Strategic commitments with misleading implicatures

Conversational implicatures were originally conceptualized to arise through mutual cooperativity between interlocutors (Grice, 1975), but a recent trend in pragmatics has sought to account for their appearance in strategic and adversarial contexts. A debate on the subject of "deceptive implicatures"– false or misleading implicatures derived from true assertions– has emerged at the intersection of linguistics and philosophy. Theoretical accounts have reached near-consensus that there is some intuitive distinction between lying, which involves truth conditions, and misleading, which involves inference. However, a growing body of experimental evidence suggests that the folk concept of lying is not so rigidly delineated by the semantics/pragmatics interface (Willemsen and Weigmann, 2017; Meibauer, 2018). And much less has been said about why, or indeed whether, this apparent distinction between lying and misleading is psychologically meaningful, or under what circumstances a speaker might exploit such a distinction when constructing deceptive utterances (Franke et al., 2020).

To address the former issue, recent theoretical (Meibauer, 2023) and experimental (Weigmann et al., 2021) works have proposed that lie judgments track *graded speaker commitments*, which motivates the present approach to the latter issue. Speaker commitment, especially pragmatic commitment, is notoriously difficult to define, but I'll consider an empirically-motivated definition involving a tradeoff between the credibility of an utterance and the reputational cost incurred if it is found to be unreliable. Then, to address the question of strategic misleading, I'll argue as follows: (1) Assuming this definition of commitment holds, implicatures must be weakly committed. (2) Deceptive speakers make weakly committed utterances to reduce the reputational cost of falsehood. Therefore, it should be the case that (3) implicatures can be used to strategically mitigate commitment, allowing the speaker to sacrifice some (but not all) of the credibility of their utterance in order to pay a reduced cost for deception.

Before beginning in earnest, I'll briefly motivate this project: as it stands, the primary offer on the table to explain the utility of misleading rather than lying, following Pinker et al. (2008), is actually that the speaker maintains *plausible deniability* with respect to implicatures. To give a basic example: Say the speaker asserts "I spent some of the money", through which they implicate that they didn't spend all of it. Since truth conditions are not affected by implicatures, the speaker may deny having intended "not all" without contradicting themself (if it were to become known that they did in fact spend all of the money). And since the implicature is derived by inference, the blame for "misunderstanding", then, falls at least partially upon the listener.

This proposal has plenty of merit, as it is intuitively the case that appearing to mislead by accident is less reputationally damaging than lying on purpose, and empirically-attested uncertainty in implicature calculation (e.g. Goodman and Frank, 2016; Degen, 2023) may well account for some of the mixed results on lie judgment tasks. But from a strategic perspective, it is also apparent that post-hoc implicature cancellation (i.e. "I didn't mean to suggest that…"), while always *logically* acceptable, may not necessarily be credible to the listener if cooperativity is not taken for granted, especially because the speaker's honesty is already in question (Pinker et al., 2008). Therefore, a supplementary analysis is motivated to account for cases where a strategic speaker might choose to implicate rather than assert falsehoods even when they *do not* expect that they will be able to convince the listener that the implicature was an accident, or when it is unlikely that the listener will be willing to take the blame for misinterpreting. As such, the commitment-based analysis which I undertake affects not the meaning of the utterance itself, but rather the extent to which the speaker has presented it as credible.

The notion of speaker commitment crops up frequently in conjunction with related linguistic phenomena such as speech acts, evidentiality, and modality, but it has historically been difficult to pin down as a concept in its own right (Moeschler, 2013). A recent analysis from Geurts (2019) characterizes commitment as a social relation between conversational participants towards a proposition, which allows the interlocutors to coordinate their actions. Others in linguistics have contended that such a social commitment is achieved through a sanction on falsehood, which provides interlocutors with a reason for acting in accordance with the truth of committed utterances (e.g. Krifka, 2020). A similar concept is regularly cited in game theory and economics, following Schelling (1960)– when players with unaligned motives attempt to coordinate their actions, they can make their threats and promises more credible by accepting a high cost for reneging. And it has also been argued from the perspective of evolutionary psychology that stable communication could not have evolved if there were no cost associated with false utterances (since deceptive communicators would easily outperform gullible communicators), and that accepting a cost for falsehood makes a speaker's utterance credible (Mercier, 2020; Vullioud et al., 2017). Therefore, an *epistemically vigilant* listener, who evaluates the credibility of the source and content of an interlocutor's testimony before forming beliefs, would then be more epistemically justified (all else equal) in accepting a proposition if it is strongly committed, and would assign a reputational cost to speakers who commit to unreliable signals (Mercier 2020; Sperber et al. 2010).

All of these accounts differ in many respects, but I'll extract the common thread which will be useful for the rest of this argument: Commitments are a means of coordinating actions

between agents whose interests are not necessarily aligned, whereby a sender can verify the credibility of their message by accepting a reputational cost for an unreliable signal. The opposite also applies– by signaling an unwillingness to commit, the sender's message becomes less credible.

From an empirical standpoint, there is some emerging evidence to support the idea that weakly committed claims are less persuasive than strongly-committed claims, but are also less reputationally costly if false. Vullioud et al. (2017), using expressions of confidence as an indicator for commitment, found that speakers who produced more confident utterances were more readily believed by listeners, but they were also more harshly judged reputationally when uttering falsely (in terms of punishment for the utterance itself, and future trust of the speaker) in comparison to speakers whose testimony was accepted on the basis of competence instead. Although they did not test the credibility component, Mazzarella et al. (2018), working under a similar hypothesis that implicatures are weakly committed, found that false information delivered by implicature was less reputationally costly than the equivalent information delivered by presupposition or explicit assertion, even when the implicature was readily calculated by the listener. And likewise, although they did not test the cost-of-falsehood component, Degen et al. (2019) and Lorson et al. (2023) varied the strength of epistemic modals and attitude reports respectively, and determined that weaker commitment was associated with attribution of a lesser evidential state by the listener, which I take to indicate lesser justification (again, all else equal) for the listener to form a belief based on the weakly committed utterance.

So how do implicatures fit into this picture? Typically, implicatures are not considered committed, since they are, as previously mentioned, definitionally defeasible (Grice, 1975). Defeasibility presents a problem according to definitions which require the speaker to defend a commitment when challenged (e.g. Viebahn, 2017), since many implicatures can, in fact, be easily denied. But it also cannot be the case that there is no reputational cost associated with falsely implicating, or else that listeners do not form beliefs based on implicatures, or else that speakers are not expected to act in accordance with pragmatically derived meaning, if implicatures are a stable, fundamental component of day-to-day conversation (following the same argument presented by Mercier, 2020). The most likely explanation, then, is that implicatures are committed to some extent (and some more than others depending on context), but for the most part less so than explicit assertions. Additionally, commitments need not be intentional (cf. Geurts, 2019, although he speculates full commitment to implicature)– which accounts for the fact that, even if the speaker truly has implicated accidentally, the listener would

still feel somewhat entitled to act in accordance with the truth of the implicated content and to apply a reputational cost if the utterance turns out to be false.

To return to our original question of why a speaker might mislead strategically instead of lying: Lying is usually theorized to involve a commitment in some respect– in fact, commitment comes free of charge in most assertion-based accounts, since speakers are taken to be committed to the truth of any asserted proposition (Meibauer, 2016; Stokke, 2016). In other words, a speaker hasn't lied simply by uttering something false– the deception involves presenting that utterance as though it is true and fully credible. In the alternative "what is said"-based accounts, which do not stipulate an assertion, a "warranting context" is generally specified instead, to rule out fiction, jokes etc, where the speaker isn't expected to endorse or stand by the propositions they put forward (Saul, 2012). From both perspectives, the commitment or truth-warrant is generally taken to be either present or absent. But given the analysis presented above, using a misleading implicature in place of an explicit lie could be useful– the listener might successfully recover the meaning of the utterance and have reason to believe it, but would not feel as though they've been "actively persuaded" towards a falsehood the way they might have in the explicit case, and therefore, punish it less severely.

The next question to address, then, is whether speakers *do* in fact moderate their commitment level strategically to suit their goals in an adversarial or noncooperative context, rather than to cooperatively report their own certainty or uncertainty around the communicated proposition. If this is indeed the case, it seems plausible that weakly-committed implicatures could be used for this purpose.

The answer to this question is more obviously "yes" in cases where the speaker *increases* commitment to make their utterances *more* credible: For example, the speaker might upgrade an assertion to a swear (Krifka, 2020), or choose an utterance of "know" over "believe" (Lorson et al., 2023) when attempting to persuade a disbelieving or adversarial interlocutor. Even adverbs that emphasize commitment, such as "literally", "genuinely" etc., seem to frequently precede claims that the speaker might expect would not be credible on their own (Krifka, 2020). However, there's still reason to believe that speakers might attempt to *lessen* their commitment to reduce the cost of falsehood, as well. Here's an example of a speaker attempting to withdraw commitment retroactively:

(1) "Well, one lesson I've learned is that just because I say something to a group and they laugh doesn't mean it's going to be all that hilarious as a post on 𝕏," Musk said. "Turns

out that jokes are WAY less funny if people don't know the context and the delivery is plain text."[1]

In this example, the speaker attempts to avoid the cost of an explicitly asserted proposition by reframing it as a joke, once interlocutors begin to hold him accountable for his statement. While this example isn't particularly *successful* in my eyes (especially since the commitment modification happens only after the cost is applied), I think it's still worth noting that weakening (or completely withdrawing) commitment in this way is a strategy that people regularly attempt to use, especially in online text-based communication– where the "meaning" of the utterance is clear, but the extent to which the speaker has presented it as true or credible might be more ambiguous without prosodic cues.

Furthermore, despite the perpetual theoretical relitigation of Bill Clinton's perjury trial following Solan and Tiersma (2005), it often goes without mention that Clinton repeatedly claimed that he couldn't *remember* the events being discussed:

(2) Q. . . . At any time were you and Monica Lewinsky together alone in the Oval Office?
A. "I don't recall. . . . She—it seems to me she brought things to me once or twice on the weekends. In that case, whatever time she would be in there, drop it off, exchange a few words and go, she was there" (Solan and Tiersma, 2005).

This type of deception is referred to in the literature as "proviso lying" (or in this case, perhaps, proviso merely-misleading), in which the speaker asserts a known-false proposition and then hedges it, either by inviting the listener to disbelieve it ("but you don't have to take my word for it"), or by claiming incomplete evidence for making the assertion ("but I was pretty drunk when I saw") (Arico and Fallis, 2013). The objective of a proviso lie is to present information that the speaker *knows* is false as though it is weakly believed-true despite a high level of uncertainty; without the proviso, the falsehood is delivered with much less uncertainty. So, upon hearing the proviso, the interlocutor has less reason, but still *some* reason, to believe the false proposition, in contrast to the unhedged assertion. The speaker's expectation is that their deception, if discovered, would feel less persuasive, and therefore less blameworthy, to the listener– essentially the same as I've proposed above for the implicature case.

[1]https://variety.com/2024/digital/news/elon-musk-deletes-post-kamala-harris-assassination-joke-1236145953/

The takeaway here is that speakers don't just modify their commitment to reflect their actual evidential or epistemic state. In strategic scenarios, they can reduce the level of commitment to a false proposition, since weakly committed utterances are less reputationally costly if unreliable. However, I will clarify that the utility of this conversational strategy is contextually dependent. For example, if the speaker knows that they are generally perceived as trustworthy by the listener, weak commitment might be sufficient to achieve their persuasive goals. But if the interlocutor is mistrustful of the speaker already, it may be necessary to increase the commitment to successfully persuade them, and then accept the full cost of lying if discovered.

From a theoretical perspective, I see no reason why such a strategy cannot apply to implicatures as well– after all, there are many examples of implicature for which the reduction of commitment is transparently the primary communicative goal. Consider for example the "recommendation letter" paradigm case from Grice (1975):

(3) Dear Sir, Mr. X's command of English is excellent, and his attendance at tutorials has been regular. Yours, etc.

If we reframe this as a "misleading implicature" case, where Mr. X is in fact an excellent philosopher whom the professor simply doesn't like, it seems to me that an utterance of (3) would be less actively harmful to X's admissions opportunities– i.e. less convincing– than a direct utterance of "Mr. X is not good at philosophy". But by my intuition, it also seems to be significantly less reputationally damaging if discovered to be false (the effect feels even more pronounced in the misleading case than in the case where Mr. X is actually a poor student). In any case, the professor's intended meaning is clear, and it would be very silly to attempt to deny it, but leaving the lie unsaid seems to weaken it significantly nevertheless.

Ultimately, though, the question at hand is empirical in nature, and the existing data is sparse. A handful of studies have shown that false implicatures are somewhat less reputationally costly than explicit lies (e.g. Mazzarella et al., 2018, Yuan and Lyu, 2022). However, most of the experimental work on misleading implicature at this point has focused on listener judgments, rather than speaker production strategies (cf. Franke et al., 2020). Further studies to verify the claim presented here might involve testing the speakers' preference for lying vs falsely implicating with trustful and mistrustful interlocutors, to determine if speakers calibrate their level of commitment depending on how likely they suspect they are to be believed. In addition, to my knowledge, no experiments have tested the credibility of

implicatures in comparison to assertions. To serve this purpose, a replication of Vullioud (2017), where the level of explicitness is varied instead of confidence, would be suitable to test both sides of the cost-credibility tradeoff discussed in this paper.

To sum up, in response to the question of why speakers might use misleading implicatures strategically, I've motivated an alternative analysis to the usual "plausible deniability" account, proposed a definition of speaker commitment based on a tradeoff between cost and credibility, and shown how weakly committed implicatures might be used to make a false claim appear trustworthy enough without seeming too deceptive if discovered. Although I think this analysis is viable, further formalization and empirical support will be required for me to commit to it completely.

References

Arico, A. J., & Fallis, D. (2013). Lies, damned lies, and statistics: An empirical investigation of the concept of lying. *Philosophical Psychology*, *26*(6), 790–816. https://doi.org/10.1080/09515089.2012.725977

Degen, J. (2023). The Rational Speech Act Framework. *Annual Review of Linguistics*, *9*(1), 519–540. https://doi.org/10.1146/annurev-linguistics-031220-010811

Degen, J., Trotzke, A., Scontras, G., Wittenberg, E., & Goodman, N. D. (2019). Definitely, maybe: A new experimental paradigm for investigating the pragmatics of evidential devices across languages. *Journal of Pragmatics*, *140*, 33–48. https://doi.org/10.1016/j.pragma.2018.11.015

Franke, M., Dulcinati, G., & Pouscoulous, N. (2019). Strategies of Deception: Under‐Informativity, Uninformativity, and Lies—Misleading With Different Kinds of Implicature. *Topics in Cognitive Science*, *12*(2), 583–607. https://doi.org/10.1111/tops.12456

Geurts, B. (2019). Communication as commitment sharing: speech acts, implicatures, common ground. *Theoretical Linguistics*, *45*(1-2), 1–30. https://doi.org/10.1515/tl-2019-0001

Goodman, N. D., & Frank, M. C. (2016). Pragmatic Language Interpretation as Probabilistic Inference. *Trends in Cognitive Sciences*, *20*(11), 818–829. https://doi.org/10.1016/j.tics.2016.08.005

Grice, H. P. (1975). Logic and Conversation. *Speech Acts*, *3*, 45. https://doi.org/10.1163/9789004368811_003

Jennifer Mather Saul. (2015). *Lying, misleading, and what is said : an exploration in philosophy of language and in ethics*. Oxford Univ Press.

Jörg Meibauer. (2014). *Lying at the Semantics-Pragmatics Interface*. Walter de Gruyter GmbH & Co KG.

Krifka, M. (2020). Layers of Assertive Clauses: Propositions, Judgements, Commitments, Acts. *Propositional Arguments in Cross-Linguistic Research: Theoretical and Empirical Issues*.

Lorson, A., Rohde, H., & Cummins, C. (2023). Epistemicity and communicative strategies. *Discourse Processes*, *60*(8), 556–593. https://doi.org/10.1080/0163853x.2023.2255494

Mazzarella, D., Reinecke, R., Noveck, I., & Mercier, H. (2018). Saying, presupposing and implicating: How pragmatics modulates commitment. *Journal of Pragmatics*, *133*, 15–27. https://doi.org/10.1016/j.pragma.2018.05.009

Meibauer, J. (2018). The Linguistics of Lying. *Annual Review of Linguistics*, *4*(1), 357–375. https://doi.org/10.1146/annurev-linguistics-011817-045634

Meibauer, J. (2023). On commitment to untruthful implicatures. *Intercultural Pragmatics*, *20*(1), 75–98. https://doi.org/10.1515/ip-2023-0004

Mercier, H. (2020). *Not Born Yesterday : The Science of Who We Trust and What We Believe*. Princeton University Press.

Moeschler, J. (2013). Is a speaker-based pragmatics possible? Or how can a hearer infer a speaker's commitment? *Journal of Pragmatics*, *48*(1), 84–97. https://doi.org/10.1016/j.pragma.2012.11.019

Pinker, S., Nowak, M. A., & Lee, J. J. (2008). The logic of indirect speech. *Proceedings of the National Academy of Sciences*, *105*(3), 833–838. https://doi.org/10.1073/pnas.0707192105

Schelling, T. C. (1960). *The strategy of conflict*. Cambridge, Mass. Harvard Univ. Pr.

Solan, L., & Peter Meijes Tiersma. (2005). *Speaking of crime : the language of criminal justice*. University Of Chicago Press.

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

Stokke, A. (2016). Lying and Misleading in Discourse. *Philosophical Review*, *125*(1), 83–134. https://doi.org/10.1215/00318108-3321731

Viebahn, E. (2017). Non-literal Lies. *Erkenntnis*, *82*(6), 1367–1380. https://doi.org/10.1007/s10670-017-9880-8

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

Wiegmann, A., & Willemsen, P. (2017). *How the truth can make a great lie: An empirical investigation of the folk concept of lying by falsely implicating.* 3516–3521. https://doi.org/10.5167/uzh-175885

Wiegmann, A., Willemsen, P., & Meibauer, J. (2022). Lying, Deceptive Implicatures, and Commitment. *Ergo an Open Access Journal of Philosophy*, *8*(0). https://doi.org/10.3998/ergo.2251

Yuan, W., & Lyu, S. (2022). Speech act matters: Commitment to what's said or what's implicated differs in the case of assertion and promise. *Journal of Pragmatics*, *191*, 128–142. https://doi.org/10.1016/j.pragma.2022.01.012