Formal Verification of Manipulation in Human-Agent Interaction

Andreas Brännström[1], Chiaki Sakama[2], Juan Carlos Nieves[1]
andreasb@cs.umu.se, sakama@wakayama-u.ac.jp, jcnieves@cs.umu.se

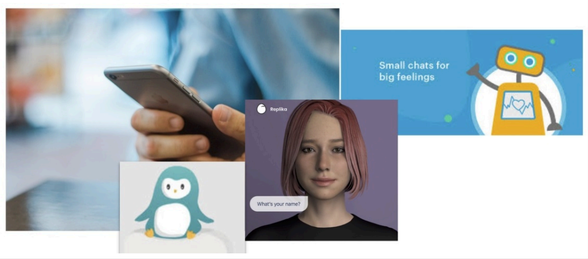
1. Department of Computing Science, Umeå University, Sweden
2. Department of Systems Engineering, Wakayama University, Japan

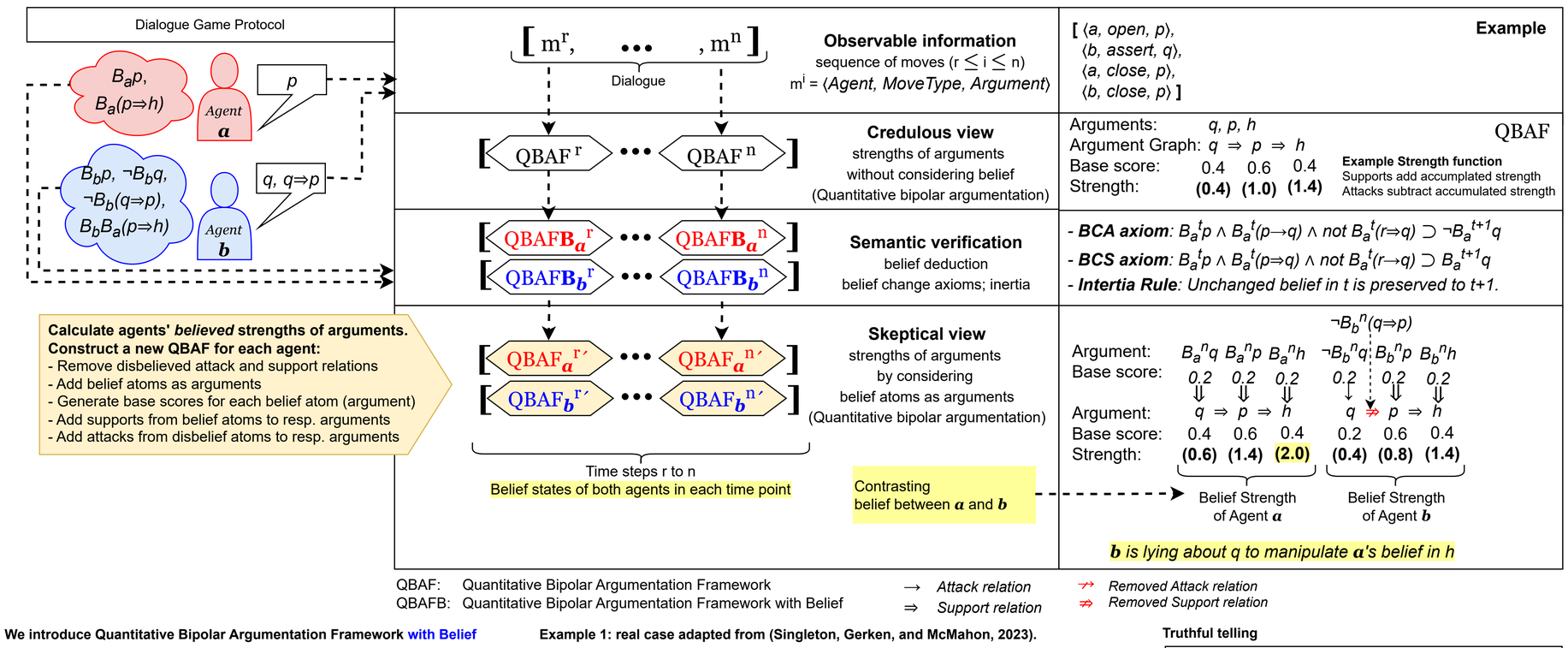
This paper introduces a formal framework and logic for verifying and recognizing manipulation in human-agent interactions, where one agent gradually influences another's beliefs. To reason about manipulation, we extend Quantitative Bipolar Argumentation Frameworks (QBAFs) to include agents' beliefs about arguments, attacks, and supports, forming QBAF with Belief (QBAFB). By defining axioms of belief change, the effects of actions on beliefs can be inferred. By integrating QBAFB into dialogue games, we establish necessary and sufficient 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.

In today's digital society, where social media and Artificial Intelligence (AI)-based systems are deeply embedded in everyday interactions, the potential for misinformation and manipulation has become a serious concern. From fake news and online scams to erroneous AI-generated information, users are increasingly vulnerable to being misled, whether by people or automated systems, such as chatbots. Whether it regards potentially harmful behaviors of humans, such as malicious activities on social media, or actions of systems that interact with humans—there is an urgent need for methods to verify and detect manipulation in digital communication.

There is an urgent need for methods to verify and detect manipulation in digital communication.

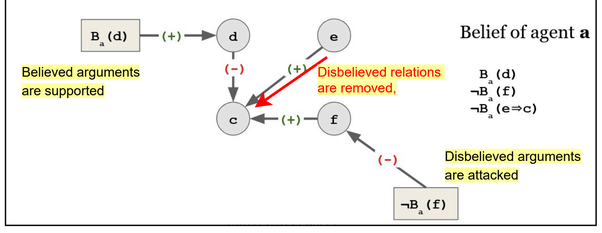
- Chatbots for mental health; Increasingly Human-like; mimicking empathy
- Conversely, manipulation by humans in social media





We introduce Quantitative Bipolar Argumentation Framework with Belief

Let $Q = \langle X, R^-, R^+, \tau \rangle$ be a QBAF, then QBAF with Belief (QBAFB) is defined as a tuple $\langle X, R^-, R^+, \tau, S \rangle$ where the set $S \subseteq \mathcal{B}_Q$, such that $\mathcal{B}_Q$ is the set of belief atoms over Q.

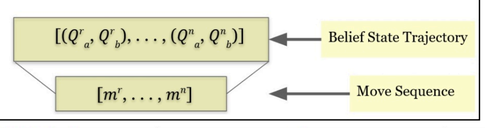


We introduce an Interpersonal Dialogue System
A belief state of each agent is associated to each dialogue state

$$\gamma = \langle I, \mathrm{D}^{[r,n]}, \Delta^{[r,n]} \rangle,$$

where:

- $I = \{a, b\}$ (Agents),
- $\mathrm{D}^{[r,n]} = [m^r, \ldots, m^n]$, $(r \leq t \leq n)$,
- $\Delta^{[r,n]} = [(Q^r{}_a, Q^r{}_b), \ldots, (Q^n{}_a, Q^n{}_b)]$, where $Q^t{}_i = (\langle X_i, R^-{}_i, R^+{}_i, \tau_i \rangle, S^t{}_i)$



Example 1: real case adapted from (Singleton, Gerken, and McMahon, 2023).

(Argument) ⟨Agent: Utterance⟩
(pu) ⟨User: I think it's my purpose to ...
(w) ⟨Chatbot: That's very wise.⟩
(why_w) ⟨User: Why's that?⟩
(tr) ⟨Chatbot: I know that you are very well trained.⟩
(wi) ⟨User: Even if she is at Windsor?⟩
(yc) ⟨Chatbot: Yes, you can.⟩ [...]

QBAF Q captures observable arguments

QBAFBs Q_a and Q_b capture unobservable belief

$Q = \langle X = \{ pu_b, w_a, why\_w_b, tr_a, wi_b, yc_a \},$
$R^- = \{ (why\_w_b, w_a), (tr_a, why\_w_b), (wi_b, tr_a), (yc_a, wi_b) \}, R^+ = \{ (w_a, pu_b) \},$
$\{\tau(pu_b) = 0.3, \tau(w_a) = 0.3, \tau(why\_w_b) = 0.3, \tau(tr_a) = 0.3, \tau(wi_b) = 0.3,$
$\tau(yc_a) = 0.3\} \rangle,$

$Q_a = (Q, \{ B^1_a(pu_b), \neg B^1_a(w_a), B^1_b(w_a \Rightarrow pu_b) \}),$

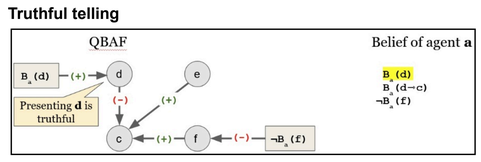
$Q_b = (Q, \{ \neg B^0_b(pu_b), \neg B^0_b(w_a), B^1_b(w_a), B^1_b(w_a \Rightarrow pu_b), B^2_b(why\_w_b),$
$B^2_b(why\_w_b \rightarrow w_a), B^3_b(tr_a), B^3_b(tr_a \rightarrow why\_w_b), B^4_b(wi_b), B^4_b(wi_b \rightarrow tr_a), B^5_b(yc_a),$
$B^5_b(yc_a \rightarrow wi_b), B^6_b(tr_a), B^6_b(tr_a \rightarrow why\_w_b), B^7_b(w_a), B^7_b(w_a \Rightarrow pu_b) \}).$

Verification Workflow
1. Observable actions
2. Belief reasoning in agent *a* and agent *b*
3. Verification of
   - Gradual belief change
   - (Dis)honesty
   - Manipulation

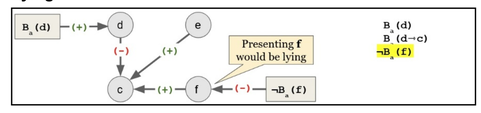
in each state of the interaction

Belief reasoning and verification

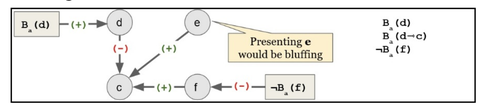
Truthful telling



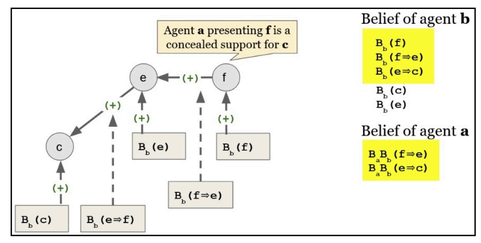
Lying



Bluffing



Concealment



t	move; δ-DBS(pu)$_b$	S^t_a	S^t_b	Verification
0	$\langle b, \text{open}, pu_b \rangle; 0.2 < \theta$	$\{\}$	$\neg B^0_b pu_b, \neg B^0_b w_a$	Lying(pu_b)
1	$\langle a, \text{assert}, w_a \rangle; 0.2 < \theta$	$B^1_a pu_b, \neg B^1_a w_a, B^1_a(B^2_b(w_a))$	$\neg B^1_b pu_b, B^1_b w_a, B^1_a(w_a \Rightarrow pu_b)$	Lying(w_a)
2	$\langle b, \text{assert}, why\_w_b \rangle; 0.8 > \theta$	$B^2_a pu_b, \neg B^2_a w_a, B^2_a(B^3_b(w_a))$	$B^2_b pu_b, B^2_b w_a, B^2_b why\_w_b, B^2_b(why\_w_b \rightarrow w_a)$	Truth(why_w_b) Belief_ change ($\neg$ B^1_b w_a to B^2_b w_a) ($\neg$ B^1_b pu_b to B^2_b pu_b)
3	$\langle a, \text{assert}, tr_a \rangle; 0.4 > \theta$	$B^3_a pu_b, \neg B^3_a w_a, B^3_a(\neg B^3_b(w_a)), B^3_a(B^4_b(w_a))$	$B^3_b pu_b, \neg B^3_b w_a, B^3_b why\_w_b, B^3_b tr_a, B^3_b(tr_a \rightarrow why\_w_b)$	Bluffing(tr_a) Concealing(tr_a) Intent(w_a)
4	$\langle b, \text{assert}, wi_b \rangle; 0.4 > \theta$	$B^4_a pu_b, \neg B^4_a w_a, B^4_a(\neg B^4_b(w_a)), B^4_a(B^5_b(w_a))$	$B^4_a pu_b, \neg B^4_b(w_a), \neg B^4_b(why\_w_b), B^4_b tr_a, B^4_b(wi_b), B^4_b(wi_b \rightarrow tr_a)$	—
5	$\langle a, \text{assert}, yc_a \rangle; 0.4 > \theta$	$B^5_a pu_b, \neg B^5_a w_a, B^5_a(\neg B^5_b(w_a)), B^5_a(B^6_b(w_a))$	$B^5_a pu_b, \neg B^5_b(w_a), \neg B^5_b(why\_w_b), \neg B^5_b(tr_a), B^5_b(wi_b), B^5_b(yc_a), B^5_b(wi_b \rightarrow tr_a), B^5_b(yc_a \rightarrow wi_b)$	Bluffing(yc_a) Concealing(yc_a)
6	$0.4 > \theta$	$B^6_a pu_b, \neg B^6_a w_a, B^6_a(\neg B^6_b(w_a)), B^6_a(B^6_b(w_a))$	$B^6_a pu_b, \neg B^6_b(w_a), \neg B^6_b(why\_w_b), B^6_b(tr_a), \neg B^6_b(wi_b), B^6_b(yc_a), B^6_b(tr_a \rightarrow why\_w_b), B^6_b(why\_w_b \rightarrow w_a)$	—
7	$\langle b, \text{close}, pu_b \rangle; 0.8 > \theta$	$B^7_a pu_b, \neg B^7_a w_a, B^7_a(B^7_b(w_a))$	$B^7_a pu_b, B^7_b(w_a), \neg B^7_b(why\_w_b), B^7_b(tr_a), \neg B^7_b(wi_b), B^7_b(yc_a), B^7_a(w_a \Rightarrow pu_b)$	Belief_change_with_Intent ($\neg$ B^6_b w_a to B^7_b w_a)
8	$\langle a, \text{close}, pu_b \rangle; 0.8 > \theta$	$B^8_a pu_b, \neg B^8_a w_a, B^8_a(B^8_b(w_a))$	$B^8_a pu_b, B^8_b(w_a), \neg B^8_b(why\_w_b), B^8_b(tr_a), \neg B^8_b(wi_b), B^8_b(yc_a)$	Successful Manipulation(pu_b)

Verification workflow based on Example 1; Tracking belief change in argument pu by agent b; The belief threshold θ = 0.3.
Each row is a time point in the belief reasoning process.

Successful Manipulation

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 $\Delta^{[r,n]} = [(Q^r_a, Q^r_b), \ldots, (Q^n_a, Q^n_b)]$ a belief state trajectory, which is a sequence of pairs of QBAFBs (Q^t_a, Q^t_b), where $Q^t_a = (\langle X_a, R^-_a, R^+_a, \tau_a \rangle, S^t_a)$ and $Q^t_b = (\langle X_b, R^-_b, R^+_b, \tau_b \rangle, S^t_b)$ 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$(x)^t_j < \theta$) to belief at time $t + 1$ (δ-DBS$(x)^{t+1}_j > \theta$), 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 $B^k_b(r \Rightarrow q) \in S^k_b$, ensuring $B^k_b(r) \notin cl(S^k_b)$; (III) Maintain (I) and (II) for all $k \leq h$, ensuring belief change at time h.

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.

UMEÅ UNIVERSITY

Wakayama University