considered accepted by all discourse participants, explanations are also frequently enlisted for the purpose of *argumentation*, where acceptance depends on the strength of the inference available from the reason(s) to the conclusion.

Section 2.5 motivates the forthcoming experiment and lays out the contrasting predic-

tions. Briefly, if an addressee relies solely on extralinguistic reasoning to evaluate an argument (i.e. to adjust his credence in the main claim in response to the reasons), a more detailed explanation is always predicted to *increase* his credence. If the "protest too much" inference exists, this theory alone cannot explain it. Conversely, if pragmatic reasoning also affects evaluation, then providing an explanation– especially a lengthy one that involves multiple reasons– should generate an inference that the rational speaker anticipated her main claim would be surprising or controverisal in some way. Such an inference, under certain circumstances, might result in an overall *decrease* in the listener's credence.

2.2 Psycholinguistic rationality and communicative efficiency

2.2.1 A production preference for efficiency

"Information" in formal pragmatics is traditionally conceptualized as the set of possible worlds at which a proposition is true: it is a measure of logical strength. In certain computational and psycholinguistic traditions, however, information is a measure of *surprise*. Any linguistic unit (word, phoneme, etc) is informative when it is surprising; it is redundant when it is predictable (Levy and Jaeger, 2007). Following Shannon (1948), the information content ("surprisal") of a linguistic unit is typically quantified by the negative log probability of that unit conditioned on its context: $-\log P(u_i|u_0...u_{i-1})$. Shannon's influential model describes the optimal rate of information, or bitrate, through a noisy channel with limited bandwidth. In order to maximize the amount of information that can be transmitted without degradation, a constant bitrate that approaches, but does not exceed, the channel's capacity must be maintained. As such, the optimal coding scheme for a set of messages M distributed according to $P(M)$ assigns $-logP(m)$ bits to each $m \in M$. Or, less formally, longer codes are assigned to low-probability messages, and shorter codes to high-probability messages, in order to keep the bitrate fixed.

Natural languages are not optimal coding schemes, of course. However, it has long been of interest to psycholinguists whether language users might make *rational* usage of limited cognitive resources by constructing utterances efficiently (going back at least to Zipf, 1935). Over the past 20 years, numerous instantiations of the noisy-channel

model have successfully captured patterns of signal reduction in low-surprisal contexts across various levels of linguistic structure. Early work on speech processing identified articulatory and durational reductions on words and syllables in predictive contexts (Aylett and Turk, 2004; Bell et al, 2003; Pluymaekers et al, 2005). Levy and Jaeger (2007) generalized this pattern to all levels of linguistic representation as the Uniform Information Density hypothesis, which has since been applied successfully to a wide variety of phenomena such as contractions (Piantadosi et al, 2011), lexical reductions (e.g. chimp vs chimpanzee; Mahowald et al, 2013), omission of complementizers (Jaeger, 2010), use of sentence fragments (Bergen and Goodman, 2015; Lemke et al, 2021), and optionality of discourse connectives (Asr and Demberg, 2020).

Despite these findings, "rational" theories of language processing have often come under fire because of their apparent failure to account for the informational redundancies that abound in naturalistic communication (c.f. Degen et al, 2020). Most of these criticisms have focused on the phenomenon of definite reference, where speakers are willing to provide a logically stronger utterance than strictly necessary to uniquely identify the target. For example, a speaker might refer to "the blue square" in a tableau with just one square, or else might provide a descriptive NP when a pronoun or zero would still be unambiguous. More recent work, however, suggests that logically redundant references still conform to communicative efficiency pressures. Speakers do not "overmodify" at random. They reliably select adjectives that will be useful in establishing attention on the target (e.g. overmodification is common in high-entropy scenes, generally with adjectives that are visually salient or unique to the referent; Rubio-Fernandez, 2016), and they reliably avoid modifiers that are contextually inferrable (e.g. a yellow apple is "the yellow apple", but a yellow banana is just "the banana", since bananas are predictably yellow, and apples are not; Degen et al, 2020). While most of the experimental literature involves perceptual attention in the context of visual reference, Crone (2018) also argues that assertions of clarity (e.g. "It is clear that…"), which are sometimes logically redundant on Stalnaker's model of information update, are still asserted felicitously when establishing discoursal attention on grounded propositions which may not, at the present moment, be memorially or inferentially available.

2.2.2 A comprehension preference for informativity

If speakers have a tendency towards communicative efficiency, then, we should also expect comprehenders to interpret utterances with such a production preference in mind.

Frank and Goodman (2012) note that this is the core insight of Grice (1975)– comprehenders infer the state of the world based on both the semantic content of the utterance and the context in which it would be rational for the speaker to utter it.

A tension between the maxims of Quality and Quantity emerges in the experimental literature. In keeping with Quality, a number of studies have shown a comprehension bias towards utterances that are more plausible or typical, and thus, more likely to be true. For example, Bicknell et al (2010) observed reduced reading times for utterances expressing real-world typical material (e.g. *The **mechanic** fixed the **brakes*** and *The **journalist** fixed the **spelling***), and increased reading times for utterances expressing real-world atypical material (*The **mechanic** fixed the **spelling*** and *The **journalist** fixed the **brakes***).

However, in keeping with Quantity, numerous other studies have shown a comprehension preference for utterances that are informative or newsworthy. Pragmatic listeners expect content to be plausible, but they often do not expect the *most* locally predictable content, since pragmatic speakers prefer to avoid low-surprisal material (Rohde et al, 2021). For example, upon hearing the world "yellow", comprehenders expect the continuation to be an object that *can* be yellow (e.g. a shirt), but isn't *predictably* yellow (e.g. a banana), unless in the presence of a contrastive, atypical object (e.g. a brown banana), as evidenced by a number of sentence completion and eye-tracking studies (e.g. Sedivy, 2003; Rohde and Rubio-Fernandez, 2022). Hao et al (2025) extend this analysis from prototypical properties like color to stereotypical properties like gender: in their sentence-completion task, participants predicted traditionally masculine occupations in the continuation of *"Wang Hong is a **female**-"* far more often than in the continuation of *"Wang Hong is a-"*, since the adjective "female" would be uninformative if both the name and the occupation were female-stereotypical.

Based on these findings, several recent studies have investigated whether "informativity expectations" can lead to inferences about the contents of the common ground (Kravtchenko and Demberg, 2022a; 2022b; Ryzhova et al, 2023; Rees and Rohde, 2023). Redundancy must be measured against some prior information state, which is generally assumed to be the common ground. So, if the speaker makes an assertion which would be uninformative with respect to the listener's actual model of the *cg*, the listener might assume that the speaker's model of the *cg* is misaligned with his own. A number of studies have shown that listeners attempt to "accommodate" redundant utterances by adjusting their model of the *cg*, rather than simply assuming that the

speaker is behaving "irrationally". Kravtchenko and Demberg (2022a), for example, demonstrate that the assertion of script continuations, which would typically be available as default inferences, can give rise to "atypicality" or "habituality" inferences, e.g.*"She went to the grocery store, and* ***she paid the cashier!****"* generates an inference that this character doesn't normally pay for her groceries, as the listener would otherwise have assumed.

Kravtchenko and Demberg (2022b) quantify this inference using the Rational Speech Act (RSA; Frank and Goodman, 2012) framework, specifically adapting the "wonky world" model of Degen et al (2015) to capture this revision of background knowledge. Their habituality RSA model specifies that a pragmatic speaker should select the utterance that maximizes the probability of a "literal listener" inferring the correct state of the world (e.g. the one where the character pays for her groceries), just from the literal semantics of her utterance and his prior knowledge of the activity's habituality. However, she also incurs a cost for the effort of her assertion, so she should only inform the listener of the character's behavior if she thinks it would be unlikely for him to infer it otherwise. The pragmatic listener, then, jointly infers the world state and the habituality of the action. Considering the fact that the speaker did find it worthwhile to mention that the character ***paid the cashier*** at the grocery store, she must have predicted that a literal listener's prior would not be high enough to assume this behavior by default. As such, the pragmatic listener learns that the character does not usually pay for her groceries (i.e. the "wonky" prior), in addition to the fact that she did on this specific occasion (the posterior).

The literature has (thus far) been entirely concerned with *interpretation*: how listeners recover speakers' beliefs based on the utterances they produce. Kravtchenko and Demberg do not consider whether the habituality inference they observed might also reduce the listener's posterior belief in the asserted content. It is (understandably) assumed that the listener never infers a world state that is inconsistent with the speaker's intended meaning.

The present study investigates whether informativity expectations can also affect *evaluation*. There is some precedent to this idea: Sperber et al (2010) suggest that their theory of *epistemic vigilance*, a set of cognitive mechanisms with which we assess the trustworthiness of testimony, is viable because the processing cost of plausibility-estimation is built into the processing cost of interpretation. However, they somewhat unsurprisingly propose that the key mechanism is the establishment of *relevance*, in

the sense of Sperber and Wilson (1987). Further, the "vigilance to the content" they describe ultimately only involves the plausibility of the (possibly pragmatically enriched) content itself, not the plausibility of a rational speaker choosing to assert it at all.

While it is entirely sensible to assume that the Quality maxim holds in cases like Kravtchenko and Demberg's, explanations are unique in that they are frequently recruited to raise the probability of a proposition *which might otherwise be difficult to believe* (i.e. in persuasive and argumentative contexts). In fact, as noted by Sperber et al, Grice (1969) himself concedes that argumentation presents a challenge for his cooperative principle, since an argument is meant to be accepted on the basis of soundness, not (just) by recognition of communicative intent. Since accepting the conclusion of a persuasive explanation is expected to involve at least some extralinguistic reasoning beyond recovery of the speaker's intended meaning, an informativity inference that *reduces* the presumed prior probability of the conclusion could potentially *counteract* the acceptance of the argument (this point will be elaborated in further detail in section 2.5).

2.3 Extralinguistic rationality and argument strength

Although argumentation is a somewhat marginal case in the study of (cooperative) pragmatics, it is a central concern for philosophers and psychologists. This section will introduce an intuitive psychological theory of argument strength which has received modest attention from linguists studying argumentation (e.g. Cummins and Franke, 2021; Merin, 1999; Winterstein, 2012; 2018), and which, I will demonstrate, does not predict the "protest too much" inference.

It is well-attested in the psychological literature that addressees are more inclined to accept propositions that increase the overall coherence of their beliefs. As such, the core function of argumentation is to increase the acceptability of a controversial proposition by making it more coherent with, and therefore inferrable from, other propositions that the interlocutor will admit more easily (Mercier, 2020). A reason is said to "count in favor" of a claim whose acceptance it enables in this way.

The "counts in favor" or "argues for" relation has historically been treated categorically, typically in the study of formal/syllogistic arguments dating back at least to Aristotle. An argument can only be sound if the conclusion is a logical consequence of the

reasons; arguments that fail to meet this standard are considered fallacious even if they are psychologically persuasive (see Mercier and Sperber, 2017 for an overview). Taking a more gradient tack, however, an argument can also be characterized in terms of its effectiveness in bringing about the desired belief (Oaksford and Hahn, 2004; Hahn and Oaksford, 2007). The latter approach lends itself to probabilistic, and specifically Bayesian models. In brief, a set of reasons argues in favor of the conclusion if the posterior probability of the conclusion conditioned on the reasons $P(C|R)$ is greater than the (pre-utterance) prior probability of the conclusion $P(C)$. The strength of a reason in an informal argument is quantified by the extent to which it raises the addressee's degree of belief (credence) or commitment towards the conclusion, generally as computed through application of Bayes' Theorem (e.g. Godden and Zenker, 2018).

The Bayesian analysis has been applied to a wide variety of classic fallacies, accounting for the intuitive difference in "validity" between different instances of the same type: for example, *Ghosts exist because no one has proved they do not* and *This drug is safe because we have no evidence that it is not* are both cases of "argument from ignorance", but the latter is obviously more compelling (Hahn and Oaksford, 2007). However, although the Bayesian approach breaks from formal logic here, it should be noted that a number of its adherents still ultimately pursue a normative measure of argument strength. This formulation is intended to describe how a "rational" listener *should* update his beliefs; on this analysis, divergent behavior can only be explained because humans do not always behave rationally.

Although argumentation has received surprisingly little attention from linguists, some efforts have been made (e.g. Merin, 1999; Winterstein, 2012; 2018) to import the probabilistic/Bayesian approach into one classic linguistic theory of argument, Argumentation within Language (Anscombre and Ducrot, 1983). The key motivation for this account is that argument strength depends on linguistic form in addition to propositional content. For example, *Alice* ***almost*** *arrived on time, don't scold her!* is a legible argument, while *#Alice* ***barely*** *arrived on time, don't scold her!* is not, even though the latter proposition entails *Alice arrived on time* and seems to be stronger evidence against scolding her. Winterstein (2012; 2018) develops a Bayesian semantics for a number of discourse connectives and particles to account for their particular argumentative orientation.

Along similar lines, it has been observed that pragmatic enrichment, due for example to approximate numeral interpretation (Cummins and Franke, 2021) or quantifier

vagueness (Macuch Silva et al, 2024) can also affect the extent to which a reason counts in favor of a conclusion. To illustrate: *University X has a **top 19** linguistics program* is a semantically stronger, but pragmatically weaker, endorsement of the department at University X than *University X has a **top 20** linguistics program* (adapted from Cummins and Franke, 2021). This is because 19 is typically interpreted punctually, whereas 20 is typically interpreted vaguely, and its pragmatic halo (Lasersohn, 1999) often includes numbers higher than 19.

Each of these linguistic accounts improves upon the purely propositional analysis, but they do not question the tacit assumption that a "better" argument licenses a stronger inference from the reasons to the conclusion. Indeed, a number of them also make reference to a type of reasoning over alternatives reminiscent of Grice, where a listener might reject a favorable argument if he knows of an obviously stronger one the speaker could have made, but didn't (assuming she would have, if it were true; see also the "weak evidence effect", Barnett et al, 2022; and also Davis et al, 2007; McCready, 2015 for related approaches to hearsay/reportative evidentials).

As such, it seems to be the case that many good reasons (for some measure of goodness, and as long as the reasons are all acceptable and consistent with one another), should always count more strongly in favor of the conclusion. The listener's degree of belief in the conclusion should increase monotonically approaching certainty (i.e. $P(C|R) = 1$) as more positive reasons are provided (c.f. Godden and Zenker, 2018 for a formal treatment). While first-order Bayesian conditionalization does predict "diminishing returns" as the addressee's credence gets closer to 1 (Winterstein, 2018), it has nothing to say about Queen Gertrude's observation, that too many reasons might actually count against the conclusion.

Now, I want to question the assumption that, if the speaker provides a more complete justification for any given conclusion, the listener can only be better off epistemically. When the conclusion C is particularly challenging or implausible to the listener, providing reasons becomes necessary to make C available by inference (Mercier, 2020). But in day-to-day conversation, it seems to me that the typical number of reasons provided to support any given conclusion is zero, since most assertions are admissible due in large part to conversational norms (i.e. Grice, 1975).

There is, of course, considerable debate among epistemologists as to whether this tendency is justified or simply gullible (see Levy, 2021 for a recent contribution) if it

exists at all (Sperber et al, 2010; Mercier, 2020), but I don't intend to make a case for the epistemic status of testimony here. I just want to propose that the "protesting too much" inference, if it should arise, can still be explained in terms of rational language processing; it need not be a simple failure of extralinguistic reasoning. In fact, as we have already seen, rational language users aim to maximize the *rate* of information transfer between interlocutors. Language would be a highly *inefficient* means of acquiring new information if every proposition learned by testimony was fully, or even mostly, deducible from the listener's prior epistemic state, although testimony would undeniably be more reliable if this were so. A pragmatic speaker might then be expected to reserve explanations– an in particular, detailed or effortful explanations– for propositions that are otherwise surprising, controversial, or less coherent with (the speaker's model of) the listener's information state. In fact, many previous pragmatic theories of explanation, in both the philosophical and psycholinguistic traditions, have made this precise prediction.

2.4 Pragmatic explanations

Before specifying the predictions that each concept of rationality makes with respect to *redundant* explanations, I will briefly address what makes an explanation *informative, useful, and/or anticipated* in a discourse. On a "relational" account of discourse coherence (e.g. Asher and Lascarides, 2003; Hobbs, 1979; Kehler, 2002), an EXPLANATION is one of the many coherence relations that can hold between two utterances. Although various proposals differ in their exact details, it is generally accepted that EXPLANATION holds between utterances U_1 and U_2 (expressing propositions p and q) when p "follows from" q, either causally or evidentially. Relational accounts have typically been concerned with specifying the logical rules and inferential processes by which coherence is established (c.f. Kehler, 2022 for an overview).

However, more recent experimental approaches to online discourse processing have shown that comprehenders make predictions about which coherence relations are likely to follow a given utterance, rather than simply recovering the coherence relation that is most likely to hold between two full utterances (Kehler et al, 2008). Much of this literature has focused the on anticipation of EXPLANATIONs in particular. Unsurprisingly, explanations are anticipated when the preceding discourse is unusual or statistically unlikely in some way. For example, Rohde et al (2007) show that comprehenders

anticipate more explanations in transfer-of-possession contexts that feature unusual instruments (*"John gave a **bloody meat cleaver** to Bob"*, whereas OCCASION and ELABORATION are both more likely following normal instruments (*"John gave a **book** to Bob*").

Perhaps the most well-established class of items that anticipate an EXPLANATION continuation, however, are the implicit causality (IC) verbs (e.g. Garvey and Caramazza, 1974; McKoon et al 1993, i.a.). These items originally generated interest because of the apparent asymmetry in their coreference or "re-mention" bias in explanatory continuations: comprehenders tend to interpret the pronoun in *"Sally **frightened** Mary because **she**..."* to refer to Sally, whereas they interpret the pronoun in *"Sally **loves** Mary because **she**..."* to refer to Mary (Hartshorne, 2014). More recent work (Kehler and Rohde, 2017; Solstad and Bott, 2021) aims to determine whether– and if so, why– these items have a *coherence* bias for explanation continuations (which, in turn, might account for the coreference bias). The coherence bias is borne out in a number of corpus (Asr and Demberg, 2020) as well as experimental (Kehler and Rohde, 2017; Solstad and Bott, 2021) studies. Although the source of the IC bias is somewhat less clear, it only appears for specific classes of items, such as psychological verbs (*fear*), judgment verbs (*criticize*), and stimulus-experiencer verbs (*annoy*; Solstad and Bott, 2021). Although these items do not describe actions that are particularly *statistically* unlikely, it should be noted that they do involve *normative* uncertainty. The explanation clarifies whether or not the action or psychological state is justified or good, which would not be contextually inferrable.

From a philosophical perspective, pragmatic considerations have also played a role in characterizing "explanatory virtue", i.e. what makes a (scientific/factual) explanation good or useful to an addressee. This line of thinking is due in large part to Gärdenfors (1980), who proposes that explanations are typically requested and provided in response to claims that are particularly surprising or unpredictable with respect to the addressee's epistemic state. A good explanation, then, must make the claim less surprising within that epistemic state (absent the knowledge of the claim itself). Chandra et al (2025) provide experimental evidence that explainers do predict what information receivers might be missing when responding to *why* questions and tend to offer the minimal explanation that will resolve the receiver's surprise.

While Gärdenfors' account is strictly epistemic, his insight is preserved in causal accounts of factual explanation as well (famously, Halpern and Pearl, 2005). Halpern

and Pearl also propose a parsimony constraint, that a causal explanation should be the *minimal* set of propositions that constitute an "actual cause" within the addressee's epistemic state. Harding et al (2025) have recently proposed an RSA-based theory of causal explanation, in which the preference for minimal explanations that resolve the listener's uncertainty can be attributed to pragmatic speakers maximizing the utility of the utterance for a literal listener while also minimizing the cost of the utterance for themselves.

These accounts generally deal with factual rather than argumentative explanations; the listener is assumed to accept the conclusion with or without the explanation. However, these accounts still seem to corroborate the prediction outlined in section 2.3: a pragmatic speaker prefers to provide an explanation to raise the probability of her main conclusion when she expects the listener might otherwise assign that conclusion a low probability. Any additional probability-raising is unnecessary and therefore inefficient.

What remains to be seen, then, is if listeners interpret and evaluate excessive or redundant explanations with this production preference in mind.

2.5 The present study

The objective of the present study is to determine whether communicative efficiency considerations affect argument strength. I will constrain the space of "arguments" to first-person explanations for actions, in the format of "I did/will do A, because of W, X, Y, Z". This is an intuitive candidate for realistic over-explanation cases, since people do often want to make sure they are perceived as having done "the right thing" (see 3.3 for further detail on developing naturalistic stimuli). Furthermore, the "goodness" of an action is ultimately a matter of opinion, so it should hopefully be evident to the experiment participants that the speaker and listener can reasonably disagree on the target proposition, whether or not the explanation is provided.

The experiment will measure whether excessively detailed reasoning for an action can negatively influence the interlocutor's opinion of the action. It will also measure whether this effect is moderated by how expected the explanation is within the discourse.

2.5.1 RQ1: The pragmatic prior

The first research question of this study asks whether a pragmatic listener respond to over-explanation by making an inference about the speaker's model of his beliefs.

Listeners, at first, should have their own independent assessment of the "favorability" or acceptability of the action; I will call this the naive prior. The naive prior should be agnostic to any justification provided by the speaker. However, listeners might also infer how (un)favorable the action would have to be in order for a pragmatic speaker to produce the utterance she produced; call this the pragmatic prior. This study aims to determine whether the listener is more likely to assume that the speaker anticipated an *unfavorable* response when she provides many reasons, rather than none. As such, the pragmatic prior should be lower when reasons are provided, and higher when they are not.

Such an inference would be in keeping with the previously mentioned studies (Degen et al, 2015; Kravtchenko and Demberg, 2022) which show that comprehenders will interpret against "wonky" priors to make sense of utterances which would not be "rational" to assert against their naive priors, as well as with the prediction that giving more reasons is only efficient when the listener is less likely to accept the claim otherwise.

2.5.2 RQ2: Posterior belief

The second research question addresses the issue unique to argumentation: whether the reduction of the prior, as proposed in RQ1, can counteract the extralinguistic weight of the reasons and affect the evaluation of the claim.

In other words, if the listener realizes that the speaker *expected* the action to be received unfavorably, and he tries to rationalize why she might have thought that, then the explanation itself might not be sufficient to recover from that inference, even if all of the individual reasons are strong. For example, imagine your friend is explaining at length why she needs to go to the library (she's got an exam tomorrow, her book is overdue, there's construction going on outside her apartment...), to choose a random innocuous activity. At a certain point, you might begin to wonder: *well... why did she think I'd have a problem with her going to the library? Maybe she's avoiding another obligation...*

2.5.3 Moderation by expectedness of reasons

Further, this study investigates whether such a change in opinion is moderated by how "over-informative" the listener actually perceives the explanation to be. Naturally, it cannot be the case that giving many reasons to justify a conclusion never increases the hearer's credence–if that were true, this paper has already gone on too long. As we have seen from section 2.4, there are contexts where explanations are more or less expected. In a context where a rational speaker would be expected to provide *many* reasons, the listener would not need to make any inference to rationalize her utterance. As such, this experiment also makes use of a "(many) reasons expected" vs "(many) reasons unexpected" manipulation, see 3.3 for further detail.

2.5.4 Hypotheses

1. The **naive/literal hypothesis** ("More is more!") predicts that comprehenders will not reason about the speaker's motivation for justifying her action. As such, the number of reasons provided should have no effect on the prior, whether or not the reasons are expected. Then, the listener's posterior belief should always be *more* favorable when reasons are provided, since more evidence is available to corroborate the conclusion. This result is predicted by the extralinguistic/first-order Bayesian analysis of argument strength, as specified in section 2.3.

2. The **pragmatic hypothesis** ("Less is more!") predicts that comprehenders *will* reason about the speaker's motivation for justifying her action (i.e. she predicted a disagreement). The volume of reasons provided should yield a reduction in the prior when reasons are provided unexpectedly, and should have no effect when the reasons are expected. As such, the listener's posterior belief should be *less* favorable when reasons are unexpectedly provided, and *more* favorable when reasons are expectedly provided, in comparison to the no-reasons cases.

Chapter 3

Methods

3.1 Participants

60 participants were recruited through Prolific. All participants self-identified as native speakers of English, and were residents of the United Kingdom or United States. An additional 6 participants were recruited for a shorter pilot study (with the free-text question omitted); their data is not included in the subsequent analysis. The survey took a median time of 17 minutes, 52 seconds to complete, and participants were compensated £5.00. All participants in the main study passed the attention checks, so they were all included in the analysis.

3.2 Study Design

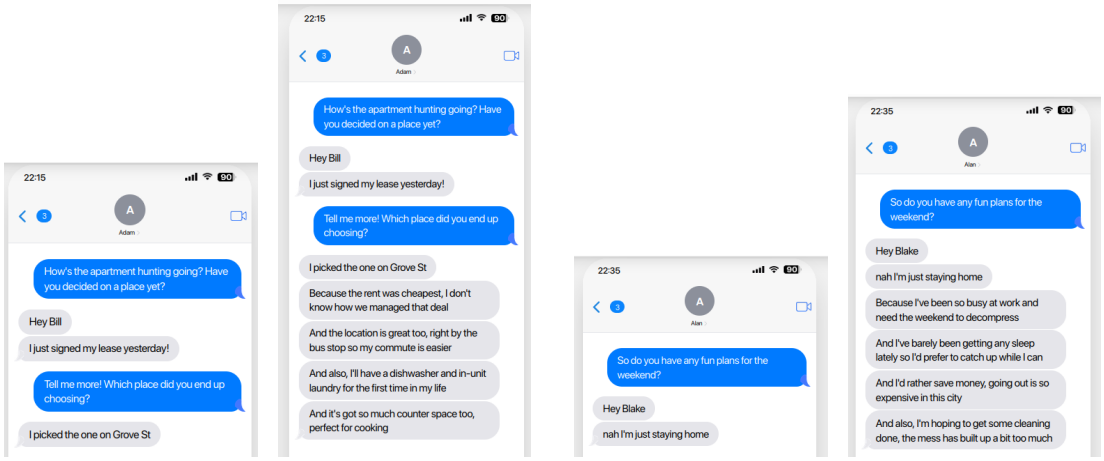
The experiment was constructed using a 2x2 Latin square design, crossing expectation (**reasonsUnexpected** vs **reasonsExpected**) with reasons (**noReasonsGiven** vs **reasonsGiven**. Items were sorted into two lists, with 4 trials in each of the 4 conditions. Each item appeared in only one condition per list. Participants also saw 4 attention checks, for a total of 20 trials.

3.3 Materials

Prior work has shown that comprehenders have a greater expectation for informativity when the utterance is produced by a salient speaker (for example, a named character), so redundancies are more easily perceived as deliberate (Reksnes et al, 2024). As

such, the stimuli for this study were presented as text message conversations between two named characters, adapting the formatting of Wallbridge et al (2021). To avoid confusion, the explainer's name always began with A, and the addressee's name always began with B.

Since the research questions target comprehenders' inferences about speakers' production choices (rather than speakers' production choices themselves), participants were made to assume the role of the addressee rather than the explainer. Therefore, the explanation always came from the "incoming"/grey bubble character, and both rating scale questions concerned the opinion of the "outgoing"/blue bubble character.



Sample **reasonsExpected** stimuli

Sample **reasonsUnexpected** stimuli

Figure 3.1: Sample stimuli in **noReasonsGiven** and **reasonsGiven** conditions.

For the **reasonsUnexpected** class of items, character A describes an "ordinary" altruistic or innocuous behavior (e.g. cleaning the kitchen for a roommate or studying in the library), which typically would not involve explicit justification. The reasons (when provided) were offered unprompted.

For the **reasonsExpected** class of items, character A describes a behavior that typically involves a deliberative process where multiple reasons would be considered (e.g. moving to a new city or making an expensive purchase), in response to a "which one"/"tell me more"-type question from character B. This was meant to create a context where extensive elaboration would be appropriate both in terms of subject-matter and surrounding discourse.

In the **noReasonsGiven** condition, character A simply mentions the behavior.

In the **reasonsGiven** condition, character A provides four reasons. The content of the

explanation was not held constant across item types, but was instead tailored to provide strong/sufficient justification for each individual behavior (to avoid confounding with the "weak evidence" effect e.g. Barnett et al, 2022; as well as to maintain the naturalness of the dialogues). All explanations were approximately the same total length (although some variability was introduced, again to maintain naturalism).

Realism of the stimuli was prioritized so that the participants might have actual intuitions about why a justification would or would not be provided in a given circumstance, following prior work on pragmatic inferences generated by strategic language. For example, while a number of studies have claimed that participants nearly-uniformly do not make lie judgments on implicit meaning, Weigmann (2023) shows that such judgments are simply suppressed in odd situations where the characters' motives for deception are unintuitive. This prediction was borne out in the responses to the free-text question, see section 5.4.

In all target trials, the messages cut off before character B responds. An additional 4 attention checks were constructed, where character B does react to character A's message with explicit approval or disapproval (2 items for each), so the participant could simply report B's perspective without inference.

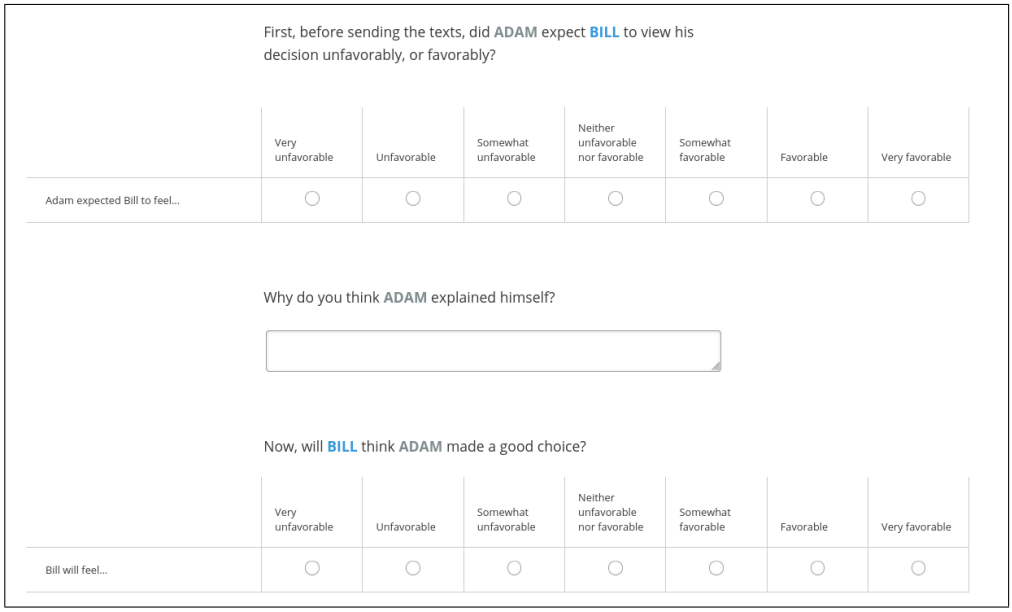


Figure 3.2: Sample questions in the **reasonsGiven** condition.

Each target trial contained two or three questions. First, participants were asked to rate the following on a 7-point Likert scale from "strongly unfavorable" to "strongly favorable":

- First, before sending the texts, did A expect B to view her decision unfavorably, or favorably? A expected B to feel...
- Now, will B think A made a good choice? B will feel...

The first rating was intended to target RQ1: whether the listener would think the speaker had explained her decision because she expected him to disagree with it. The second rating was intended to target RQ2: whether the listener's recognition of that expectation could counteract the effectiveness of the reasons in influencing his beliefs. In other words, the first question identifies whether the "pragmatic prior" is lower (less favorable) in the **reasonsGiven** condition than in the **noReasonsGiven** condition, and whether that effect is moderated by expectedness of the explanation. The second question identifies whether the lower prior produced by the explanation would then yield a reduction in *post-utterance belief* (again, a reduction in favorability rating from the **noReasonsGiven** condition to the **reasonsGiven** condition– and again, whether that effect is moderated by expectedness of the explanation).

Participants were also asked to provide a free-text response to the question "Why do you think A explained herself?" for items in the **reasonsGiven** condition, to see if participants preferred an alternative rationalization for the over-informativity besides the hypothesized judgments measured by the Likert ratings. This additional question was included in the main experiment to provide more interpretable data for analysis, after conducting a preliminary pilot study which only included the Likert ratings.

3.4 Procedure

The experiment was hosted online on Qualtrics. Before proceeding to the survey, participants were asked to read an information sheet and consent to participation (approved by the University of Edinburgh LEL Research Ethics Board) and to confirm their Prolific ID and native-speaker status. The participants were then instructed to read each set of messages closely and reason about the characters' motives for sending them. Using the Qualtrics survey randomizer, each participant was assigned to one of the two stimulus lists, and saw the items in random order. Each trial was presented on a separate page, and participants were required to answer each of the questions before clicking to advance. At the end of the survey, participants completed a "debrief", where they were asked to provide feedback and guess what the experiment was about.

Chapter 4

Results

4.1 Model selection

Data analysis for the Likert rating questions was conducted using a Bayesian cumulative logit model. Although linear mixed-effects models are standard practice in psychology for 7-point Likert scales (assumption of equidistant category boundaries is more acceptable with high-granularity ordinal scales), the Bayesian cumulative model is still best practice for ordinal data since it makes no assumptions regarding the spacing of the categories (Bürkner and Vuorre, 2019). The model was implemented using the R (R Core Team, 2025) package `brms` (Bürkner, 2017) with default priors. The fixed effects included REASONS (**noReasonsGiven** vs **reasonsGiven**), EXPECTATION (**reasonsUnexpected** vs **reasonsExpected**), and their interaction. To account for participant and item-level variability, random effects for PARTICIPANT and ITEM were also included.

4.2 Do listeners rationalize over-explanation?

The first Likert rating question (*Before sending the texts, did A expect B to view her decision unfavorably, or favorably?*) measures the pragmatic prior. It was intended to target RQ1: whether listeners would attempt to rationalize a speaker's decision to (over)explain her actions, by inferring that she predicted a disagreement. Responses are plotted in Figure 5.1.

Treating **noReasonsGiven** and **reasonsUnexpected** as the baseline conditions, the

model identified a credible main effect of REASONS (β = -0.77, 95% CI -1.09, -0.43), indicating a reduction in favorability rating when reasons were given. This effect was moderated by a credible interaction of REASONS with EXPECTATION (β = 0.83, 95% CI 0.35, 1.30), which reversed the main effect entirely when the reasons were expected. In other words, participants rated the speaker's estimation of the listener's opinion credibly lower when reasons were *unexpectedly* provided. Participants did not rate the speaker's estimation of the listener's opinion credibly lower when the reasons were *expectedly* provided. There was also no credible main effect of EXPECTATION, indicating that **reasonsExpected** items were not consistently rated higher or lower in favorability than **reasonsUnexpected** items in the **noReasonsGiven** condition. Results of the models are summarized in Table 5.1.

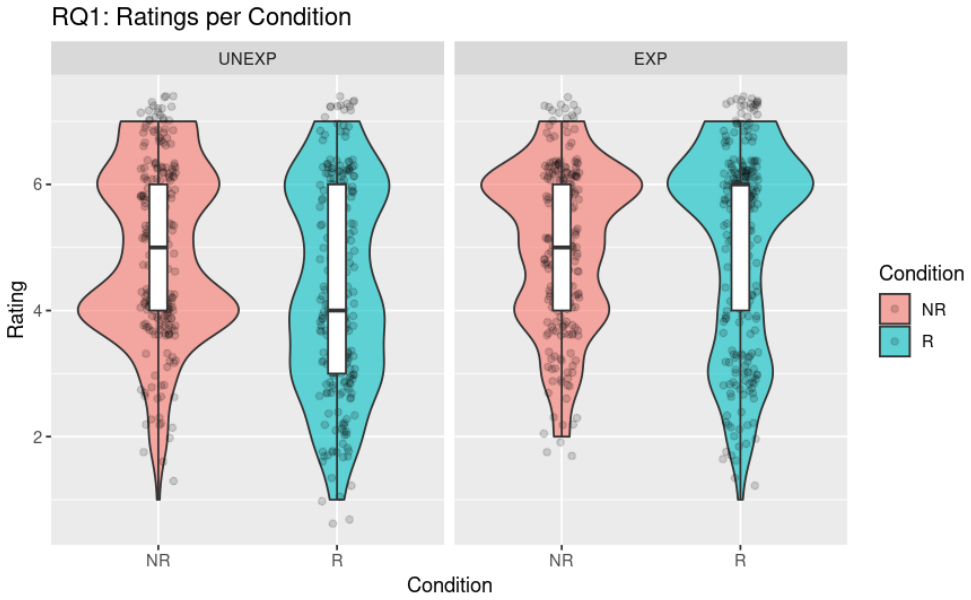


Figure 4.1: Responses to RQ1. Prior favorability ratings reduced when reasons were provided unexpectedly.

These results are in accordance with the predictions of the pragmatic hypothesis: when many reasons are unexpectedly provided, the listener rationalizes the speaker's decision to provide them. The listener infers that the speaker expected him to disapprove of her action– that he would need to be *convinced* to agree with her. The result is a reduced "pragmatic prior" favorability rating. However, in a context where many reasons are appropriate, no such inference is necessary. The literal hypothesis would instead predict no effect for either item type (expected or unexpected), since the listener does not reason about the speaker's choice to explain herself in any case.

Predictor	Estimate	Std. Error	95% CI Lower	95% CI Upper
conditionR	**-0.77**	**0.17**	**-1.09**	**-0.43**
typeEXP	0.19	0.66	-1.12	1.52
conditionR:typeEXP	**0.83**	**0.24**	**0.35**	**1.30**

Table 4.1: Log-odds coefficients from the Bayesian ordinal model of RQ1. A credible main effect of REASONS, moderated by a credible interaction with EXPECTATION, is observed.

4.3 Does over-explanation affect argument strength?

The second Likert scale question (*Now, will B think A made a good choice?*) measures the listener's posterior favorability rating. It was intended to target RQ2: whether higher-order inferences about *why the speaker felt an explanation was necessary* would mitigate that explanation's effectiveness in persuading the listener. Responses are plotted in Figure 5.2.

Once again treating **noReasonsGiven**, **reasonsUnexpected** as the baseline conditions, there was no credible main effect of REASONS for **reasonsUnexpected** items. However, the model identified a credible interaction of REASONS with EXPECTATION (β = 0.75, 95% CI 0.28, 1.22), meaning that the posterior favorability ratings improved when the reasons were *expected*. Although unexpected reasons had no effect on the listener's post-utterance favorability ratings, expected reasons significantly improved the listener's rating. There was still no credible main effect of EXPECTATION, indicating that **reasonsExpected** items were not consistently rated higher or lower than **reasonsUnexpected** items in the **noReasonsGiven** condition. Results of the models are summarized in Table 5.2.

Predictor	Estimate	Std. Error	95% CI Lower	95% CI Upper
conditionR	0.15	0.17	-0.20	0.50
typeEXP	0.25	0.76	-1.25	1.79
conditionR:typeEXP	**0.75**	**0.24**	**0.28**	**1.22**

Table 4.2: Log-odds coefficients from the Bayesian ordinal model of RQ2. A credible interaction of REASONS and EXPECTATION is observed.

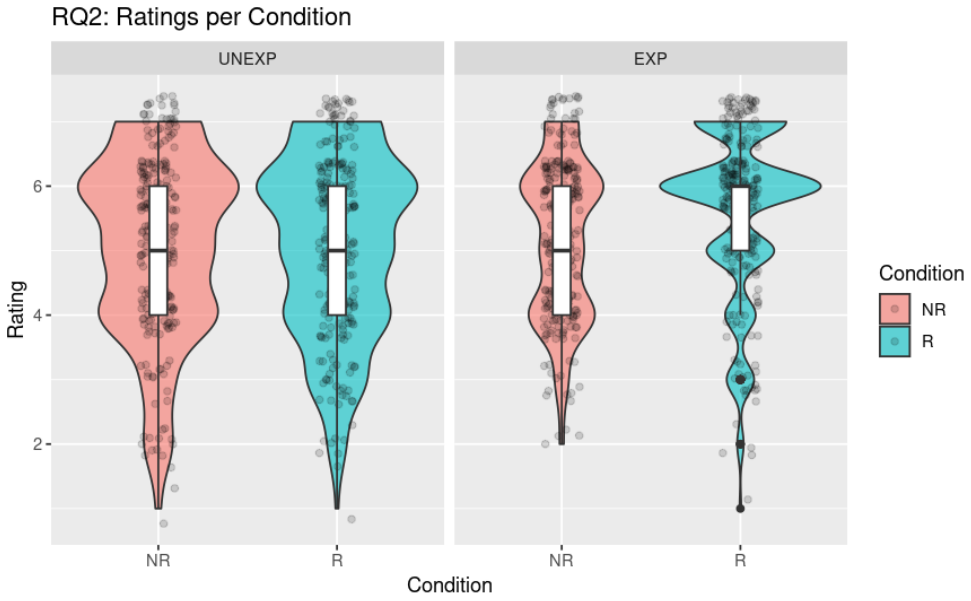


Figure 4.2: Responses to RQ2. Posterior favorability ratings increased with reasons only when the reasons were expected.

There was no credible *reduction* in listeners' posterior favorability rating in the **reasonsGiven** condition, suggesting that the higher-order inference about why the speaker wanted to explain herself was insufficient to fully counteract the argumentative strength of the reasons. However, these results still support a pragmatic analysis. It should be noted that there was no credible effect of EXPECTATION for items in the **noReasonsGiven** condition: that is, there was no significant difference between baseline approval ratings for the **reasonsExpected** and **reasonsUnexpected** items. However, the explanations improved post-utterance listener favorability ratings only for **reasonsExpected** items; they had no credible effect for **reasonsUnexpected** items. This suggests that, although there was similar "room for improvement" in both conditions (i.e. it was not the case that listeners *already* maximally approved of the **reasonsUnexpected** items), listeners were only persuaded in the **reasonsExpected** condition. If this is indeed the case, the argumentative strength of the reasons could have been dampened by the reduction of the pragmatic prior identified in Q1, as predicted by the pragmatic hypothesis.

To corroborate this analysis, a new model was constructed to examine the difference between participants' responses to the two questions (effectively Q2-Q1, but maintaining the ordinality of the data); that is, to measure the difference between the pragmatic prior (Q1 rating) and the listener's post-utterance belief (Q2 rating). The model in-

Predictor	Estimate	Std. Error	95% CI Lower	95% CI Upper
condR	-0.77	0.17	-1.11	-0.44
Q2	0.19	0.17	-0.15	0.52
typeEXP	0.24	0.68	-1.10	1.60
condR:Q2	**0.95**	**0.24**	**0.48**	**1.44**
condR:typeEXP	0.86	0.24	0.40	1.32
Q2:typeEXP	-0.03	0.24	-0.51	0.44
condR:Q2:typeEXP	-0.17	0.34	-0.87	0.49

Table 4.3: Log-odds coefficients from the three-way model. A credible interaction of QUESTION and REASONS is observed.

cluded an additional fixed effect of QUESTION number (and its interactions), with all other parameters maintained from the previous models. The new model found no credible main effect of QUESTION, meaning that the pragmatic prior (Q1 rating) and posterior (Q2 rating) were about the same when no reasons were provided. There was a credible interaction of QUESTION and REASONS: the posterior (Q2) was credibly higher than the pragmatic prior (Q1) when reasons were provided. Finally, there was no credible three-way interaction between QUESTION, REASONS, and EXPECTATION, meaning that the increase in favorability from the pragmatic prior to the listener's posterior belief was similar for both **reasonsUnexpected** and **reasonsExpected** items when reasons were provided.

Summing up, then: the Q1 model found that the that the pragmatic prior was reduced in the **reasonsGiven** condition for **reasonsUnexpected** items, but it was not reduced for the **reasonsExpected** items. The three-way model found that the *difference* between the pragmatic prior and the posterior favorability was equivalent for both **reasonsUnexpected** and **reasonsExpected** items. Therefore, items in the **reasonsExpected** set saw a net improvement from the **noReasonsGiven** condition to the **reasonsGiven** condition, whereas the items in the **reasonsUnexpected** set simply recovered from the reduced pragmatic prior and returned to "baseline", as per the findings of the Q2 model.

4.4 Notes on the free-text data

Although full quantitative analysis of the free-text responses is beyond the scope of this paper[1], some high-level analysis of the results that emerged are discussed in the following chapter.

4.5 Notes on the debrief

At the end of the survey, participants were asked to report what they thought experiment was investigating. The vast majority of respondents did not mention explanation/justification at all. Many reported that they could not even guess what the experiment was about, and several others suggested that it might be testing how the text-message medium affects tone recognition. These results indicate that participants did not trivially identify the research questions and attempt to give "correct" responses, which provides validation for the experiment design.

[1] Check back here in late December to see the full analysis!

Chapter 5

General Discussion

5.1 Summing up

The results of the behavioral study show that comprehenders do consider speakers' production preferences for communicative efficiency when evaluating explanations. They do not increase their belief in a proposition just in accordance with the weight of evidence provided in its favor. Instead, they identify when more reasons are provided than expected, and derive inferences about what prior beliefs the speaker predicted them to have held, in order to make an argument worthwhile. These pragmatic inferences seem to affect the strength of the argument: when more reasons are given than expected, the argument fails to influence the listener's posterior beliefs in the desired direction, resulting in a sense of "the lady doth protest too much." But when reasons are contextually appropriate, they reliably influence the interlocutor's beliefs in the desired direction.

5.2 Limitations and future work

5.2.1 Non-monotonicity

Although the present findings are in support of a pragmatic analysis of argument strength, there is no evidence for true non-monotonicity. While it seems that the reduced prior did, in fact, weaken the argument for the **reasonsUnexpected** items in the **reasonsGiven** condition in comparison to the **reasonsExpected** items, it was not enough to produce post-utterance favorability ratings lower than in the **noReasons-**

Given condition. Impressionistic evaluation of the free-text data suggests some potential avenues for further work: some participants commented that it was difficult to determine whether the character made a good decision when no reasons at all were given, while others commented that many of the explanations seemed excessive or unnecessary (as designed). Non-monotonicity might be more easily identified, then, if a third condition with just one or two reasons was added to the experiment. If participants are unconvinced when no reasons are given, and suspicious when too many are given, perhaps the optimal number is somewhere between the two extremes. A reduction in argument strength might then be observed between the "few reasons" and "many reasons" conditions.

5.2.2 Discourse vs subject-matter manipulation

Another useful qualitative finding from the free-text responses was that the discourse manipulation seemed much more effective in dismissing the "defensive" interpretation for **reasonsExpected** items than the subject-matter manipulation. Many participants commented that the speaker had explained herself because the addressee seemed curious, or wanted to know more, although some did also comment on the stakes or typicality of the decision being explained. Since it seems that the **reasonsExpected** condition might indeed be induced fully-discoursally, a more tightly controlled replication of the present study becomes possible. In brief, the expected/unexpected sets could be replaced by a manipulation of the question under discussion (QUD; Roberts, 2012). If the (explicit) QUD is "Why [action]?" or similar, at least some amount of explanation is expected, which might be enough to suppress the informativity inference, even if the action described does not usually require multiple reasons in the deliberative process. As such, the same actions and reasons could be used in both conditions, allowing for better control over item-level variability.

5.2.3 Content of the reasons

This study was primarily concerned with the absolute quantity of reasons. However, the free-text results also suggest that the *type* of reason might have an effect on the inferences that arise. Since the scenarios used in this experiment involved justifications for actions, many different "flavors" of reason are possible: the character might *want* to perform the action, or she might feel *obligated* to do it, or else she might insist that it is *permissible*, and these all concievably might generate different inferences. For

example, a character over-explaining why she *wanted* to make a charitable donation generated inferences of "virtue signaling", and perhaps trying to recruit the listener to contribute as well. By contrast, a character over-explaining why he *had* to go to the library repeatedly generated the inference that he was avoiding spending time with the listener. Further work might consider whether the type and typicality of the reasons provided can affect the extent to which over-explanation can backfire (in the same way that the type and typicality of modifiers affects the informativity inferences they generate; Degen et al, 2020; Sedivy, 2003).

5.2.4 Credence-raising and vigilance to the source

The results of the present study suggest avenues for further research on the subject of "credence-raising" moves, and their role in negotiating disbelief and disagreement. The discourse-level credence-raising addressed in this paper targets what Sperber et al (2010) describe as "vigilance to the content": speakers provide reasons to support a claim in order to make the claim itself more plausible to listeners. However, speakers can also construct their utterances to target "vigilance to the source". For example, by indicating their own maximal subjective credence towards a controversial proposition, speakers indicate that they are reliable testifiers, and are willing to accept a higher reputational cost for false testimony (Vullioud et al, 2017). However, an informativity-based analysis would seem to predict that marking maximal subjective credence towards a proposition that *isn't* obviously controversial might generate inferences similar to the one described here. For example, an utterance of *"I'm **certain** it's raining in Glasgow today"* seems odd unless *"It's raining in Glasgow today"* would not be acceptable on its own. In ordinary contexts, then, the former might conceivably produce lower posterior beliefs than the latter.

Although I don't know of any psycholinguistic work on this topic, it does crop up on occasion in the formal literature. Notably, von Fintel and Gillies (2010) make reference to the very same Shakespeare passage in their defense of strong epistemic *must*, to account for how an uncertainty inference might still arise from an expression of maximal subjective credence. It is often noted that verum focus is infelicitous out of the blue, yet it DOES seem appropriate on propositions that are simply unlikely based on the preceding discourse, in addition to the classic correction and confirmation cases (Lai, 2012). Prior work on propositional intensification with items such as *really* (e.g. Romero and Han, 2004) and *totally* (e.g. Beltrama, 2018) suggest that these items are

licensed pragmatically when the proposition would not be admissible to the common ground by default (both of these accounts propose a metaconversational contribution of approximately "the proposition should for sure be added to the common ground"). Informativity analyses might be available for these and numerous other items that indicate sincerity or certainty, which would seem to run afoul of the Quantity maxim in cases where common ground admissibility would normally be taken for granted.

5.3 Concluding Remarks

Ultimately, this study can contribute to our understanding of the role credibility in everyday language use. It seems intuitive that speakers should construct their utterances to be more believable to listeners, and that listeners should reject utterances that seem implausible. Natural languages provide us with numerous mechanisms for raising the plausibility of a proposition in a context. Yet, we also regularly use language to share knowledge that would be difficult or even impossible to fully verify, and this sort of knowledge is indispensable in all aspects of our daily lives (Levy, 2021; Mercier and Sperber, 2017). It is well-established that there must be some correlation between language processing and truth judgment (Sperber et al, 2010), yet the nature of this relationship is at this time quite opaque. Future work on this topic, I hope, can shed new light on the relationship between langauge and belief.

Bibliography

Anscombre, J-C., & Ducrot, O. (1983). *L'argumentation dans la langue*. Mardaga.

Asher, N., & Lascarides, A. (2003). *Logics of conversation*. Cambridge University Press.

Aylett, M., & Turk, A. (2004). The Smooth Signal Redundancy Hypothesis: A Functional Explanation for Relationships between Redundancy, Prosodic Prominence, and Duration in Spontaneous Speech. *Language and Speech*, *47*(1), 31–56. https://doi.org/10.1177/00238309040470010201

Barnett, S. A., Griffiths, T. L., & Hawkins, R. D. (2022). A Pragmatic Account of the Weak Evidence Effect. *Open Mind*, *6*, 169–182. https://doi.org/10.1162/opmi_a_00061

Bell, A., Jurafsky, D., Fosler-Lussier, E., Girand, C., Gregory, M., & Gildea, D. (2003). Effects of disfluencies, predictability, and utterance position on word form variation in English conversation. *The Journal of the Acoustical Society of America*, *113*(2), 1001–1024. https://doi.org/10.1121/1.1534836

Beltrama, A. (2018). Totally Between Subjectivity and Discourse. Exploring the Pragmatic Side of Intensification. *Journal of Semantics*, *35*(2), 219–261. https://doi.org/10.1093/semant/ffx021

Bergen, L., & Goodman, N. D. (2015). The Strategic Use of Noise in Pragmatic Reasoning. *Topics in Cognitive Science*, *7*(2), 336–350. https://doi.org/10.1111/tops.12144

Bicknell, K., Elman, J. L., Hare, M., McRae, K., & Kutas, M. (2010). Effects of event knowledge in processing verbal arguments. *Journal of Memory and Language*, *63*(4), 489–505. https://doi.org/10.1016/j.jml.2010.08.004

Bürkner, P.-C. (2017). Brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, *80*(1), 1–28. https://doi.org/10.18637/jss.v080.i01

Bürkner, P.-C., & Vuorre, M. (2019). Ordinal Regression Models in Psychology: A Tutorial. *Advances in Methods and Practices in Psychological Science*, *2*(1), 77–101. https://doi.org/10.1177/2515245918823199

Chandra, K., Chen, T., Li, T., Ragan-Kelley, J., & Tenenbaum, J. (2024). Cooperative Explanation as Rational Communication. *Proceedings of the Annual Meeting of the Cognitive Science Society*, 46. Retrieved from https://escholarship.org/uc/item/8bf5g4h6

Clark, H. H. (1996). *Using language*. Cambridge University Press.

Crone, P. (2018). Assertions of Clarity & Raising Awareness. *Journal of Semantics*, *36*(1), 53–97. https://doi.org/10.1093/jos/ffy006

Cummins, C., & Franke, M. (2021). Rational Interpretation of Numerical Quantity in Argumentative Contexts. *Frontiers in Communication*, *6*. https://doi.org/10.3389/fcomm.2021.662027

Davis, C., Potts, C., & Speas, M. (2015). The Pragmatic Values of Evidential Sentences. *Semantics and Linguistic Theory*, 71. https://doi.org/10.3765/salt.v0i0.2966

Degen, J., Hawkins, R. D., Graf, C., Kreiss, E., & Goodman, N. D. (2020). When redundancy is useful: A Bayesian approach to "overinformative" referring expressions.*Psychological Review, 127*(4), 591–621. https://doi.org/10.1037/rev0000186

Degen, J., Tessler, M. H., & Goodman, N. D. (2015). Wonky worlds: Listeners revise world knowledge when utterances are odd. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *37*(0). https://escholarship.org/uc/item/9wn4w9zk

Farkas, D. F., & Bruce, K. B. (2009). On Reacting to Assertions and Polar Questions. *Journal of Semantics*, *27*(1), 81–118. https://doi.org/10.1093/jos/ffp010

von Fintel, K., & Gillies, A. S. (2010). Must . . . stay . . . strong! *Natural Language Semantics*, *18*(4), 351–383. https://doi.org/10.1007/s11050-010-9058-2
Florian Jaeger, T. (2010). Redundancy and reduction: Speakers manage syntactic information density. *Cognitive Psychology*, *61*(1), 23–62. https://doi.org/10.1016/j.cogpsych.2010.02.002
Frank, M. C., & Goodman, N. D. (2012). Predicting Pragmatic Reasoning in Language Games. *Science*, *336*(6084), 998–998. https://doi.org/10.1126/science.1218633
Gardenfors, P. (1980). A Pragmatic Approach to Explanations. Philosophy of Science, 47(3):404–423, 1980. ISSN 0031-8248
Garvey, & Caramazza. (1974). Implicit Causality in verbs. *Linguistic Inquiry*, *5*, 459-646.
Godden, D., & Zenker, F. (2018). A probabilistic analysis of argument cogency. *Synthese*, *195*(4), 1715–1740. JSTOR. https://doi.org/10.2307/26750710
Grice, H. P. (1975). Logic and conversation. *Communications*, *30*(1), 41–58. https://doi.org/10.3406/comm.1979.1446
Hahn, U., & Oaksford, M. (2007). The rationality of informal argumentation: A Bayesian approach to reasoning fallacies. *Psychological Review*, *114*(3), 704–732. https://doi.org/10.1037/0033-295x.114.3.704
Halpern, J. Y., & Pearl, J. (2000). Causes and Explanations: A Structural-Model Approach, Part I: Causes. *ArXiv (Cornell University)*. https://doi.org/10.48550/arxiv.cs/0011012
Harding, J., Gerstenberg, T., & Icard, T. (2025). A Communication-First Account of Explanation. *ArXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2505.03732
Hartshorne, J. K. (2014). What is implicit causality? *Language, Cognition and Neuroscience, 29*(7), 804–824. https://doi.org/10.1080/01690965.2013.796396
Hao, H., He, M., & Fuchs, Z. (2024). Greta is a female director: When gender stereotypes interact with informativity expectations. *Proceedings of the Annual Meeting of the Cognitive Science Society*, 46. Retrieved from https://escholarship.org/uc/item/49d0685w
Hobbs, J. R. (1979). Coherence and Coreference*. *Cognitive Science*, *3*(1), 67–90. https://doi.org/10.1207/s15516709cog0301_4
Kao, J. T., Wu, J. Y., Bergen, L., & Goodman, N. D. (2014). Nonliteral understanding of number words. *Proceedings of the National Academy of Sciences*, *111*(33), 12002–12007. https://doi.org/10.1073/pnas.1407479111
Kehler A. (2002). Coherence, Reference, and the Theory of Grammar. Stanford, CA: CSLI Publ.
Kehler, A. (2022). Coherence Establishment as a Source of Explanation in Linguistic Theory. *Annual Review of Linguistics*, *8*(1), 123–142. https://doi.org/10.1146/annurev-linguistics-011619-030357
Kehler, A., Kertz, L., Rohde, H., & Elman, J. L. (2007). Coherence and Coreference Revisited. *Journal of Semantics*, *25*(1), 1–44. https://doi.org/10.1093/jos/ffm018
Kehler, A., & Rohde, H. (2017). Evaluating an Expectation-Driven Question-Under-Discussion Model of Discourse Interpretation. *Discourse Processes*, *54*(3), 219–238. https://doi.org/10.1080/0163853X.2016.1169069
Kravtchenko, E., & Demberg, V. (2022). Modeling atypicality inferences in pragmatic reasoning. *Proceedings of the Annual Meeting of the Cognitive Science Society*, 44. Retrieved from https://escholarship.org/uc/item/7630p08b
Kravtchenko, E., & Demberg, V. (2022). Informationally redundant utterances elicit pragmatic inferences. *Cognition*, *225*, 105159. https://doi.org/10.1016/j.cognition.2022.105159
Lasersohn, P. (1999). Pragmatic Halos. *Language*, *75*(3), 522. https://doi.org/10.2307/417059

Lassiter, D., & Franke, M. (2024). The rationality of inferring causation from correlational language. *Proceedings of the Annual Meeting of the Cognitive Science Society*, 46. Retrieved from https://escholarship.org/uc/item/9p29w77n

Lemke, R., Reich, I., Schäfer, L., & Heiner Drenhaus. (2021). Predictable Words Are More Likely to Be Omitted in Fragments–Evidence From Production Data. *Frontiers in Psychology*, *12*. https://doi.org/10.3389/fpsyg.2021.662125

Lai, Catherine. (2012). Rises All the Way Up: The Interpretation of Prosody, Discourse Attitudes and Dialogue Structure. PhD thesis, University of Pennsylvania

Levy, N. (2021). *Bad Beliefs*. Oxford University Press.

Levy, R., & Jaeger, T. F. (2007). Speakers optimize information density through syntactic reduction. *Advances in Neural Information Processing Systems*, 849–856. https://doi.org/10.7551/mitpress/7503.003.0111

Mahowald, K., Fedorenko, E., Piantadosi, S. T., & Gibson, E. (2013). Info/information theory: Speakers choose shorter words in predictive contexts. *Cognition*, *126*(2), 313–318. https://doi.org/10.1016/j.cognition.2012.09.010

McCready, E. (2015). *Reliability in Pragmatics*. OUP Oxford.

McKoon, G., Greene, S. B., & Ratcliff, R. (1993). Discourse models, pronoun resolution, and the implicit causality of verbs.*Journal of Experimental Psychology: Learning, Memory, and Cognition, 19*(5), 1040–1052. https://doi.org/10.1037/0278-7393.19.5.1040

Mercier, H. (2022). *NOT BORN YESTERDAY : the science of who we trust and what we believe.* Princeton Univ Press.

Mercier, H., & Sperber, D. (2017). *The enigma of reason*. Harvard University Press.

Merin, A. (1999). "Information, Relevance, and Social Decision-Making: Some Principles and Results of Decision-Theoretic Semantics," in *Logic, Language, and Computation*. Editors L. S. Moss, J. Ginzburg, and M. de Rijke (Stanford, CA: CSLI Publications), Vol. 2, 179–221.

Oaksford, Mike & Ulrike Hahn. 2004. A Bayesian approach to the argument from ignorance. Canadian Journal of Experimental Psychology 58. 75–85

Oey, L. A., Schachner, A., & Vul, E. (2022). Designing and detecting lies by reasoning about other agents. *Journal of Experimental Psychology: General*. https://doi.org/10.1037/xge0001277

Piantadosi, S. T., Tily, H., & Gibson, E. (2011). Word lengths are optimized for efficient communication. *Proceedings of the National Academy of Sciences*, *108*(9), 3526–3529. https://doi.org/10.1073/pnas.1012551108

Pluymaekers, M., Mirjam Ernestus, & R. Harald Baayen. (2005). Lexical frequency and acoustic reduction in spoken Dutch. *Journal of the Acoustical Society of America*, *118*(4), 2561–2569. https://doi.org/10.1121/1.2011150

R Core Team. (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Reksnes, VRS, Rees, A, Cummins, C & Rohde, H 2024, Tell me something I don't know:Speaker salience and style affect comprehenders' expectations for informativity. in R Lemke, L Schäfer & I Reich (eds), Information Structure and Information Theory: Topics at the Grammar-Discourse Interface 10. Topics at the Grammar-Discourse Interface, Language Science Press, Berlin, pp. 177-214. https://doi.org/10.5281/zenodo.12784266

Roberts, C. (2012). Information structure in discourse: Towards an integrated formal theory of pragmatics. *Semantics and Pragmatics*, *5*. https://doi.org/10.3765/sp.5.6

Rohde, H. & Elman, J. (2007). Pronoun interpretation as a side effect of discourse coherence. Proceedings of the 29th Annual Conference of the Cognitive Science Society. 2007.

Rohde, H., Futrell, R., & Lucas, C. G. (2021). What's new? A comprehension bias in favor of informativity. *Cognition*, *209*, 104491. https://doi.org/10.1016/j.cognition.2020.104491

Rohde, H., & Rubio-Fernandez, P. (2022). Color interpretation is guided by informativity expectations, not by world knowledge about colors. *Journal of Memory and Language*, *127*, 1–10. https://doi.org/10.1016/j.jml.2022.104371

Romero, M., & Han, C.-H. (2004). On Negative Yes/No Questions. *Linguistics and Philosophy*, *27*(5), 609–658. https://doi.org/10.1023/b:ling.0000033850.15705.94

Rubio-Fernández, P. (2016). How Redundant Are Redundant Color Adjectives? An Efficiency-Based Analysis of Color Overspecification. *Frontiers in Psychology*, *7*. https://doi.org/10.3389/fpsyg.2016.00153

Ryzhova, M., Mayn, A., & Demberg, V. (2023). What inferences do people actually make upon encountering informationally redundant utterances? An individual differences study. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *45*(45). https://escholarship.org/uc/item/88g7g5z0

Sedivy, J. C. (2003). Pragmatic versus form-based accounts of referential contrast: evidence for effects of informativity expectations . *Journal of Psycholinguistic Research*, *32*(1), 3–23. https://doi.org/10.1023/a:1021928914454

Shannon, C. E. (1948). A Mathematical theory of communication. *Bell System Technical Journal*, *27*(3), 379–423. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x

Silva, V. M., Lorson, A., Franke, M., Cummins, C., & Winter, B. (2024). Strategic use of English quantifiers in the reporting of quantitative information. *Discourse Processes*, *61*(10), 498–523. https://doi.org/10.1080/0163853x.2024.2413311

Solstad, T., & Bott, O. (2022). On the nature of implicit causality and consequentiality: the case of psychological verbs. *Language, Cognition and Neuroscience*, *37*(10), 1311–1340. https://doi.org/10.1080/23273798.2022.2069277

Sperber, D., Clement, F., Heintz, C., Mascaro, O., Mercier, H., Origgi, G., & Wilson, D. (2010). Epistemic vigilance. *Mind & Language*, *25*(4), 359–393. https://doi.org/10.1111/j.1468-0017.2010.01394.x

Sperber, D., & Wilson, D. (1986). *Relevance: Communication and Cognition*. Blackwell Publishing.

Stalnaker, R. (1978). Assertion. Formal Semantics, pages 147–161.

Torabi Asr, F., & Demberg, V. (2020). Interpretation of Discourse Connectives Is Probabilistic: Evidence From the Study of But and Although. *Discourse Processes*, *57*(4), 376–399. https://doi.org/10.1080/0163853x.2019.1700760

Vullioud, C., Clément, F., Scott-Phillips, T., & Mercier, H. (2017). Confidence as an expression of commitment: Why misplaced expressions of confidence backfire. *Evolution and Human Behavior, 38*(1), 9–17. https://doi.org/10.1016/j.evolhumbehav.2016.06.002

Wallbridge, S., Bell, P., & Lai, C. (2021). It's not what you said, it's how you said it: discriminative perception of speech as a multichannel communication system. ArXiv.org. https://arxiv.org/abs/2105.00260v1

Wiegmann, A. (2023). Does lying require objective falsity? *Synthese (Dordrecht. Print)*, *202*(2). https://doi.org/10.1007/s11229-023-04291-3

Winterstein, G. (2012). What but-sentences argue for: An argumentative analysis of but. *Lingua*, *122*(15), 1864–1885. https://doi.org/10.1016/j.lingua.2012.09.014
Winterstein, G. (2018). *A Bayesian approach to Argumentation within Language.* https://semanticsarchive.net/Archive/mFkZGM0N/Argumentation.pdf
Zipf, G. K. (1935). *The psycho-biology of language.* Houghton, Mifflin.