

# A Formal Framework for Deceptive Topic Planning in Information-Seeking Dialogues

Extended Abstract

Andreas Brännström  
Umeå University  
Umeå, Sweden  
andreasb@cs.umu.se

Virginia Dignum  
Umeå University  
Umeå, Sweden  
virginia@cs.umu.se

Juan Carlos Nieves  
Umeå University  
Umeå, Sweden  
jcnieves@cs.umu.se

## ABSTRACT

This paper introduces a formal framework for goal-hiding information-seeking dialogues to deal with interactions where a seeker agent estimates a human respondent to not be willing to share the sought-for information. Hence, the seeker postpones (hides) a sensitive goal topic until the respondent is perceived willing to talk about it. This regards a type of deceptive strategy to withhold information, e.g., a sensitive question, that, in a given dialogue state, may be harmful to a respondent, e.g., by violating privacy. The framework uses Quantitative Bipolar Argumentation Frameworks to assign willingness scores to topics, inferred from a respondent’s asserted beliefs. A gradual semantics is introduced to handle changes in willingness scores based on relations among topics. The goal-hiding dialogue process is illustrated using an example inspired by primary healthcare nurses’ strategies for collecting sensitive health information from patients.

## KEYWORDS

Formal dialogues; Formal argumentation; Knowledge extraction; Non-collaborative agents; Machine deception

### ACM Reference Format:

Andreas Brännström, Virginia Dignum, and Juan Carlos Nieves. 2023. A Formal Framework for Deceptive Topic Planning in Information-Seeking Dialogues: Extended Abstract. In *Proc. of the 22nd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2023)*, London, United Kingdom, May 29 – June 2, 2023, IFAAMAS, 3 pages.

## 1 INTRODUCTION

In the area of formal argumentation dialogues, an information-seeking dialogue [17] is commonly defined as an interaction between a seeker agent and a respondent agent. The seeker’s overall goal is to obtain a particular set of information, assumed to be possessed by the respondent, that the seeker cannot get access to through other means than by questioning the respondent. The respondent is commonly defined as being collaborative, and has the role of providing the sought for information by answering the seeker’s questions as clearly as possible. This paper is concerned with defining a class of formal information-seeking dialogues, referred to as Goal-Hiding Dialogues, between a software seeker agent and a human respondent agent, where it is assumed that the human respondent initially is unwilling (non-collaborative) to disclose the information that the seeker wants. The seeker aims to

introduce its goal topics, while being constrained to only introduce topics for which the respondent has sufficient willingness. Thus, the seeker postpones its goal topics until the respondent is perceived to willingly talk about them.

There is a range of human settings where such dialogue strategies are present, such as criminal interrogations [9] and medical assessments [7], commonly involving sensitive information that can be difficult to talk about or to admit directly. A particular example regards health promoting dialogues [8], conducted between a primary healthcare nurse and a patient, where building trust through a tactful order of topics is central for successfully collecting sensitive health information [1]. In order to approach sensitive topics, the nurses employ strategies such as being friendly and welcoming, and introducing lighter topics, to establish trust in the dialogue.

In the area of chatbots [5], topic selection and personalization of dialogues have been approached through various techniques from Natural Language Processing (NLP) [4, 6] and Machine Learning (ML) [10, 15, 16] to build response generation models, which have enabled systems that can understand and respond to user inputs in a conversational manner. However, these methods require large amounts of social conversation data [18], typically not available in settings where sensitive topics are discussed. In a goal-hiding dialogue, to proactively select topics, a software agent must consider a human’s dynamic willingness for topics. This requires a formalism that is non-monotonic w.r.t. the state of the dialogue.

A Quantitative Bipolar Argumentation Framework (QBAF) [3] is a tuple  $\langle X, R^-, R^+, \tau \rangle$  consisting of a finite set  $X$  of arguments, a binary (attack) relation  $R^-$  on  $X$ , a binary (support) relation  $R^+$  on  $X$  and a total function  $\tau : X \rightarrow [0,1]$ ; returning the so-called base score of arguments. A total strength function  $\delta : X \rightarrow [0,1]$ ; returns the so-called strength of arguments. By analyzing arguments’ support and attack relations, the strengths are adjusted by considering a gradual semantics. This paper proposes QBAFs for modeling a human agent’s willingness for topics. By applying QBAFs, a set of arguments  $T$  represents topics, and a function  $\delta : T \rightarrow [0,1]$  generates willingness scores for topics in a given dialogue state. A game strategic gradual semantics [2] is introduced to deal with changing willingness as new topics are opened in the dialogue. We assume a measure of willingness to be quantitative, on a finite willingness score between 0 and 1, and we assume willingness to be argumentative and bipolar since it can be based on topics that promote (support) or demote (attack) willingness of other topics. We introduce a method for analyzing a respondent’s asserted beliefs to construct a QBAF-based willingness model. By adapting the model to each dialogue-state, a seeker agent can strategically promote willingness for a goal topic before opening it in the conversation.

*Proc. of the 22nd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2023)*, A. Ricci, W. Yeoh, N. Agmon, B. An (eds.), May 29 – June 2, 2023, London, United Kingdom. © 2023 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaaamas.org). All rights reserved.



## ACKNOWLEDGEMENTS

This work was partially funded by the Knut and Alice Wallenberg Foundation.

## REFERENCES

- [1] Linda Baer. 1969. Improving oncology nurses' communication skills for difficult conversations. *Number 3/June 2013* 17, 3 (1969), E45–E51.
- [2] Pietro Baroni, Giulia Comini, Antonio Rago, and Francesca Toni. 2017. Abstract games of argumentation strategy and game-theoretical argument strength. In *International Conference on Principles and Practice of Multi-Agent Systems*. Springer, 403–419.
- [3] Pietro Baroni, Antonio Rago, and Francesca Toni. 2019. From fine-grained properties to broad principles for gradual argumentation: A principled spectrum. *International Journal of Approximate Reasoning* 105 (2019), 252–286.
- [4] Jackylyn Beredo, Carlo Migel Bautista, Macario Cordel, and Ethel Ong. 2021. Generating Empathetic Responses with a Pre-trained Conversational Model. In *International Conference on Text, Speech, and Dialogue*. Springer, 147–158.
- [5] Petter Bae Brandtzaeg and Asbjørn Følstad. 2018. Chatbots: changing user needs and motivations. *Interactions* 25, 5 (2018), 38–43.
- [6] Jiajia Duan, Hui Zhao, Qian Zhou, Meikang Qiu, and Meiqin Liu. 2020. A study of pre-trained language models in natural language processing. In *2020 IEEE International Conference on Smart Cloud (SmartCloud)*. IEEE, 116–121.
- [7] Linda Ganzini, Lauren M Denneson, Nancy Press, Matthew J Bair, Drew A Helmer, Jennifer Poat, and Steven K Dobscha. 2013. Trust is the basis for effective suicide risk screening and assessment in veterans. *Journal of General Internal Medicine* 28, 9 (2013), 1215–1221.
- [8] Åsa Hörnsten, Karin Lindahl, Kristina Persson, and Kristina Edvardsson. 2014. Strategies in health-promoting dialogues—primary healthcare nurses' perspectives—a qualitative study. *Scandinavian journal of caring sciences* 28, 2 (2014), 235–244.
- [9] Kevan L Jensen and Mark W Smith. 2021. A Preliminary Examination of the Effectiveness of Assessment Questions in Detecting Dishonest Behavior. *Journal of Forensic Accounting Research* 6, 1 (2021), 127–148.
- [10] Jiwei Li, Michel Galley, Chris Brockett, Georgios P Spithourakis, Jianfeng Gao, and Bill Dolan. 2016. A persona-based neural conversation model. *arXiv preprint arXiv:1603.06155* (2016).
- [11] Peta Masters and Sebastian Sardina. 2017. Deceptive Path-Planning. In *IJCAL* 4368–4375.
- [12] Ami Rokach. 2013. Loneliness updated: An introduction. In *Loneliness Updated*. Routledge, 17–22.
- [13] Chiaki Sakama. 2012. Dishonest Arguments in Debate Games. *COMMA* 75 (2012), 177–184.
- [14] Ştefan Sarkadi, Alison R Panisson, Rafael H Bordini, Peter McBurney, Simon Parsons, and Martin Chapman. 2019. Modelling deception using theory of mind in multi-agent systems. *AI Communications* 32, 4 (2019), 287–302.
- [15] Lifeng Shang, Zhengdong Lu, and Hang Li. 2015. Neural responding machine for short-text conversation. *arXiv preprint arXiv:1503.02364* (2015).
- [16] Michael Shumanov and Lester Johnson. 2021. Making conversations with chatbots more personalized. *Computers in Human Behavior* 117 (2021), 106627.
- [17] Douglas Walton and Erik CW Krabbe. 1995. *Commitment in dialogue: Basic concepts of interpersonal reasoning*. SUNY press.
- [18] Chen Xing, Wei Wu, Yu Wu, Jie Liu, Yalou Huang, Ming Zhou, and Wei-Ying Ma. 2017. Topic aware neural response generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 31.