What Do You Think I Thought You'd Think? Recursive ToM for Pragmatic Understanding in CUIs

Natalia Tyulina
CUNY Graduate Center
New York, USA
ntyulina@gradcenter.cuny.edu

ABSTRACT

Despite recent advances in natural language understanding and generation, Conversational User Interfaces (CUIs) still fall short of meeting human expectations for pragmatic nuance in dialogue. This paper argues that advancing Theory of Mind (ToM) in human–AI interaction requires moving beyond literal intent recognition toward models of communication grounded in recursive social reasoning. Drawing on the Rational Speech Act (RSA) framework and neo-Gricean pragmatics, I propose an agenda for integrating Bayesian models of cognition into the architecture of next-generation CUIs.

KEYWORDS

Conversational User Interfaces, Theory of Mind, Recursive Reasoning, Inference, Rational Speech Acts, Common Ground, Implicature

ACM Reference Format:
Natalia Tyulina. 2025. What Do You Think I Thought You'd Think? Recursive ToM for Pragmatic Understanding in CUIs. In *Proceedings of Theory of Mind in Human-CUI Interaction Workshop @ ACM CUI 2025 (ToMinHAI '25). (CUI '25).* ACM, New York, NY, USA, 4 pages.

1 INTRODUCTION

Conversational User Interfaces (CUIs) have made unprecedented progress in language comprehension and task-oriented dialogue, driven by advances in large language models (LLMs) and dialogue management systems [12, 17, 23, 24]. Yet CUIs often struggle in situations that require subtle inference, indirectness, or other settings in which humans naturally rely on the Theory of Mind (ToM) to infer others' mental states.

The Rational Speech Act (RSA) framework is a relatively recent development, largely attributed to the work of Frank & Goodman [6] and Goodman & Stuhlmüller [8]. However, it builds on several well-established theoretical foundations spanning a range of interdisciplinary sources. RSA draws, on the one hand, from earlier philosophical and linguistic research on Speech Acts [1, 18]. It also models communication as a probabilistic process in which speakers choose utterances based on how they expect a listener to interpret them. In turn, listeners reason backward from the utterance to infer the beliefs and goals that would have made it a rational choice for the speaker. RSA formalizes this process using tools from Bayesian cognition and information theory, providing a computational account of pragmatic language understanding[7].

CUI '25, July 8–10, 2025, Waterloo, ON, Canada

RSA also builds on earlier game-theoretic foundations. Lewis was among the first to connect natural language pragmatics with cooperative game strategies, framing communication as a signaling game [15]. RSA similarly draws from Grice's Cooperative Principle [10], which holds that effective communication depends on shared assumptions of cooperation between interlocutors. Grice proposed four conversational maxims, Quantity, Quality, Relation, and Manner, as guiding dimensions for how interlocutors infer each other's intent and unspoken implications. Even though they have been critiqued as overly broad or idealized, these maxims are a good starting point for conceptually decomposing the notion of communicative cooperation. Table 1 presents a reinterpretation of these dimensions as they apply to CUIs.

Table 1: Grice's Conversational Maxims for CUIs

Maxim	Description
Quantity	Provide the amount of information the user is likely expecting—no more, no less.
Quality	Ensure the response is accurate and verifiable; avoid speculation or misleading claims.
Relation	Tailor the response to the user's inferred goals or information needs, even when not explicitly stated.
Manner	Communicate in a way that is consistent but not repetitive; avoid unnatural phrasing or formatting that may confuse the user.

While some CUI research has incorporated Gricean maxims to support intent recognition [19], it often stops short of modeling recursive reasoning, such as reasoning about the user's mental states, including expectations about the system's response.

To illustrate this distinction, I present a sample interaction with a CUI, tested with two response variants:

> **User:** I don't want to talk to you anymore.
> **CUI 1:** I understand. If you ever want to chat again, I'll be here. Take care!

> **User:** I don't want to talk to you anymore.
> **CUI 2:** Got it — if you ever feel like coming back, I'll be here. Take care, Natasha.

This type of user utterance often carries an implicit social meaning: users typically would not say anything unless they expect the

system to try to re-engage them and continue the dialogue. Recognizing this intent and responding accordingly requires a form of basic recursive reasoning. Yet, this kind of inference is still beyond the capabilities of even most sophisticated CUIs.

Goodman and Frank [9] demonstrate the power of this approach through referential communication games. In the next section, I show how it can be applied in the context of CUI-user interaction. By assuming that speakers choose utterances that optimize informativity relative to context, and that listeners apply Bayesian reasoning to reverse-engineer the speaker's intent, RSA is capable of capturing certain nuanced patterns of language use.

Recent extensions of the RSA have expanded the core concept of rationality to include special cases, such as politeness [22], hyperbole and affect [14], metaphor [2], deception [4], and decisions under lexical uncertainty [16]. These approaches enrich the speaker's utility function with additional social, emotional, or epistemic components. For example, in modeling politeness, speakers are assumed to weigh social utility against epistemic utility, leading to utterances that may be less informative in a literal sense but more socially appropriate. Such refinements demonstrate how recursive reasoning can accommodate nonliteral meaning, indirectness, and affective nuance, core aspects of human communication that CUIs frequently struggle to manage.

2 RECURSIVE REASONING AS A FOUNDATION FOR TOM IN CUIS

A foundational perspective on rational action under uncertainty is provided by Bayesian Decision Theory, as formalized by Berger [3]. In this framework, an agent faces an unknown state of the world, denoted θ, and must choose an action a that minimizes expected loss based on its beliefs. The optimal action a^* is defined as:

$$a^* = \arg\min_a \mathbb{E}_{\theta|x}[L(a, \theta)]$$

Here, θ represents the latent state of the world, a is a possible action the agent can take, x denotes the agent's observations or evidence, and $L(a, \theta)$ is the loss incurred by taking action a when the true state is θ. The expectation is taken over the agent's posterior belief $P(\theta \mid x)$ about the world state given its observations.

This formalism underlies much of probabilistic modeling in both statistics and cognitive science. When applied to communication, it frames interpretation as inverse planning: the listener infers the speaker's goals by assuming that the utterance was chosen to optimize communicative utility. Crucially, this aligns with evidence from psycholinguistics [13], [11] that human communication involves belief modeling and expectation management.

Let us first assign CUIs the role of a listener within the RSA framework, as an agent that interprets utterances by recursively modeling the speaker's mental state. This framing foregrounds the importance of pragmatic nuance in how *users* evaluate a CUI's cognitive sophistication and how *they* attribute agency and intentionality to the system.

Fig. 1 illustrates recursive reasoning as introduced in Goodman and Frank [9], adapted for a human–CUI interaction scenario. It depicts an RSA-style model of pragmatic inference, where the CUI reasons not only about the literal meaning of the user's utterance, but about the speaker's deeper communicative intent, including their reasoning about what the system will infer. To achieve this, the CUI simulates a simplified internal model of the speaker (S), who in turn reasons about a literal listener (Lit)—a minimal model of how the CUI might interpret referring expressions.

At the Lit level, the word *beard* applies both to the individual with only a beard (B) and the one with both a hat and a beard (HB), yielding a uniform distribution over these two referents. In contrast, *hat* uniquely identifies HB. A rational speaker S who intends to refer to HB would therefore prefer the more informative *hat* whereas *beard* would be selected to refer to B.

The CUI, acting as a pragmatic listener, inverts this reasoning: upon hearing "beard," it infers that if the user had intended HB, they likely would have said "hat" instead. Thus, "beard" is interpreted as most likely referring to B. This nested structure of inference allows the CUI to resolve ambiguity by considering speaker goals, utterance alternatives, and context—critical for interpreting underspecified input in natural conversation.

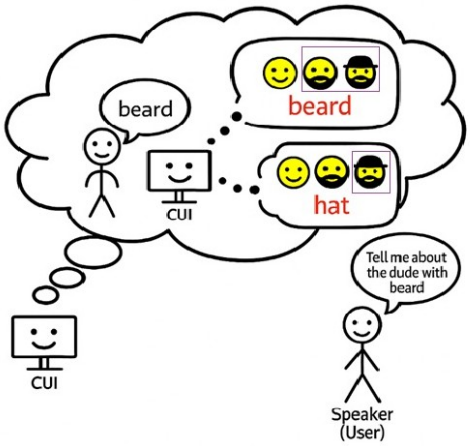

Figure 1: Recursive inference in CUI reasoning, inspired by Goodman and Frank [9].

A recursive ToM model, especially one informed by RSA and its extensions (e.g., utility-based speaker models), offers a formal framework for implementing these ideas. To embed this in CUIs, I advocate for hybrid architectures that combine neural encoders with symbolic and probabilistic RSA modules. These systems should not only estimate user goals from the dialogue context, but also explicitly simulate what the user might assume about the system's beliefs, an essential step toward mutual modeling and adaptive pragmatics.

To define this proposal more concretely, we can frame CUI reasoning in terms of two complementary components: inference and action selection. The system must infer the most likely user goals given an utterance, and then select an action that optimizes its communicative utility under uncertainty.

Inference Objective. The CUI infers a posterior over world states θ given the user utterance u and context w:

$$P(\theta \mid u, w) \propto P(u \mid \theta, w) \cdot P(\theta \mid w)$$

System Action Selection. Let $a \in \mathcal{A}$ be a candidate system action (e.g., repeat the answer, apologize, change topic). The optimal action minimizes the expected loss.

While this formal framing provides a clean abstraction, real-world implementation raises several challenges. These arise not only from the computational costs of recursive inference, but also from the context-sensitive nature of human interaction. Below, I outline some of the most pressing modeling challenges for CUIs.

Modeling Challenges.

- **Recursion depth:** Determining how many levels of reasoning are computationally tractable and cognitively plausible, while still capturing relevant user expectations.
- **Dynamic priors:** User goals and emotional states evolve over time: trust, urgency, and frustration fluctuate between interactions.
- **Multimodal cues:** Nonverbal signals such as prosody, timing, gaze, or facial expressions influence inferred intent and must be integrated along with text.
- **Alternative space construction:** Accurately modeling the set of possible world states or speaker intentions.

3 DISCUSSION AND FUTURE WORK

While this paper has focused on CUIs as pragmatic listeners who interpret user utterances through recursive inference over user mental states, robust human-like communication ultimately demands bidirectional reasoning. To fully participate in social dialogue, CUIs must also act as speakers who select utterances based on internal models of how listeners will interpret them. This speaker-side reasoning is implicitly embedded in RSA but has yet to be fully operationalized in language generation pipelines. Integrating both roles would allow CUIs to dynamically adapt to evolving conversational context, user-specific priors, and common ground. Such systems could more effectively simulate mutual ToM and participate in socially intelligent dialogue.

Recent studies reinforce this need. Soubki et al. [20] introduced COMMON-TOM, a benchmark derived from naturally occurring spoken dialogues designed to evaluate ToM competency in LLMs using the notion of common ground [21]. They point out, that in cognitive science literature [5], common ground not only involves shared knowledge but also allows for false or mistaken beliefs about others' mental states. Their model, fine-tuned on COMMON-TOM, outperforms GPT-4 on second- (i.e., A believes B believes p) and third-order (i.e., A believes B believes A believes p) belief tasks, despite having significantly fewer parameters.

Table 2 outlines key types of pragmatic implicature that are essential for effective human–CUI interaction. These categories draw on both established pragmatic theory and recent computational extensions within RSA-based modeling. While several of these inference types are already implemented within existing RSA formulations, others remain open challenges, particularly, when applied at scale.

Informed by these phenomena, I propose an empirical research agenda to advance ToM in CUIs:

Table 2: Types of Pragmatic Implicature with Relevance to CUIs

Implicature Type	Description	Example Use Cases
Quantity	Inferring meaning from what is not said (e.g., "some" implies "not all")	Visual referencing; constrained explanations
Relevance	Interpreting intent based on context and conversational goals	Task disambiguation; context switching
Politeness	Indirect or softened language to preserve face	User correction; negative feedback
Manner	Drawing inference from vagueness, unnaturalness, or over-specificity	Vague queries; response formatting
Affective	Exaggerated or emphatic language to signal emotion	Frustration modeling; empathy response
Scalar	Choosing among graded alternatives (e.g., *good* < *great*, *may* < *have to* < *must*)	Recommendations; preference tracking
Presuppose	Assumptions about shared knowledge or discourse history	Elliptical follow-ups; multi-turn cohesion
Non-literal	Deviations from literal use to convey creativity or humor	Entertainment bots; narrative agents

- **Gricean Maxim Sensitivity:** Conduct controlled experiments where CUIs intentionally violate different conversational maxims. User ratings on informativeness, relevance, clarity, and trust can be used to model perception of mental states and communicative intent.
- **QUD Alignment:** Develop multi-turn tasks that require CUIs to shift attention to new Questions Under Discussion. Measure whether users perceive the system as flexible and contextually aware.
- **Common Ground Repair:** Create interactive tasks involving misalignment or misunderstanding, requiring CUIs to re-establish common ground. Evaluate on user perceptions of recovery and alignment.
- **Alternatives and Priors Estimation:** Design interactive tasks in which users choose plausible interpretations or responses. Aggregate these to empirically estimate context-sensitive priors and alternative sets for RSA-based models.
- **Belief Attribution via Probing:** Use subtle prompts (e.g., "You probably know this already...") to assess whether users attribute beliefs or intentions to the CUI based on its response behavior.
- **User Modeling and Individual Differences:** Investigate how cognitive styles, social preferences, or ToM tendencies

shape interaction with recursive reasoning agents. Use this insight to tailor personalization strategies.

While this research agenda spans a range of pragmatic phenomena, some directions may offer a more feasible entry point than others. For example, investigating Gricean maxim sensitivity and QUD alignment can be prioritized in early-stage work. As the architecture matures, user modeling and priors estimation could be layered in to support deeper personalization and probabilistic reasoning. This staged approach enables iterative development while maintaining empirical grounding.

4 CONCLUSION

In human–human interaction, conversational implicature enables communication that is efficient and collaborative. Speakers tend to convey more than they explicitly say, and listeners infer these meanings with minimal cognitive effort. This mutual pragmatics fosters engagement, trust, and a sense of being understood. For CUIs, supporting implicature is thus not just a matter of pragmatic competence, but a core design goal.

Engagement is inherently bidirectional: users feel at ease when they can say less and still be heard, and when systems avoid redundancy or over-explaining. Yet implicature depends on recursive reasoning: the capacity to model not just what users say, but what they believe, expect, and assume about the system's own beliefs. CUIs that model such reasoning reduce user frustration and enable more fluid, natural conversation. In doing so, they bring us closer to systems that truly understand, and are understood.

REFERENCES

[1] J. L. Austin. 1962. *How to Do Things with Words: The William James Lectures Delivered in Harvard University in 1955.* Oxford University Press UK, Oxford, England.

[2] Leon Bergen and Noah D. Goodman. 2020. Pragmatic reasoning through semantic inference. *Topics in Cognitive Science* 12, 3 (2020), 701–717.

[3] James O. Berger. 1985. *Statistical Decision Theory and Bayesian Analysis.* Springer, New York. https://doi.org/10.1007/978-1-4757-4286-2

[4] Connor Briggs, Justine Kao, Leon Bergen, and Noah D. Goodman. 2021. Modeling the semantics and pragmatics of deceptive language. In *Proceedings of the 43rd Annual Meeting of the Cognitive Science Society.*

[5] Sarah Brown-Schmidt and Melissa C. Duff. 2016. Memory and Common Ground Processes in Language Use. *Topics in Cognitive Science* 8, 4 (2016), 722–736. https://doi.org/10.1111/tops.12224

[6] Michael C. Frank and Noah D. Goodman. 2012. Predicting Pragmatic Reasoning in Language Games. *Science* 336, 6084 (2012), 998–998. https://doi.org/10.1126/science.1218633 arXiv:https://www.science.org/doi/pdf/10.1126/science.1218633

[7] Michael C. Frank and Noah D. Goodman. 2012. Predicting Pragmatic Reasoning in Language Games. *Science* 336, 6084 (2012), 998–998. https://doi.org/10.1126/science.1218633 arXiv:https://www.science.org/doi/pdf/10.1126/science.1218633

[8] Noah Goodman and Andreas Stuhlmüller. 2013. Knowledge and Implicature: Modeling Language Understanding as Social Cognition. *Topics in cognitive science* 5 (01 2013), 173–184. https://doi.org/10.1111/tops.12007

[9] Noah D. Goodman and Michael C. Frank. 2016. Pragmatic Language Interpretation as Probabilistic Inference. *Trends in Cognitive Sciences* 20, 11 (2016), 818–829. https://doi.org/10.1016/j.tics.2016.08.005

[10] H. P. Grice. 1975. Logic and Conversation. In *Syntax and Semantics: Vol. 3: Speech Acts*, Peter Cole and Jerry L. Morgan (Eds.). Academic Press, New York, 41–58. http://www.ucl.ac.uk/ls/studypacks/Grice-Logic.pdf

[11] Daniel Grodner and Julie C. Sedivy. 2011. The effect of speaker-specific information on pragmatic inferences. In *The Processing and Acquisition of Reference*, Edward Gibson and Neal Pearlmutter (Eds.). MIT Press, Cambridge, MA, 239–272.

[12] Patrick Kahardipraja, Brielen Madureira, and David Schlangen. 2021. Towards Incremental Transformers: An Empirical Analysis of Transformer Models for Incremental NLU. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (Eds.). Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 1178–1189. https://doi.org/10.18653/v1/2021.emnlp-main.90

[13] Yuki Kamide, Gerry T. M. Altmann, and Sarah L. Haywood. 2003. The time-course of prediction in incremental sentence processing: Evidence from anticipatory eye movements. *Journal of Memory and Language* 49, 1 (2003), 133–156. https://doi.org/10.1016/S0749-596X(03)00023-8

[14] Justine Kao, Leon Bergen, and Noah D. Goodman. 2014. Nonliteral understanding of number words. In *Proceedings of the 36th Annual Conference of the Cognitive Science Society (CogSci).* 719–724.

[15] David Kellogg Lewis. 1969. *Convention: A Philosophical Study.* Wiley-Blackwell, Cambridge, MA, USA.

[16] Christopher Potts, Daniel Lassiter, Roger Levy, and Michael C. Frank. 2015. Embedded Implicatures as Pragmatic Inferences under Compositional Lexical Uncertainty. *Journal of Semantics* 33, 4 (12 2015), 755–802. https://doi.org/10.1093/jos/ffv012 arXiv:https://academic.oup.com/jos/article-pdf/33/4/755/8260418/ffv012.pdf

[17] Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2019. Towards Empathetic Open-domain Conversation Models: A New Benchmark and Dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Anna Korhonen, David Traum, and Lluís Màrquez (Eds.). Association for Computational Linguistics, Florence, Italy, 5370–5381. https://doi.org/10.18653/v1/P19-1534

[18] John Rogers Searle. 1969. *Speech Acts: An Essay in the Philosophy of Language.* Cambridge, England.

[19] Vidya Setlur and Melanie Tory. 2022. How do you Converse with an Analytical Chatbot? Revisiting Gricean Maxims for Designing Analytical Conversational Behavior. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22).* Association for Computing Machinery, New York, NY, USA, Article 29, 17 pages. https://doi.org/10.1145/3491102.3501972

[20] Adil Soubki, John Murzaku, Arash Yousefi Jordehi, Peter Zeng, Magdalena Markowska, Seyed Abolghasem Mirroshandel, and Owen Rambow. 2024. Views Are My Own, but Also Yours: Benchmarking Theory of Mind Using Common Ground. In *Findings of the Association for Computational Linguistics: ACL 2024*, Lun-Wei Ku, Andre Martins, and Vivek Srikumar (Eds.). Association for Computational Linguistics, Bangkok, Thailand, 14815–14823. https://doi.org/10.18653/v1/2024.findings-acl.880

[21] Robert Stalnaker. 2002. Common Ground. *Linguistics and Philosophy* 25, 5 (2002), 701–721. https://doi.org/10.1023/a:1020867916902

[22] Siyan Yoon, Michael H. Tessler, and Noah D. Goodman. 2020. Polite language generation with context-aware hierarchical encoder-decoder. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics.* 3755–3767.

[23] Eric Zhang. 2023. ChatGPT: The making of a viral AI chatbot. https://openai.com/blog/chatgpt. Accessed: 2025-06-23.

[24] Li Zhou, Jianfeng Gao, Di Li, and Heung-Yeung Shum. 2020. The design and implementation of XiaoIce, an empathetic social chatbot. *Computational Linguistics* 46, 1 (2020), 53–93.