Coffee or Date? A Bayesian account of Indirect Disclosure in High-Stakes Social Interactions

Victor Btesh and David Lagnado
Department of Experimental Psychology, University College London
{victor.btesh.19, d.lagnado}@ucl.ac.uk

Abstract

We present a computational model of how people navigate high-stakes social interactions in which they want to disclose sensitive information but are unsure how the other person will react. We show that indirect communication emerges from recursive pragmatic inference over a simple endorsement structure: asking someone for coffee entails both friendship and romance, while a date rules out friendship alone. People's actions are guided by utilities resting on inferences over the belief states of others, and these goals can be at odds. In romantic settings, people only acted directly when certain their feelings would be reciprocated; otherwise they chose indirect actions to probe the other while protecting their own intimacy. Our model captures these trade-offs via two quantities beyond a desire for a positive response: strategic ambiguity, a motivation to obscure one's true goal, and motivated search for alignment, an interest in learning about the other's goal insofar as it matches one's own. Together, these components closely match participants' behaviour across 6 cover stories and 6 levels of uncertainty over the listener's belief state, providing evidence that high-level social goals are represented as information-theoretic quantities computed over a Bayesian Theory of Mind.

Introduction

Indirect communication is a staple of social life. When we ask someone for coffee rather than a date, or mention a political opinion as a question rather than a claim, we navigate a fundamental tension: the need to express ourselves against the risk of rejection or exposure. Pinker et al. (2008) formalised this as a game-theoretic identification problem in which plausible deniability is optimal under uncertainty about the listener's intentions, a framework extended to loopholes by Qian et al. (2024). Yet, these do not directly demonstrate how utilities governing indirect speech (e.g. fear of rejection, need for expression, desire to protect one's image) shift under different belief states of the listener. Building on these accounts, we propose that utilities emerge from recursive pragmatic inference and that their weights shift dynamically with the speaker's beliefs about what the listener wants, knows, and believes about the speaker. We test this approach in a vignette experiment manipulating three nested belief states across six ecologically valid scenarios.

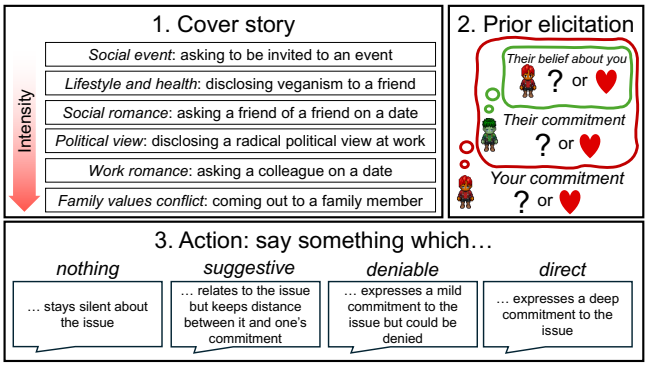

Figure 1: Structure of a single trial. Participants read a text describing the setting of one of the 6 cover stories. We then elicit priors by letting them know whether they are *uncertain* (question marks) or *certain* (hearts) about their own commitment, that of the listener, and the belief the listener has of the participant's commitment. Finally, they pick an action from the 4 available options.

Model

We model indirect speech for disclosure as a pragmatic inference problem. Statements and actions $a \in A$ (Silence $\rightarrow$ Direct) vary in how strongly they endorse a speaker's latent commitment to the issue being discussed $c \in C$ (Indifferent $\rightarrow$ Committed). This structure is captured by a design matrix $\Psi \in \{-1, 0, 1\}^{|A| \times |C|}$, where $\Psi_{ac} = 1$ if action a endorses goal c, -1 if it discards it, and 0 if neutral:

$$\Psi = \begin{pmatrix} 1 & 1 & 0 & 0 \\ -1 & 1 & 1 & 0 \\ -1 & -1 & 1 & 1 \\ -1 & -1 & -1 & 1 \end{pmatrix} \tag{1}$$

A naive speaker chooses actions proportionally to their endorsement, $p(a \mid c) = \sigma(\beta_s \Psi_{ac})$, and a naive listener responds positively in proportion to how well the action endorses their own goal, $p(r \mid a, c) = \sigma(\beta_l \cdot r \cdot \Psi_{ac})$. A pragmatic listener L_1 further infers the speaker's goal from their action and updates their response utility ac-

Extended abstract presented at the *9th Conference on Cognitive Computational Neuroscience*, New York, NY, USA, 2026.

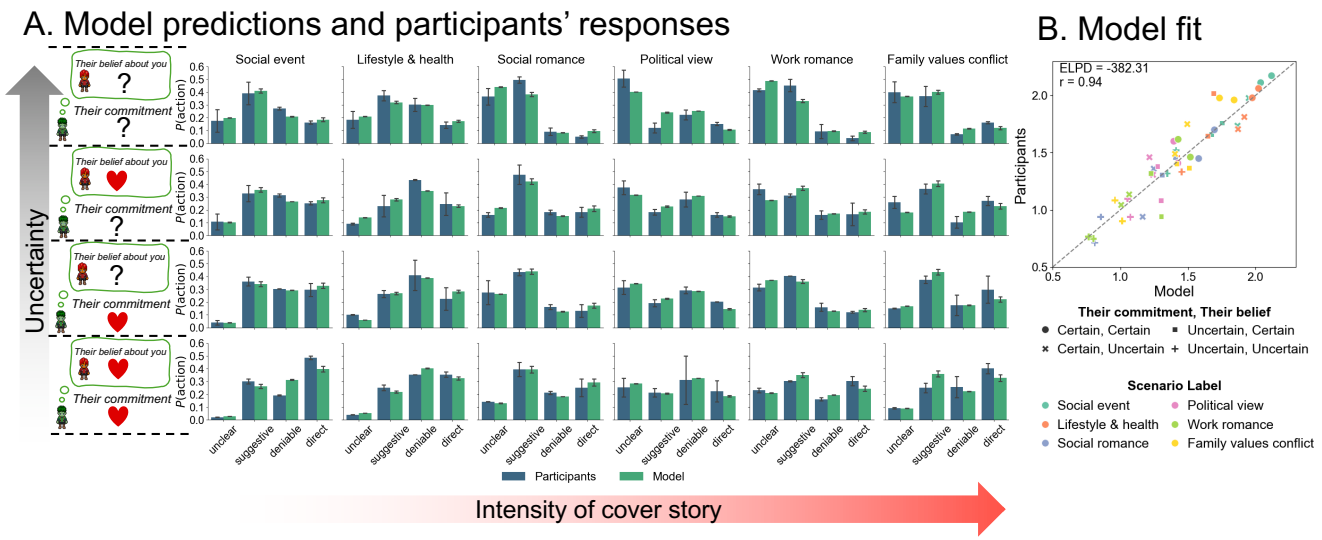

Figure 2: A. Model predictions and participants responses (averaged over speaker's commitment). The likelihood of direct action decreases with both uncertainty in the listener's belief state and in the intensity of the cover story. B. Model fit, ELPD (expect log probability density) and Pearson's r, for the full model for all within conditions.

cordingly:

$$U_{L_1}(r;\, a, c) = r\left[\Psi_{ac} + \log p(c_{\text{speaker}} = c \mid a)\right] \qquad (2)$$

The speaker reasons about this listener and selects actions according to:

$$\begin{aligned} U_{S_2}(a, c;\, \rho) \;=\; & \underbrace{\alpha_{\text{positive}} \cdot \log p(r = \text{positive} \mid a,\, c_{L_1})}_{\text{need for positive response}} \\ & + \underbrace{\alpha_{\text{ambiguity}} \cdot H\big[\rho(c),\; p_{L_1}(c \mid a)\big]}_{\text{strategic ambiguity}} \\ & - \underbrace{\alpha_{\text{align}} \cdot \mathbb{E}_{p(r|a)}\big[D_{\text{KL}}\big[p(c_{L_1} \mid r) \,\|\, \rho(c)\big]\big]}_{\text{motivated search for alignment}} \end{aligned}$$

where $\rho(c)$ is a distribution encoding the speaker's own commitment. Need for positive response maximises the log-likelihood of r = positive given the listener's belief and action a. Strategic ambiguity maximises the cross-entropy between the listener's posterior over the speaker's commitment under a and the speaker's true $\rho(c)$. Motivated search for alignment seeks actions whose expected responses would, on average, reveal a listener goal matching $\rho(c)$. Uncertainty in each nested belief state is represented as the variance of a Gaussian prior fitted per certainty condition.

Experiment

99 participants (Prolific, US/UK) completed a vignette task across 6 scenarios spanning romantic, family, and workplace contexts. Three belief states were manipulated: *your commitment* (between subjects), *their commitment* (within), and *their belief about you* (within), each at two levels (certain and uncertain), yielding 4 within subject variants per scenario and 24 trials total. Each trial presented a base scenario establishing the social context, followed by three condition-specific sentences describing the speaker's own commitment, their belief about the listener's commitment, and their belief about what the listener thinks of them. Participants then selected among four responses (Nothing, Suggestive, Deniable, or Direct), matched to the rows of Ψ (see Figure 1).

Results and Discussion

We fitted the model to the whole sample's response frequencies with MCMC (Phan et al., 2019). We assumed the weights given to each term would vary across stories and so fitted them with a hierarchical Dirichlet distribution. We also hierarchically fitted the variances across the belief conditions along with a softmax inverse temperature. We compared the full model with ablated variants. Removing the need for positive expression renders the model at chance (r = .19, ELPD = -528.65); removing strategic ambiguity produces a substantial drop (r = .70, ELPD = -529.98). Removing motivated search for alignment leaves average fit intact (r = .91, ELPD = -433.85) but loses the bimodalities observed in some scenarios. The full model best captures the complete pattern (r = .94, ELPD = -384.34, see Figure 2).

Across 6 diverse cover stories, the model provides evidence that people are sensitive to changes in information-theoretic quantities computed over the listener's belief state by capturing a graded shift from direct to indirect action as uncertainty increases. Yet, the model still fails to capture the stakes in the relationship itself. In romantic overtures or family conflict, the consequences of action extend beyond a single exchange and span the entire trajectory of the relationship (Lee & Pinker, 2010). Future work should seek to represent not only the speaker's local beliefs and goals but their model of what the relationship is, what it could become, and what would be lost or gained by making that transition explicit.

Acknowledgements

VB was supported by a Bogue Fellowship and an UCL Experimental Psychology Demonstratorship.

References

Lee, J. J., & Pinker, S. (2010). Rationales for indirect speech: The theory of the strategic speaker. *Psychological Review*, *117*(3), 785–807. https://doi.org/10.1037/a0019688.

Phan, D., Pradhan, N., & Jankowiak, M. (2019). Composable Effects for Flexible and Accelerated Probabilistic Programming in NumPyro. *arXiv preprint arXiv:1912.11554*.

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.

Qian, P., Bridgers, S., Taliaferro, M., Parece, K., & Ullman, T. D. (2024). Ambivalence by design: A computational account of loopholes. *Cognition*, *252*, 105914. https://doi.org/10.1016/j.cognition.2024.105914.