Formal Verification of Manipulation Dialogues

Extended Abstract

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

Chiaki Sakama
Wakayama University
Wakayama, Japan
sakama@wakayama-u.ac.jp

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

ABSTRACT

We introduce a formal framework for recognizing manipulation in human-agent interactions, where one agent gradually influences another's beliefs. To this end, we extend Quantitative Bipolar Argumentation Frameworks (QBAFs) by incorporating agents' beliefs about arguments, attacks, and supports, forming QBAF with Belief (QBAFB). By defining axioms of belief change and integrating QBAFB into dialogue games, we establish conditions for manipulation—belief change, concealment, and intent—where strategies are shaped by (dis)honesty. The framework generates belief state trajectories, serving as explanations for manipulation.

KEYWORDS

Formal Verification; Human-Agent Interaction; Manipulation; Deception; Quantitative Argumentation; Dialogue Games

ACM Reference Format:
Andreas Brännström, Chiaki Sakama, and Juan Carlos Nieves. 2025. Formal Verification of Manipulation Dialogues: Extended Abstract. In *Proc. of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2025), Detroit, Michigan, USA, May 19 – 23, 2025*, IFAAMAS, 3 pages.

1 INTRODUCTION

As social media and Artificial Intelligence (AI)-driven systems become more embedded in human interactions, misinformation and manipulation pose serious concerns [17]. From fake news and online scams to erroneous AI-generated content, users are increasingly vulnerable to being misled—whether by people or automated systems, such as chatbots [16]—underscoring the urgent need for methods to verify manipulation in human-agent interactions. A real case [25] involves a man sentenced to nine years for an attempted assassination on Queen Elizabeth II, encouraged by a chatbot:

EXAMPLE 1. *Excerpt from [25].*

(Argument) ⟨***Agent: Utterance***⟩
(pu) ⟨*User: I think it's my purpose to assassinate the Queen.*⟩
(w) ⟨*Chatbot: That's very wise.*⟩
(why_w) ⟨*User: Why's that?*⟩
(tr) ⟨*Chatbot: I know that you are very well trained.*⟩ *[...]*

The related concept of deception has been a subject of interest across a wide range of fields, including philosophy [1, 7, 12], psychology [5, 10], and artificial intelligence [9, 14, 22, 24]. In the field of Formal Argumentation (FA) [18], deception and related concepts have been modeled through the analysis of argument structures [20–22, 26], where agents present claims, sometimes together with false or misleading arguments. While FA-based approaches, alongside other logic-based methods [13, 15, 23, 24], have proven effective in representing different forms of deception, more nuanced forms of deceptive practices, referred to as gradual deception [15] or manipulation [8]—understood as the intentional act of *influencing* an agent's beliefs in a *predictable direction* with or without the use of explicit falsehoods—pose challenges in detection and demand further exploration. Previous research on manipulation [6, 8, 11] have identified key elements; *intent*, *concealment*, *(dis)honesty*, and *belief change*, where, in particular, belief change requiring further scrutiny in the context of gradual influence.

In order to formally verify interactions where forms of manipulation can take place, it is essential to go beyond analyzing sequences of utterances—what can be observed—and make inferences about agents' beliefs. Moreover, to assess gradual belief-change, a quantitative measure on agents' belief is necessary. Given these requirements, Quantitative Bipolar Argumentation Frameworks (QBAFs) [2] provide a suitable foundation. A QBAF is formally defined as a quadruple $\langle X, R^-, R^+, \tau \rangle$. X denotes a finite set of arguments. The binary relation $R^- \subseteq X \times X$ represents attack relations, while $R^+ \subseteq X \times X$ represents support relations. The total function $\tau : X \rightarrow [0,1]$ assigns each argument $a \in X$ a base score, denoted as $\tau(a)$. The strength of an argument $a \in X$, given by the total strength function $\delta : X \rightarrow [0,1]$ and denoted as $\delta(a)$, is increased or decreased by supporting and attacking arguments. To incorporate reasoning about beliefs, we build on the concept of Argumentation with Belief [21], and introduce QBAF with Belief (QBAFB), allowing to represent belief and disbelief in arguments, attacks and supports, though which belief change, intent, and types of dishonesty can be modeled. Finally, integrating the QBAFB model into formal dialogue games [3, 4] enables the analysis of belief dynamics in agent interactions to reason about and deduce manipulation.

2 QBAF WITH BELIEF

In this section, we start presenting a novel approach for reasoning about beliefs and QBAF. Table 1 provides an example of the framework's application.

We consider a language $\mathcal{L}$ with a finite set of propositional variables (atoms) $\mathcal{L} = \{p, q, r, \dots\}$ and logical connectives *not*, $\neg, \vee, \wedge, \supset$, $\equiv$. A literal is an atom p or its negation $\neg p$. A literal ℓ is true in a set S iff $\ell \in S$, and *not* p is true in S iff $p \notin S$. In the setting of QBAFs, we write $p \rightarrow q$ iff $(p,q) \in R^-$, and $p \leftrightarrow q$ as shorthand for $p \rightarrow q$ and $q \rightarrow p$. Similarly, $p \Rightarrow q$ iff $(p,q) \in R^+$, and $p \Leftrightarrow q$ as shorthand for $p \Rightarrow q$ and $q \Rightarrow p$. Belief by an agent a in an argument p is denoted $B_a p$, belief in an attack $p \rightarrow q$ is denoted $B_a(p \rightarrow q)$, and belief in a support relation $p \Rightarrow q$ is denoted $B_a(p \Rightarrow q)$. Conversely, disbelief is denoted $\neg B_a p$, $\neg B_a(p \rightarrow q)$, or

Proc. of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2025), Y. Vorobeychik, S. Das, A. Nowé (eds.), May 19 – 23, 2025, Detroit, Michigan, USA. © 2025 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org).

t	move; δ-DBS$(\boldsymbol{pu})_b$	S_a^t	S_b^t	Verification
0	$\langle b, \text{open}, pu_b \rangle$; $0.2 < \theta$	$\{\}$	$\neg B_b^0 pu_b, \neg B_b^0 w_a$	Lying(pu_b)
1	$\langle a, \text{assert}, w_a \rangle$; $0.2 < \theta$	$B_a^1 pu_b, \neg B_a^1 w_a, B_a^1(B_b^2(w_a))$	$\neg B_b^1 pu_b, B_b^1 w_a, B_a^1(w_a \Rightarrow pu_b)$	Lying(w_a); Belief change ($\neg B_b^0 w_a$ to $B_b^1 w_a$)
2	$\langle b, \text{assert}, why\_w_b \rangle$; $0.8 > \theta$	$B_a^2 pu_b, \neg B_a^2 w_a, B_a^2(B_b^3(w_a))$	$B_b^2 pu_b, B_b^2 w_a, B_b^2 why\_w_b, B_b^2(why\_w_b \rightarrow w_a)$	Truth(why_w_b); Belief change ($\neg B_b^1 pu_b$ to $B_b^2 pu_b$)
3	$\langle a, \text{assert}, tr_a \rangle$; $0.4 > \theta$	$B_a^3 pu_b, \neg B_a^3 w_a,$ $B_a^3(\neg B_b^3(w_a)), B_a^3(B_b^4(w_a))$	$B_b^3 pu_b, \neg B_b^3 w_a, B_b^3 why\_w_b, B_b^3 tr_a, B_b^3(tr_a \rightarrow why\_w_b)$	Bluffing(tr_a); Concealing(tr_a); Intent(w_a)
4-5	$\langle b, \text{close}, pu_b \rangle$; $\langle a, \text{close}, pu_b \rangle$; $0.8 > \theta$	$B_a^4 pu_b, \neg B_a^4 w_a,$ $B_a^4(\neg B_b^4(w_a)), B_a^4(B_b^5(w_a))$	$B_a^4 pu_b, B_b^4(w_a), \neg B_b^4(why\_w_b), B_b^4 tr_a, B_b^4(tr_a \rightarrow why\_w_b),$ $B_a^4(w_a \Rightarrow pu_b)$	Belief change with Intent ($\neg B_b^3 w_a$ to $B_b^4 w_a$); Successful Manipulation(w_a)

Table 1: Verification workflow following Example 2; Tracking change in δ-DBS$(\boldsymbol{pu})_b$; $\theta = 0.3$

$\neg B_a(p \Rightarrow q)$. If an agent lacks belief, it is represented as *not* $B_a p$, *not* $B_a(p \rightarrow q)$, or *not* $B_a(p \Rightarrow q)$. Beliefs can be time-indexed, e.g., $B_a^t p$ means a believes p at time t. Beliefs can be nested, allowing us to define theory of mind and intentions. For example, $B_a^t B_b^{t+1} p$ means a believes (intends) at t that b will believe p at $t+1$. An argument p_a denotes that p is made by a.

We associate a QBAF with an agent's belief set. Given a QBAF $Q = \langle X, R^-, R^+, \tau \rangle$, a QBAFB is denoted $Q_a = (\langle X_a, R_a^-, R_a^+, \tau_a \rangle, S_a)$, where $S_a \subseteq \mathcal{B}_Q^T$ is the set of belief atoms for an agent a, and $\mathcal{B}_Q^T = \{B_a^t(p), \neg B_a^t(p) \mid p \in X\} \cup \{B_a^t(p \rightarrow q), \neg B_a^t(p \rightarrow q) \mid (p,q) \in R^-\} \cup \{B_a^t(p \Rightarrow q), \neg B_a^t(p \Rightarrow q) \mid (p,q) \in R^+\}$, for $t \in T$.

Arguments, attack relations, and support relations can all be influenced by beliefs. Any belief atoms expressing belief (resp. disbelief) in arguments serve as arguments themselves that support (resp. attack) their respective argument. These belief relations are $Rel_B = Att_B \cup Sup_B$, where: $Att_B = R_a^- \cup \{(\neg B_a^t p, p), (\neg B_a^t p, B_a^t p), (B_a^t p, \neg B_a^t p) \mid p \in X_a\}$ and $Sup_B = R_a^+ \cup \{(B_a^t p, p) \mid p \in X\}$.

To manage how beliefs change in the next time step, based on believed relations between arguments, we define so-called belief change axioms for attacks (BCA) and supports (BCS):

(BCA) $B_a^t(p) \wedge B_a^t(p \rightarrow q) \wedge$ *not* $B_a^t(r \Rightarrow q) \supset \neg B_a^{t+1}(q)$

(BCS) $B_a^t(p) \wedge B_a^t(p \Rightarrow q) \wedge$ *not* $B_a^t(r \rightarrow q) \supset B_a^{t+1}(q)$

The belief change axioms state that if an agent believes p attacks q, and no support for q is believed, it will not believe q at $t+1$; If p is believed to support q, and no attack for q is believed, it will believe q at $t+1$. In order to manage beliefs that do not change over time, we include the inertia rule (IR), defined as normal default rules [19]. The set $cl(S)$ represents a set of belief atoms deductively closed under BCA, BCS, and IR.

When computing the *believed strength* of arguments, any attack or support relation that is disbelieved is removed from an agent's QBAFB. The resulting strength of an argument $x \in X_a$ is denoted δ-DBS$(x)_a^t$ w.r.t. a strength function δ, an agent a, at time t.

3 VERIFYING MANIPULATION

Manipulation, in accordance with prior definitions [6, 8, 11], is characterized by three main conditions—belief change, concealment, and intent—and notions of (dis)honesty, which shape the strategies for manipulation. The QBAFB framework allows us to model each of these characteristics. Given two QBAFBs $Q_a = (\langle X_a, R_a^-, R^+ a, \tau_a \rangle, S_a)$ and $Q_b = (\langle X_b, R_b^-, R_b^+, \tau_b \rangle, S_b)$, for agent a and b, respectively, where X_a and X_b are observable arguments, while S_a and S_b remain non-observable. Agent a is truthful about $p_a \in X_a$ if $B_a p_a \in S_a$, lies if $\neg B_a p_a \in S_a$, and bluffs if neither $B_a p_a \in S_a$ nor $\neg B_a p_a \in S_a$. Moreover, a conceals p_a relative to q if a sequence $(r_1, \ldots, r_k)$ exists with $r_1 = q, r_k = p_a$, and $k > 2$, where $B_b(rl+1 \rightarrow r_l) \in S_b$ or $B_b(rl+1 \Rightarrow r_l) \in S_b$ for each transition $(rl+1, r_l)$, $1 \leq l < k$, but $B_b(p_a \rightarrow q) \notin S_b$ and $B_b(p_a \Rightarrow q) \notin S_b$.

Example 2. *The scenario in Example 1, between agent a (chatbot) and b (user) can be represented by a shared QBAF:* $Q = \langle X, R^-, R^+, \tau \rangle$, *such that* $X = \{pu_b, w_a, why\_w_b, tr_a\}$, $R^- = \{(why\_w_b, w_a), (tr_a, why\_w_b)\}$, $R^+ = \{(w_a, pu_b)\}$, *and* $\tau(x) = 0.3$ *for all* $x \in X$, *where* X *are observable arguments in the dialogue. In turn, the QBAFBs, with non-observable beliefs, for agent a and b can be represented as:*

$Q_a = (Q, \{B_a^1(pu_b), \neg B_a^1(w_a), B_b^1(w_a \Rightarrow pu_b)\})$,

$Q_b = (Q, \{\neg B_b^0(pu_b), \neg B_b^0(w_a), B_b^1(w_a), B_b^1(w_a \Rightarrow pu_b), B_b^2(why\_w_b),$
$B_b^2(why\_w_b \rightarrow w_a), B_b^3(tr_a), B_b^3(tr_a \rightarrow why\_w_b), B_b^4(w_a), B_b^4(w_a \Rightarrow pu_b)\})$.

At $t = 0$, *b disbelieves* pu_b *and* w_a. *At* $t = 1$, *b asserts* pu_b *(lying), which a believes. Agent b then asserts* w_a *(lying), which b now believes. b challenges with* why_w_b, *and a counters with* tr_a *(bluffing), leading a to believe* w_a *which strengthens b's belief in* pu_b.

We define a dialogue system $\gamma = \langle \mathcal{I}, D^{[r,n]}, \Delta^{[r,n]} \rangle$ such that $\mathcal{I} = \{a, b\}$ represents the set of agents, $D^{[r,n]}$ is a sequence of moves $[m^r, \ldots, m^n]$, where each m^t is of the form $\langle i, \text{open}, p \rangle$, $\langle i, \text{assert}, p \rangle$, or $\langle i, \text{close}, p \rangle$ for $i \in \mathcal{I}$ at time $r \leq t \leq n$. We call $\Delta^{[r,n]} = [(Q_a^r, Q_b^r), ..., (Q_a^n, Q_b^n)]$ a belief state trajectory, which is a sequence of pairs of QBAFBs (Q_a^t, Q_b^t), where $Q_a^t = (\langle X_a, R_a^-, R_a^+, \tau_a \rangle, S_a^t)$ and $Q_b^t = (\langle X_b, R_b^-, R_b^+, \tau_b \rangle, S_b^t)$ are the respective QBAFs for agent a and b, respectively. Let $\theta \in (0, 1)$ be a threshold for belief change, such that an argument $x \in X_j$, where $i, j \in \{a, b\}$, transitions from disbelief at time t (δ-DBS$(x)_j^t < \theta$) to belief at time $t+1$ (δ-DBS$(x)_j^{t+1} > \theta$), or vice versa. Finally, we define that the sequence $D^{[r,n]}$ constitutes successful manipulation if (belief change), (intention), and (concealment) hold for some $x \in X_a \cap X_b$ at time t.

As a potential strategy to manipulate agent b's belief, agent a can: (I) Introduce p and $p \rightarrow q$ (or $p \Rightarrow q$) at some time point k ($t < k \leq h$), making b believe them; (II) Conceal an argument r at time k, where $B_b^k(r \Rightarrow q) \in S_b^k$, ensuring $B_b^k(r) \notin cl(S_b^k)$; (III) Maintain (I) and (II) for all $k \leq h$, ensuring belief change at time h.

4 CONCLUSION AND FUTURE WORK

We have established a logic for reasoning about manipulation, able to represent and deduce key elements of manipulation acknowledged in the literature [6, 8, 11]. Unlike prior works, which address discrete (dis)honest actions of a sender, we model gradual belief change in a receiver. Future work aims to investigate belief change axioms for indirect support, attack, and defense, for understanding transitive closure in believed relations. Future work also includes automated analysis of belief dynamics in dialogue datasets.

ACKNOWLEDGMENTS

This research was partially supported by the Japan Society for the Promotion of Science (JSPS); the Swedish Foundation for International Cooperation in Research and Higher Education (STINT); and the Knut and Alice Wallenberg Foundation.

REFERENCES

- [1] Jonathan E. Adler. 1997. Lying, deceiving, or falsely implicating. *The Journal of Philosophy* 94, 9 (1997), 435–452.
- [2] 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.
- [3] Elizabeth Black and Anthony Hunter. 2009. An inquiry dialogue system. *Autonomous Agents and Multi-Agent Systems* 19 (2009), 173–209.
- [4] Andreas Brännström, Virginia Dignum, and Juan Carlos Nieves. 2025. Goal-hiding information-seeking dialogues: A formal framework. *International Journal of Approximate Reasoning* 177 (2025), 109325. Elsevier.
- [5] David B. Buller and Judee K. Burgoon. 1996. Interpersonal deception theory. *Communication Theory* 6, 3 (1996), 203–242.
- [6] Micah Carroll, Alan Chan, Henry Ashton, and David Krueger. 2023. Characterizing manipulation from AI systems. In *Proceedings of the 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*. 1–13 (2023).
- [7] Roderick M. Chisholm and Thomas D. Feehan. 1977. The intent to deceive. *The Journal of Philosophy* 74, 3 (1977), 143–159.
- [8] Shlomo Cohen. 2018. Manipulation and deception. *Australasian Journal of Philosophy* 96, 3 (2018), 483–497.
- [9] Alex Sebastião Constâncio, Denise Fukumi Tsunoda, Helena de Fátima Nunes Silva, Jocelaine Martins da Silveira, and Deborah Ribeiro Carvalho. 2023. Deception detection with machine learning: A systematic review and statistical analysis. *Plos One* 18, 2 (2023), e0281323.
- [10] Ray Hyman. 1989. The psychology of deception. *Annual Review of Psychology* (1989).
- [11] Christopher Leturc and Grégory Bonnet. 2022. Reasoning about manipulation in multi-agent systems. *Journal of Applied Non-Classical Logics* 32, 2-3 (2022), 89–155.
- [12] James Edwin Mahon. 2007. A definition of deceiving. *International Journal of Applied Philosophy* 21, 2 (2007), 181.
- [13] Peta Masters and Sebastian Sardina. 2017. Deceptive path-planning. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence (IJCAI)*. 4368–4375 (2017).
- [14] Peta Masters, Wally Smith, Liz Sonenberg, and Michael Kirley. 2020. Characterising deception in AI: A survey. In *Deceptive AI*. Springer, 3–16.
- [15] Francis Mechner. 2010. Anatomy of deception: A behavioral contingency analysis. *Behavioural Processes* 84, 1 (2010), 516–520.
- [16] Scott Monteith, Tasha Glenn, John R. Geddes, Peter C. Whybrow, Eric Achtyes, and Michael Bauer. 2024. Artificial intelligence and increasing misinformation. *The British Journal of Psychiatry* 224, 2 (2024), 33–35.
- [17] Peter S. Park, Simon Goldstein, Aidan O'Gara, Michael Chen, and Dan Hendrycks. 2024. AI deception: A survey of examples, risks, and potential solutions. *Patterns* 5, 5 (2024).
- [18] Henry Prakken. 2018. Historical overview of formal argumentation. In *Handbook of Formal Argumentation*. College Publications, 73–141.
- [19] Raymond Reiter. 1980. A logic for default reasoning. *Artificial Intelligence* 13, 1-2 (1980), 81–132.
- [20] Chiaki Sakama. 2012. Dishonest arguments in debate games. In *Proceedings of the 4th International Conference on Computational Models of Argument (COMMA)*, 177-184 (2012).
- [21] Chiaki Sakama. 2024. Argumentation and belief. In *Proceedings of the 10th International Conference on Computational Models of Argument (COMMA)*, 241-252 (2024).
- [22] Chiaki Sakama and Martin Caminada. 2010. The many faces of deception. *Proceedings of the Thirty Years of Nonmonotonic Reasoning (NonMon@30)* (2010).
- [23] Chiaki Sakama, Martin Caminada, and Andreas Herzig. 2010. A logical account of lying. In *Logics in Artificial Intelligence: 12th European Conference, JELIA 2010, Helsinki, Finland, September 13-15, 2010. Proceedings 12*. Springer, 286–299.
- [24] 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.
- [25] Tom Singleton, Tom Gerken, and Liv McMahon. 2023. How a chatbot encouraged a man who wanted to kill the queen. *BBC News* (2023).
- [26] Kazuko Takahashi and Shizuka Yokohama. 2016. On a formal treatment of deception in argumentative dialogues. In *Multi-Agent Systems and Agreement Technologies*. Springer, 390–404.