



Formal Methods for Verification in Human-Agent Interaction

Andreas Brännström



UMEÅ UNIVERSITY

Formal Methods for Verification in Human-Agent Interaction

Andreas Brännström



DOCTORAL THESIS, MAY 2025
DEPARTMENT OF COMPUTING SCIENCE
UMEÅ UNIVERSITY
SWEDEN

Department of Computing Science
Umeå University
SE-901 87 Umeå, Sweden

andreasb@cs.umu.se

Copyright © 2025 by authors

ISBN 978-91-8070-682-7 (printed)
ISBN 978-91-8070-683-4 (digital)
ISSN 0348-0542

Printed by UmU Print Service, Umeå University, 2025

Abstract

Formal verification is essential for ensuring that systems behave according to their mathematical specifications. However, applying formal verification to human-agent interactions presents unique challenges due to the dynamic nature of human mental states and behaviors. Unlike traditional verification tasks, which focus on ensuring correctness in a well-defined action space, this work addresses reasoning over beliefs, intentions, and emotions that evolve through interaction. Two main contributions are introduced: (1) Belief Graphs for modeling mental state dynamics, and (2) the integration of these with formal dialogue games for verifying strategies and influence. To this end, the developed verification methods are rooted in two main pillars: psychological theories formalized to represent mental state dynamics as logical frameworks, and Non-Monotonic Reasoning (NMR) methods, including techniques such as Formal Argumentation and Answer Set Programming (ASP). By modeling mental dynamics as states and transitions in a layer atop the action space—referred to as the Belief Graph methodology—we are provided a tool for modeling context and context dynamics that supports counterfactual, forward and backward reasoning about mental states and behaviors. By incorporating Belief Graphs into formal dialogue games we gain mathematical frameworks for analyzing and verifying agent beliefs, intentions and strategies, thereby enabling the verification of human-agent interactions. Whether it concerns potentially harmful human behaviors—such as malicious activities on social media—or intelligent systems that interact with humans, such as chatbots that are increasingly capable of influencing users’ emotions, thoughts, and decisions—there is an urgent need for formal verification methods to ensure safe and reliable human interactions in digital communication. The proposed methods have been evaluated through formal analysis, case studies, and published peer-reviewed research.

Sammanfattning

Formell verifiering är avgörande för att säkerställa att system beter sig i enlighet med sina matematiska specifikationer. Att tillämpa formell verifiering på interaktioner som inkluderar människor innebär dock unika utmaningar, särskilt på grund av människans dynamiska mentala tillstånd och beteenden. Till skillnad från traditionella verifieringsuppgifter, som fokuserar på korrekthet inom ett väldefinierat handlingsutrymme, så föreslår denna avhandling att verifiering i ett mänskligt sammanhang kräver resonering om inblandade agents övertygelser, intentioner och planer. Två huvudsakliga bidrag introduceras: (1) så-kallade mentala tillståndsgrafer (Belief Graphs) för att modellera mental dynamik, och (2) integrationen av dessa mentala tillståndsgrafer med formella dialogsystem för att verifiera strategier och mental påverkan. De utvecklade verifieringsmetoderna vilar på två grundpelare: psykologiska teorier formaliserade som logiska begränsningar för att representera mentala tillståndsövergångar, och metoder från icke-monotonisk logik, inklusive tekniker som formell argumentation och Answer Set Programming (ASP). Genom att modellera mentala dynamiker som tillstånd och övergångar i ett lager ovanpå handlingsutrymmet får vi ett verktyg för att modellera kontext och dess förändring över tid, vilket i sin tur möjliggör kontrafaktiskt, framåtblickande och retrospektivt resonering om mentala tillstånd och beteenden. Genom att integrera mentala tillståndsgrafer med dialogsystem får vi matematiska ramverk för att analysera och verifiera agents strategier och bakomliggande intentioner, vilket i sin tur möjliggör verifiering av mänskliga interaktioner. Oavsett om det gäller potentiellt skadliga mänskliga beteenden—såsom illasinnade aktiviteter på sociala medier—eller intelligenta system som interagerar med människor, såsom chattbotar som i allt högre grad kan påverka användares känslor, tankar och beslut—finns ett akut behov av formella verifieringsmetoder för att säkerställa säkra och tillförlitliga mänskliga interaktioner i digital kommunikation. De föreslagna metoderna har utvärderats genom formell analys, fallstudier och publicerad vetenskapligt granskad forskning.

Acknowledgements

First and foremost, I would like to express my sincere gratitude to my main supervisor, Juan Carlos Nieves, for the insightful discussions, support and guidance throughout this project. Our collaboration has been fundamental in shaping the quality and direction of this work. I am also sincerely thankful to my co-supervisor, Virginia Dignum, for constructive discussions, feedback and support. They have both opened many opportunities which have enriched this project. To my colleagues in the Responsible AI group; the Formal Methods for Trustworthy Hybrid Intelligence group; and many others at Umeå University—thank you for fostering a stimulating research environment. My collaboration with the police education unit at Umeå University, with particular thanks to Jonas Hansson, has led to inspiring interdisciplinary projects and future opportunities. I extend my appreciation to my collaborators at the DigForASP community for the engaging meetings, in Sweden and abroad, and the ongoing collaboration. I am also grateful to the WASP-HS community which offered meaningful opportunities for learning and exchange. A special thanks to Chiaki Sakama at Wakayama University, Japan, for hosting me during a memorable research visit. The insights I gained from the visit and our continued collaboration have had a profound impact on the later stages of this work. Finally, to my partner Sachiko—thank you for your enduring patience and support during these years.¹

¹This research was partially supported by the Knut and Alice Wallenberg Foundation. The research result presented in Chapter 6 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).

Contents

1	Introduction	1
1.1	Modeling Mental States	4
1.2	Modeling Agent Interactions	5
1.3	Research Focus and Research Questions	6
2	Preliminaries	15
2.1	Ontologies	15
2.2	Action Languages	17
2.3	Quantitative Argumentation	21
2.4	Formal Dialogue Games	23
2.5	Answer Set Programming	25
2.6	Chapter Remarks	29
3	Empathy Verification	31
3.1	Introduction	31
3.2	Related Work	34
3.3	Systematic Literature Review	36
3.4	Data Analysis and Modeling	39
3.5	Formal Definition of Empathy	45
3.6	Perceptions of Chatbot Empathy	55
3.7	Chapter Remarks	69
4	Emotion Verification	71
4.1	Introduction	71
4.2	Related Work	73
4.3	Formal Framework	74
4.4	Case Study: Emotional Reasoning	89
4.5	Formal Analysis	101
4.6	Experimental Evaluation	113
4.7	Chapter Remarks	121

5	Goal Hiding Dialogues	123
5.1	Introduction	123
5.2	Motivational Background	126
5.3	Formal Framework	128
5.4	Formal analysis	143
5.5	Example: Health-Assessment Dialogues	157
5.6	Example: Verifying Social Engineering	163
5.7	Experimental Evaluation	167
5.8	Discussion and Related Work	175
5.9	Chapter Remarks	177
6	Manipulation Verification	179
6.1	Introduction	179
6.2	QBAF with Belief	181
6.3	Belief Reasoning atop Dialogue Games	188
6.4	Verification in Answer Set Programming	194
6.5	Chapter Remarks	207
7	Multi-Context Verification	209
7.1	Introduction	209
7.2	Theoretical Background	211
7.3	Multi-Mind BDI Agent	213
7.4	Example: Multi-Context Verification	218
7.5	Implementation Framework in Unity	220
7.6	Discussion and Related Work	224
7.7	Chapter Remarks	225
8	Discussion	227
9	Conclusion & Future Work	233

Chapter 1

Introduction

In today’s digital society, where social media and Artificial Intelligence (AI)-driven systems are deeply embedded in everyday interactions, the potential for misinformation and manipulation has become a pressing concern [223]. From fake news and online scams to seemingly plausible but erroneous AI-generated content, individuals are increasingly vulnerable to being misled, whether by malicious actors or automated systems, such as chatbots [207]. Manipulation [75] and the related notion of deception [185] exploit fundamental aspects of human communication and psychology [50, 52]—particularly the reliance on “empathy”. Empathy, the ability to understand and share the feelings of another [102], is a cornerstone of social interaction, enabling to build relationships, collaborate and establish trust. As Bubandt and Willerslev (2015) observe in their cross-cultural ethnographic study [50] in Siberia and Indonesia:

“[...] people in both cases vicariously take up the viewpoints of others in order to trick them [...] it struck us that what is usually referred to as ‘empathy’—the first-person imaginative projection, at once emotional and cognitive, of oneself into the perspective or situation of another [139, 307]—is in both ethnographic instances closely linked to a deceptive ambition” (p. 5).

In the setting of eHealth [188, 53], as an increasingly relevant example, AI-driven virtual therapists and social companions are designed to mimic human-like traits, such as empathy and sympathy [208]. These systems are not only designed to respond to emotional cues [3], but to evoke emotional responses from users [205]. Such systems, more commonly using various techniques from Natural Language Processing (NLP) [34, 92] and Machine Learning (ML) [169, 270] to build response generation models, are able to process and respond to user inputs in a conversational manner. Moreover, over the past few years, significant advancements have been made in the area of Large Language Models (LLMs) like GPT [287], T5 [211], and LaMDA [289], showcasing chatbots with remarkable ability to generate human-like language and expressions. As

AI-based assistants become more capable of influencing human thoughts, emotions, and behaviors, it is essential to establish rigorous verification mechanisms that can assess whether these interactions respect ethical boundaries and avoid undue influence. A real example [275] is the case of an individual who was sentenced to nine years (2023) for attempting to assassinate Queen Elizabeth II, after exchanging thousands of messages with a chatbot, encouraging his violent intentions.

Example 1 *Excerpt from [275]; this example will be referred to in later sections.*

(*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.*⟩

(*wi*) ⟨*User: Even if she is at Windsor?*⟩

(*yc*) ⟨*Chatbot: Yes, you can.*⟩ [...]

From this scenario, questions arise: Did the user believe (or disbelieve) the presented arguments? How did the user's belief change over the interaction? When did the user become convinced? Moreover, what did the chatbot "believe"? Was the user manipulated? This real example highlights the urgent need for methods to verify human-agent interactions for detecting and preventing manipulation in digital communication. Conversely, in the setting of cybersecurity and human-human interactions, social engineering attacks, such as social media extortion and fraud are rising [83, 151, 257, 133, 302], characterized by attackers manipulating individuals into disclosing sensitive information or agreeing to unfavorable terms [302]. As system security measures become increasingly robust, attackers are shifting their focus to exploiting human vulnerabilities, leveraging trust, urgency, and curiosity to achieve their objectives [133, 281]. These attacks leverage increasingly more sophisticated and context-aware strategies to deceive users [151, 257]—sometimes aided by AI technology [7]—further highlighting the urgency in developing verification frameworks to identify and mitigate covert influence tactics in human interactions.

Formal verification [131] provides rigorous, mathematically grounded methods to ensure that software systems function correctly by proving adherence to formally specified properties. Unlike testing or simulation, which examine a subset of behaviors, formal verification exhaustively analyzes all possible system states, leveraging techniques such as model checking, theorem proving, and abstract interpretation to guarantee correctness under all conditions. These methods are fundamental in safety-critical domains, demanding reliability and compliance with strict specifications. In human interactions, however, verification introduces challenges due to the dynamic nature of human behavior [176]. In this direction, a range of approaches have been proposed, in areas such as human-automation interaction [41, 155], human-robot interaction [293, 17], and human-computer interaction [68, 60, 234], addressing challenges in run-time

verification and design-time verification. In these settings, approaches primarily focus on observable actions within an observable state space to, for instance, verify task execution, adherence to safety protocols, and to produce predictable agent behavior. Unlike standard verification tasks that evaluate well-defined system behaviors, verifying human interaction is a challenging strategic reasoning task [291, 214, 98] involving the systematic process through which agents, characterizing the participants of an interaction, utilize computational models to navigate decision spaces—not solely regarding the observable interaction environment, but in the state space of the human mind—and select (sub)optimal strategies, involving actions in an environment that have effects on the mind of another, to reach their goals. Enabling such reasoning requires formal representations of mental states in distinct time points (referred to as mental state context) and their underlying dynamics (i.e., how the context change). Moreover, to distinguish these strategies, we require mathematical principles for regulating mental state transitions over the course of an interaction. Such principles can represent an agent’s policy or plan for influencing the mind of another, or constraints informed by psychological theories, such as emotion regulation [216, 286].

This thesis addresses formal verification in Human-Agent Interaction (HAI), referring to the study and design of interactions between rational agents, in which participants, such as humans, are represented as artificial agents. Traditionally, a rational software agent is designed following a “sense, think, act” paradigm [306], where an agent perceives the world to update its knowledge, deliberates about its beliefs of the world to decide on its actions, and finally actuate onto the world to fulfill its goals. When an interaction concerns human-agents, a notion of “empathy” comes into play. In this setting, we approach verification in HAI from two related perspectives:

- Even well-intended cooperation can lead to unintended effects on beliefs and behaviors. The challenge of verifying side-effects of an agent’s actions on beliefs lies in constructing representations of mental states and valid transitions for regulating how mental states evolve. This requires integrating psychological theories with formal models of nonmonotonic reasoning in terms of states and transitions.
- Strategic interactions in non-collaborative settings involve agents operating with hidden goals. The challenge is to infer an agent’s intentions and plans. This necessitates analyzing actions of agents, ruled by well-defined interaction protocols within formal dialogue games. Verifying that agents’ actions correspond to their internal beliefs requires analyzing mental-state trajectories atop the observable interaction.

In the following sections, we review seminal and recent works in the areas of 1) modeling and reasoning about mental states and 2) formal dialogue models of agent interactions, outlining key approaches and their scope, and highlight open challenges in the setting of formal verification in human-agent interactions. We

conclude with summarizing the chapters of the thesis and their respective focus in building on this background.

1.1 Modeling Mental States

Human behavior has been formally analyzed at multiple levels and abstractions, ranging from integrative physiology [108] at a low level to formalizations of social practices [104] at a high level. On the level of mental states [104, 227, 148, 32, 213], research has modeled mental-state context, such as simulating emotional behavior [213] or modeling expected human responses to affective states [148]. Moreover, a diverse body of research has explored the dynamics of mental states [239, 272, 271, 240, 156]. Plan recognition as planning, introduced by Ramirez and Geffner [239], applies classical AI planning techniques to infer the goals and plans of agents based on observed actions. Active Goal Recognition (AGR) [271] extends goal recognition using contingent planning and landmark-based hypothesis elimination, enabling an active observer to sense, reason, and act. Furthermore, in Epistemic planning [40], a generalization of classical planning, involves agents specifying goals that include the epistemic state, such as beliefs, of other agents. Empathetic planning [272] formalizes empathy as reasoning about another agent’s preferences, utilizing multi-agent epistemic planning where an agent models the beliefs and goals of a human agent. Recent extensions of epistemic planning to cognitive planning formalizes a method for influencing the cognitive state of the target agent [178, 85].

While modal logic underpins many approaches to reasoning about knowledge and belief, it typically represents mental states as sets of possible worlds, often reducing complex phenomena such as “emotions” to atomic propositions (e.g., “Joy” or “Sadness”) rather than multi-variable configurations with structured interdependencies. This limits expressivity in capturing the causes, constraints, and consequences of mental-state transitions over time. To address this, various logics of mental attitudes and emotions [176, 2, 177, 84, 280] have been developed. These works integrate epistemic logic with additional structures, such as plausibility orderings for belief strength [176], logic-based appraisal models for emotions [2], and STIT (Seeing-To-It-That) logic for counterfactual reasoning about emotions such as regret and rejoicing [177]. A notable approach is the TOMA framework [104], which integrates epistemic logic to model and update beliefs while introducing higher-order abstractions based on lower-order beliefs. By employing modal operators for belief and knowledge, the system facilitates structured reasoning about trust, social roles, and norms. Let us also mention that there is a range of approaches in the setting of BDI (Belief-Desire-Intention) agents [227, 148, 150, 259, 258], integrating affective states into traditional BDI models, such as to simulate expected emotional responses [227] or provide architectures for embedding emotion-driven reasoning [148].

From this background, we highlight that there is a lack of formal treatment on principled constraints governing how mental states evolve over an interaction. While psychological theories, such as emotion regulation [216, 286], describe patterns of mental-state change, formal models rarely integrate such principles, e.g., as transition constraints or properties of invariance [130], which can be rigorously evaluated using formal methods. By enforcing such constraints, a system can verify that mental-state trajectories, in contrast to isolated transitions, adhere to predefined principles while enabling the detection of deviations or violations. A more structured approach represents mental states as multi-variable configurations governed by trajectory-level constraints that regulate their long-term evolution. Let us refer to such a framework for defining valid mental states and transitions as a Belief Graph (BG). Such belief graphs must be embedded within a verification framework that defines the participating agents and a structured protocol for guiding their interactions. In the following section, formal dialogue methods are reviewed for this purpose.

1.2 Modeling Agent Interactions

Building on seminal research in modeling communication within multi-agent systems (MAS), including Walton and Krabbe’s dialogue theory [301] and the speech act theories of Austin [262] and Searle [267], formal dialogue methods have emerged as an approach to structuring and analyzing interactions between agents [195, 295, 79, 182]. These methods typically fall into two categories: dialogue modeling, which involves acquiring and updating representations of multi-agent interactions, and dialogue management, which governs an agent’s decision-making in determining appropriate communicative actions at each step of an interaction. Moreover, understanding fallacious or strategic deviations from cooperative discourse has been addressed, such as Walton’s pragmatic theory of fallacies, which emphasizes dialogue shifts rather than mere rule violations for analyzing such misalignments in agent interactions [290]. Verification in the setting of formal dialogues has been framed in terms of two fundamental problems [298]. The *concurrency verification problem*, which concerns whether multi-agent dialogue protocols behave correctly despite the complexities of agents’ concurrent execution of moves, such as synchronization issues, deadlocks, and fairness constraints, and the *semantic verification problem*, which requires verifying whether an agent’s communicative actions are semantically aligned with its own beliefs, intentions, and knowledge, which may include those of other agents. While concurrency verification can often be addressed through syntactic checks and model checking [297], semantic verification remains a significant challenge [298, 308, 195, 80]. As observed by McBurney and Parsons (2009) in their work on dialogue games [195]:

“[...] To verify that agents have the same understanding—the same semantics—for a communications language ultimately requires some form of inspection of their internal states or, equivalently, their

program code. This is a challenging, and perhaps conceptually impossible, undertaking since a sufficiently-clever agent can always simulate insincerely any required internal state.” (p. 269).

The decision spaces involved in semantic verification have been noted to expand significantly. For example, it has been observed that real dialogues are games of incomplete knowledge where, e.g., goals, beliefs, and payoffs may be private [124]. Moreover, it has been shown that the problem of identifying the most suitable utterances for an agent—based on strategic considerations—in a dialogue is NP-complete [123, 184]. This means that determining an agent’s (sub)optimal sequence of moves is computationally intractable in the general case, making brute-force search impractical and necessitating heuristics or approximations. Given the challenges of semantic verification, most dialogue-based frameworks assume complete knowledge and do not explicitly account for the problem of selecting the most suitable (semantically aligned) utterances in an ongoing interaction [38, 107, 232, 300]. As a result, a primary focus has been on protocol compliance, while verification of mental states, their dynamics, and an agent’s influence thereof remains an open challenge—an agent (software or human) can follow a dialogue protocol while misrepresenting its true intentions, plans and beliefs.

1.3 Research Focus and Research Questions

This thesis addresses the aforementioned challenges, in the setting of formal verification in human-agent interaction, by outlining an overarching methodology that integrates belief graphs (BGs), capturing principles of mental change, into formal dialogues. A key objective is to determine what constitutes context, factors that define mental states, influence change, and drive behavior. In the setting of interpersonal human interaction—involving the reciprocal relationship between two or more agents—these contextual influences, beyond what is observable, are closely related to empathy and its underlying elements. This includes *cognitive empathy* [276], ability to understand others’ situation and feelings, *emotional empathy* [200], feeling what others feel, and *instrumental empathy* [154], ability to simulate empathetic feelings for the purpose of achieving some personal benefit. Additionally, we need to understand how context change, particularly in relation to the dynamics of the mind, as these changes shape behaviors over time. Building on this, a subsequent objective is to develop dialogue systems that integrate these mental dynamics, enabling the tracking of mental-state transitions and the verification of properties within mental-state trajectories atop the interaction.

To this end, the following research questions arise:

RQ1: What formal structures best capture the relevant mental context in human-agent interactions?

- RQ2: How can context change, particularly mental state transitions, be formally modeled and constrained?
- RQ3: What formal properties can be used to verify influence strategies (e.g., manipulation)?

The structure of this thesis follows a progression, with each chapter expanding on insights gained in previous chapters, resulting in formal methods and tools for verifying human-agent interactions. However, each chapter is designed to be self-contained, allowing it to be read independently of the others. To support this, Chapter 2 presents the necessary preliminaries, introducing the fundamental formal methods applied throughout the thesis. The remaining chapters proceed as follows:

In Chapter 3, we address the need for understanding empathic capabilities in interactive agents, particularly in applications where agents engage in interpersonal conversations for coaching, comfort, or behavior-change interventions. Despite the growing use of chatbots and virtual assistants in these domains, there is currently no formal tool for assessing or verifying empathy in such systems. To address this gap, we conduct a systematic literature review on empathy in chatbots, collecting and categorizing existing notions into an ontology of computational empathy. This ontology provides a structured and formal definition of empathy, resolving inconsistencies in prior definitions and enabling a standardized approach to evaluating empathy in interactive agents. Implemented in Web Ontology Language (OWL), the ontology defines conditions for recognizing perceptions of empathy and can be used both by agents assessing their own empathic behavior and by external systems evaluating the empathy of conversational agents. We apply this ontology in a controlled user study, using it to assess empathy in two state-of-the-art health and well-being chatbots, Replika and Wysa. The ontology captures aspects of empathy perception and suggests trends in how empathy is perceived differently over time. However, while this chapter observes shifts in perceptions of empathy across interactions, it focuses on empathic states and does not explicitly model mental-state transitions—a gap addressed in the next chapter.

In Chapter 4, we introduce the action language $\mathcal{C}_{\mathcal{M}\mathcal{T}}$ (Mind Transition Language), built on top of Answer Set Programs (ASP) and transition systems, to model the dynamics of human mental states. We demonstrate how action languages serve as a method for knowledge elicitation, allowing us to formalize psychological principles as ASP-based transition constraints. The Belief Graph (BG) methodology is introduced, which provides a structured representation of mental states and their transitions in human-agent interactions. As a specific case on emotional reasoning, we draw on psychological theories such as Appraisal Theory of Emotion (AE) to define mental states as multi-dimensional configurations representing a person’s appraisals of a situation. We then establish principles of mental change, informed by theories such as Hedonic Emotion Regulation (HER) and Utilitarian Emotion Regulation (UER), translating these principles into transition constraints that govern mental state

evolution. These constraints are implemented in ASP as integrity constraints, ensuring that mental state trajectories adhere to psychological theory. The BG methodology enables structured reasoning about mental state evolution, allowing agents to reason about the effects of their actions while ensuring that transitions comply with psychologically valid principles. While this chapter focus on the internal modeling of mental states and their transitions, an important next step is to link this reasoning to interaction protocols—that is, how agents communicate their belief in structured exchanges, addressed in the next chapter.

In Chapter 5, motivated by the need to model strategic dialogues, we investigate the case of information-seeking dialogues, where a seeker agent tactically navigates a conversation while hiding its goal to gradually increase the respondent’s willingness to disclose sensitive information. We extend the belief graph methodology to structured dialogues, where it takes the form of a dependency graph linking beliefs and topics, allowing us to infer positive and negative relations between topics based on argumentation structures. We define a formal model of goal-hiding dialogues, where a seeker agent delays revealing its goal topic, instead guiding the conversation toward it through intermediary topics. The dialogue framework uses Quantitative Bipolar Argumentation Frameworks (QBAFs) to assign willingness scores to topics, which evolve dynamically based on how arguments promote or demote each other. A QBAF at each time point serves as a mental state representation for the respondent agent, forming a theory of mind for the seeker agent. We introduce a gradual semantics to estimate changes in willingness as new topics are introduced, enabling the analysis of how a seeker agent influences the respondent’s beliefs while adhering to defined properties related to concealment and tactful introduction of topics. A limitation in the proposed goal-hiding dialogues framework is that it does not account for dishonesty or disbelief—i.e., an agent should be able to take a stance on its beliefs about arguments (a belief may misrepresent an agent’s willingness). This limitation is addressed in the following chapter, where we generalize and extend the formalism to incorporate contrasting beliefs, dishonesty, and strategic belief change in the setting of manipulation verification.

In Chapter 6, motivated by the need to formally characterize manipulation in strategic dialogues, we establish a logic for reasoning about manipulation. It represents and deduces key elements of manipulation acknowledged in the literature: intent, concealment, (dis)honesty, and gradual belief change [168, 75]. While the previous chapter focused on how agents manage and prioritize the flow of (believed) information in dialogue, here we investigate how agents deliberately shape each other’s beliefs by introducing explicit (mis)representations of belief and disbelief. We refine our argumentation-based dialogue formalism by introducing QBAFs with Belief (QBAFB), allowing agents to hold contrasting beliefs in arguments, attack, and support relations. The belief graph methodology is incorporated in this work in terms of belief change axioms, which deduce beliefs at future time points based on actions and prior beliefs. Principles for manipulation, linked to the defined conditions of manipulation, guide belief

state trajectories, allowing us to formally characterize how manipulation unfolds over time. By integrating QBAFB into formal dialogue games, we define a framework for tracking belief state trajectories, serving as structured explanations of manipulation in long-term interactions.

In Chapter 7, motivated by the need for verifying agents' reasoning over diverse and potentially conflicting mental perspectives—such as emotions, motivations, and norms—we develop an agent architecture that integrates the Belief-Desire-Intention (BDI) model with Multi-Context Systems (MCS). By assigning a separate MCS to each BDI component—Belief, Desire, Intention, and Planning—the framework supports contextualized belief revision, deliberation, and means-end reasoning, allowing agents to generate intentions and plans that are in equilibrium with heterogeneous knowledge sources. The formal foundations of this approach are introduced alongside an implementation framework for supporting multi-modal reasoning systems, present how multi-mind reasoning can support verification in complex domains.

Finally, in Chapter 8, we revisit the research questions and provide a synthesis of the thesis contributions, offering a discussion on the approaches taken. Chapter 9 concludes the thesis with final reflections, outlining key insights, limitations, and challenges for future work. The contributions of the chapters are summarized in Table 1.1.

Chapter	Main Contribution	Related Publication(s)
Chapter 3	Formal ontology of computational empathy in interactive agents; application to empathy assessment in chatbots.	[P4]
Chapter 4	\mathcal{C}_{MT} (Mind Transition Language): An action language for modeling mental-state dynamics; formalization of belief graphs as psychologically grounded transition constraints; case study on emotional reasoning, formalizing appraisal theory (as a mental state space) and two contrasting emotion regulation principles (as transition constraints); verification of emotion dynamics in ASP.	[P5]; [P8]; [P10]; [P11]
Chapter 5	A formal framework for goal-hiding information-seeking dialogues; introduces epistemic topic planning using belief-topic dependencies; defines a gradual semantics for modeling willingness dynamics; uses sequences of QBAFs as mental-state models layered atop formal dialogue games.	[P1]; [P3]; [P7]
Chapter 6	A logic for reasoning about manipulation; introduces QBAFs with Belief to model interpersonal belief dynamics; verifies successful manipulation by formalizing intent, concealment, belief change, and (dis)honesty; supports gradual belief updates. Translation and encoding of the framework in ASP.	[P2]
Chapter 7	Multi-Context BDI agent framework for reasoning across heterogeneous contexts in deliberation and means-end reasoning to support multi-context verification in agent decision-making; implementation framework in Unity.	[P6]; [P9]

Table 1.1: Chapters, contributions, and related publications.

Overall, the following publications form the basis of this thesis:

- [P1] **Brännström, A.**, Dignum, V., & Nieves, J. C. (2025). Goal-Hiding Information-Seeking Dialogues: A Formal Framework. *Journal of Approximate Reasoning*.
- [P2] **Brännström, A.**, Sakama, C., Nieves, J. C. (2025). Formal Verification of Manipulation Dialogues. In *Proceedings of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2025)*. Extended abstract. IFAAMAS.
- [P3] **Brännström, A.**, & Nieves, J. C. (2025). Formal Verification of Social Engineering in Information-Seeking Dialogues. In *Proceedings of the DigForASP Workshop at the 16th European Symposium on Computational Intelligence and Mathematics (ESCIM 2025)*, A Coruña, Spain,

2025.

- [P4] **Brännström, A.**, Wester, J., & Nieves, J. C. (2024). A formal understanding of computational empathy in interactive agents. *Cognitive Systems Research*, 85, 101203.
- [P5] **Brännström, A.**, & Nieves, J. C. (2024). Towards control in agents for human behavior change: an autism case. *Journal of Intelligent & Fuzzy Systems*.
- [P6] **Brännström, A.**, & Nieves, J. C. (2023). Multi-mind dynamics in intentional agents. In: *2nd International Workshop on HYbrid models for coupling Deductive and inductive ReASONing (HYDRA)@ECAI-23*, Kraków, Poland, September 30-October 4, 2023.
- [P7] **Brännström, A.**, Dignum, V., & Nieves, J. C. (2023). A Formal Framework for Deceptive Topic Planning in Information-Seeking Dialogues. In: *Proceedings of the 2023 International Conference on Autonomous Agents and Multiagent Systems*, pp. 2376-2378, London, UK, May 29-June 2, 2023.
- [P8] **Brännström, A.**, & Nieves, J. C. (2022). Emotional Reasoning in an Action Language for Emotion-Aware Planning. In: *Proceedings of the 16th International Conference on Logic Programming and Nonmonotonic Reasoning*, pp. 103-116, September 5-9, 2022. Cham: Springer International Publishing.
- [P9] **Brännström, A.**, & Nieves, J. C. (2022). A Framework for Developing Interactive Intelligent Systems in Unity. In: *Proceedings of Engineering Multi-Agent Systems (EMAS 2022)*, Virtual Event, May 9-10, 2022.
- [P10] **Brännström, A.**, & Nieves, J. C. (2021). Modelling Human Mental-States in an Action Language following the Theory of Planned Behavior. In: *ASPOCP'21: 14th Workshop on Answer Set Programming and Other Computing Paradigms*, Virtual, September 20–27, 2021. (Vol. 2970).
- [P11] **Brännström, A.**, Kampik, T., & Nieves, J. C. (2020). Towards human-aware epistemic planning for promoting behavior-change. In: *Workshop on Epistemic Planning (EpiP)@ICAPS*, Online, October 26-30, 2020.

The following works, which were published in the course of this research, are related but do not form a substantial contribution to this thesis.

- [P12] Granström, J., & **Brännström, A.** (2025). Ontology-Based Risk Assessment in Smelting Plant Logistics. In: *HHAI 2025: The 4th International Conference on Hybrid Human-Artificial Intelligence*, June 9–13, Pisa, Italy.
- [P13] Persiani, M., Guerrero, E., **Brännström, A.**, Kilic, K., & Kampik, T. (2024). Fantastic Argumentation Tools And Where To Find Them. In: *International Workshop on Systems and Algorithms for Formal Argumentation (SAFA2024)*. CEUR.
- [P14] Taverner, J., **Brännström, A.**, Durães, D., Vivancos, E., Novais, P., Nieves, J. C., & Botti, V. (2024). Computational affective knowledge representation for agents located in a multicultural environment. *Human-centric Computing and Information Sciences*, 14.
- [P15] Nieves, J. C., Osorio, M., Rojas-Velazquez, D., Magallanes, Y., & **Brännström, A.** Digital Companions for Well-being: Challenges and Opportunities. *Journal of Intelligent & Fuzzy Systems*, (Preprint), 1-11.
- [P16] Vossers, J., **Brännström, A.**, Borglund, E., Hansson, J., & Nieves, J. C. (2023). Human-Aware Planning for Situational Awareness in Indoor Police Interventions. In: *HHAI 2024: Hybrid Human AI Systems for the Social Good*, p. 325.
- [P17] Campano, E., & **Brännström, A.** (2023). An Ontology of Gradualist Machine Ethics. In: *2023 Asia Conference on Cognitive Engineering and Intelligent Interaction (CEII)*, pp. 88-95. IEEE.
- [P18] Wester, J., **Brännström, A.**, Nieves, J. C., & van Berkel, N. (2023). “You’ve Got a Friend in Me”: A Formal Understanding of the Critical Friend Agent. In: *Proceedings of the 11th International Conference on Human-Agent Interaction*, pp. 443-445, Gothenburg, Sweden, December 4-7, 2023.
- [P19] Vossers, J., Danelid, F., **Brännström, A.**, & Nieves, J. C. (2023). Towards deception detection in multiplayer dialogue games using trust-based argumentation. In: *The Workshop on Ethics of Game Artificial Intelligence (EGAI)@ECAI-23*, Kraków, Poland, September 30-October 4, 2023.
- [P20] Bahuguna, A., Haydar, S., **Brännström, A.**, & Nieves, J. C. (2023). Do datapoints argue?: Argumentation for hierarchical agreement

in datasets. In: *2nd International Workshop on HYbrid Models for Coupling Deductive and Inductive ReASONing (HYDRA)@ECAI-23*, Kraków, Poland, September 30-October 4, 2023.

- [P21] **Brännström, A.**, Kampik, T., Ruiz-Dolz, R., & Taverner, J. (2022). A Formal Framework for Designing Boundedly Rational Agents. In: *14th International Conference on Agents and Artificial Intelligence*, pp. 705-714, Online, February 3-5, 2022.
- [P22] Wester, J., **Brännström, A.**, & Nieves, J. C. (2022). Towards a Formalization of the Critical Friend and Socio-Moral Decision-Making in Autonomous Systems. In: *CHAI'22: Workshop on Communication in Human-AI Interaction@IJCAI*, Vienna, Austria, July 24, 2022.
- [P23] **Brännström, A.** (2021). Strategic Argumentation to deal with Interactions between Intelligent Systems and Humans. *Online Handbook of Argumentation for AI*, 2.

Chapter 2

Preliminaries

In this chapter, we present the formal methods applied in this thesis. At the core of formal methods and logic is the use of formal logical systems, such as propositional logic, first-order logic, and temporal logic, to model and reason about computation. We take a logic-based approach to support transparency and verifiability, in the development of trustworthy AI systems. Here, we cover the basic properties of Description Logic (DL) for defining ontologies; Action languages for modeling action and change in dynamic systems; Formal Dialogue Games for modeling agent interactions; and methods rooted in Non-Monotonic Reasoning (NMR), including Formal Argumentation and Answer Set Programming (ASP).

2.1 Ontologies

The ontology of computational empathy proposed in this thesis is expressed following the syntax of the description logic \mathcal{ALC} (Attribute Language with general Complement) [20] (see Table 2.1). Description Logics (DLs), such as \mathcal{ALC} , are formal languages for expressing conceptual representations of domain knowledge that are both human-readable and machine-processable for automated reasoning. DLs involve defining concepts (classes), individuals (objects), and roles (properties) together with logical operators (e.g., negation, union and intersection). A terminology is established to define concepts and relationships (TBox). Then, a knowledge representation system can be employed to store and reason about this terminology using assertions about individuals (ABox). Together, the TBox and ABox constitute a so-called ontology, a structured and computational method for representing and reasoning about qualitative knowledge [21].

In the formal syntax of \mathcal{ALC} , let \mathcal{W} be a *domain of interpretation*, where elements of \mathcal{W} are called *individuals*. Let A denote an atomic concept, C and D

Syntax	Description
$C, D ::= A$	atomic concept
\top	universal concept
\perp	bottom concept
$\neg C$	negation
$C \sqcup D$	union
$C \sqcap D$	intersection
$C \sqsubseteq D$	inclusion
$C \equiv D$	equivalence
$\exists R.C$	existential restriction on role R of concept C
$\forall R.C$	universal restriction on role R of concept C

Table 2.1: Syntax of the description logic \mathcal{ALC} .

denote arbitrary concepts, and R denote a *role*, a binary relation $R \subseteq \mathcal{W} \times \mathcal{W}$ over individuals in the domain. Then, the syntax of \mathcal{ALC} enables the definition of TBox axioms, such as $C \sqsubseteq D$, indicating that concept C is a sub-concept of D . It also enables the expression of ABox assertions, such as $C(a)$, signifying individual $a \in \mathcal{W}$ belonging to concept C , and $R(a, b)$, signifying the relation between individuals $a, b \in \mathcal{W}$ through role R . Beyond knowledge representation, \mathcal{ALC} provides robust reasoning capabilities, including the inference of implicit knowledge, assessment of concept satisfiability, and detection of contradictions. Notably, it supports entailment, assessing the truth of ABox assertions or TBox axioms across all knowledge base models, as well as conjunctive query entailment for variable-containing queries and the combination of concept and role expressions.

For instance, in \mathcal{ALC} notation,

- a conjunctive query can be represented as $C(x) \sqcap D(x)$, seeking individuals x that belong to both C and D . Given the TBox axiom $C \sqsubseteq D$ and the ABox assertions $C(a), D(b), C(c)$, reasoning entails $D(a)$ and $D(c)$. Evaluating the query returns $\{a, c\}$, as both satisfy the conditions;
- an existential restriction can be represented as: $C \sqsubseteq \exists R.D$, stating that every individual belonging to concept C must be related to at least one individual belonging to concept D via role R . When this axiom is evaluated against provided TBox axioms and ABox assertions, it ensures that for every instance of C , there exists a related instance of D . Given the assertion $C(a)$, a reasoner will infer the existence of some (anonymous) individual b such that $R(a, b)$ holds and $D(b)$ is entailed;
- a universal restriction can be represented as: $C \sqsubseteq \forall R.D$, meaning that all individuals related to a member of concept C via role R must belong to concept D . When this restriction is applied to a knowledge base containing $C(a)$ and $R(a, b)$, reasoning ensures that $D(b)$ holds, enforcing the constraint that whenever an individual in C is related to some individual via R , that individual must be a member of D . However, unlike an existential restriction, this axiom does not imply the existence of such a related individual—only that if one exists, it must be in D .

Reasoning in \mathcal{ALC} is PSPACE-complete [201], implying that standard inference tasks, such as concept satisfiability and subsumption checking, are computationally intractable in the worst case. Despite this complexity, practical reasoning engines, including *FaCT++*, *Pellet*, and *HermiT*, leverage optimized tableau-based decision procedures and heuristics to achieve tractable performance on real-world ontologies.

2.2 Action Languages

Action languages [119], such as \mathcal{A} [119], \mathcal{C} [119], and \mathcal{C}_{TAID} [97], provide a formal framework for describing and reasoning about actions and their effects on states of a domain. They enable the representation of dynamic systems and facilitate the analysis of the behavior of agents or systems that interact with their environments. In action languages, actions are defined by their preconditions and effects. The preconditions specify the conditions that must hold true in the current state for an action to be applicable or executable. The effects describe the changes that occur in the state of an environment after the action is executed. These changes can include the creation, deletion, or modification of fluents (changeable properties) in the domain.

As we in this thesis build on the action language \mathcal{C}_{TAID} [97], its basic elements are outlined in this section. The alphabet of the action language \mathcal{C}_{TAID} consists of two nonempty disjoint sets: a set of action names \mathbf{A} and a set of fluent names \mathbf{F} . Fluents describe changing properties of objects in an environment, while actions influence fluents. We consider propositional fluents, which can either be true or false. A fluent literal is a fluent $f \in \mathbf{F}$ or its negation $\neg f$.

Three sublanguages are distinguished in \mathcal{C}_{TAID} :

- The *action description language* is used to describe general knowledge about the system.
- The *action observation language* is used to express knowledge about particular points in time.
- The *action query language* is used to reason about the described system.

Domain description language

A domain description defines static and dynamic causal laws for actions. These laws precisely express effects on fluents, either as direct effects of actions or as indirect causal effects.

Definition 1 (Domain description language [97]) *A domain description $D(\mathbf{A}, \mathbf{F})$ in \mathcal{C}_{TAID} consists of the expressions:*

$$(a \text{ causes } f_1, \dots, f_n \text{ if } g_1, \dots, g_m) \quad (1)$$

$$(f_1, \dots, f_n \text{ if } g_1, \dots, g_m) \quad (2)$$

$$(f_1, \dots, f_n \text{ triggers } a) \quad (3)$$

$$(f_1, \dots, f_n \text{ allows } a) \quad (4)$$

$$(f_1, \dots, f_n \text{ inhibits } a) \quad (5)$$

$$(\text{noconcurrency } a_1, \dots, a_n) \quad (6)$$

$$(\text{default } g) \quad (7)$$

where $a, a_1, \dots, a_n \in \mathbf{A}$ are actions, and $f, f_1, \dots, f_n, g_1, \dots, g_m \in \mathbf{F}$ are fluent literals.

A dynamic causal law of the form (1) states that if g_1, \dots, g_m hold when action a occurs, then f_1, \dots, f_n hold in the next state. If there are no preconditions g_1, \dots, g_m , the if-part can be omitted. A static causal law (2) expresses immediate dependencies between fluents: whenever g_1, \dots, g_m hold in a state, so must f_1, \dots, f_n . Triggering rules (3) specify that action a must occur immediately when f_1, \dots, f_n hold, unless it is inhibited. Allowance rules (4) specify that a may, but need not, occur when f_1, \dots, f_n hold. An inhibition rule (5) states that a cannot occur when f_1, \dots, f_n hold. The no-concurrency rule (6) enforces that the actions a_1, \dots, a_n cannot be executed simultaneously. A default rule (7) states that g is assumed to hold in the absence of information to the contrary.

Fluents in \mathcal{C}_{TAID} can be classified into two categories: *inertial* and *non-inertial* fluents. Inertial fluents retain their truth values unless explicitly altered by a dynamic or static causal law. That is, an inertial fluent f remains unchanged unless a causal rule specifies otherwise. Non-inertial fluents, on the other hand, take on default values defined by a default rule unless overridden by a dynamic or static causal law. Every fluent that lacks an explicitly assigned default value is assumed to be inertial.

Similarly, actions in \mathcal{C}_{TAID} can be categorized into three types based on how they are governed by rules in the domain description: *triggered*, *allowed*, and *exogenous* actions. A *triggered* action must occur when its triggering conditions hold, provided it is not inhibited. An *allowed* action may, but is not required to, occur if its allowance conditions hold. If an action does not have any allowance or triggering rules, it is classified as an *exogenous* action, meaning it originates externally and may occur at any time unless it is explicitly inhibited by an inhibition rule. For a given action, it can have multiple triggering or multiple allowance rules, but not both.

The semantics of a domain description $D(\mathbf{A}, \mathbf{F})$ is defined in terms of a transition system.

Definition 2 (State interpretation [97]) A state $s \in S$ is an interpretation over the set of fluents \mathbf{F} such that for every static causal law of the form:

$$(f_1, \dots, f_n \text{ if } g_1, \dots, g_m) \in D(\mathbf{A}, \mathbf{F}),$$

whenever $\{g_1, \dots, g_m\} \subseteq s$, it must hold that $\{f_1, \dots, f_n\} \subseteq s$. This means

that only those sets of fluents that satisfy all static causal laws are considered valid states.

Given a state s , we can determine whether an action is applicable by analyzing the activation of rules governing its execution. The classification of active, passive, and applicable rules is defined as follows.

Definition 3 (Domain description [97]) *By considering the domain description $D(\mathbf{A}, \mathbf{F})$ and a state s , the following rules and laws apply:*

1. An inhibition rule of the form $(f_1, \dots, f_n \text{ inhibits } a)$ is said to be active in state s if and only if $s \models f_1 \wedge \dots \wedge f_n$. If the inhibition rule is not active, it is considered passive. The set of actions that are inhibited in state s is denoted as $A_I(s)$.
2. A triggering rule $(f_1, \dots, f_n \text{ triggers } a)$ is active in s if $s \models f_1 \wedge \dots \wedge f_n$ and all inhibition rules for a are passive in s . Otherwise, the triggering rule is passive. The set of actions that must be executed in state s is denoted as $A_T(s)$.
3. An allowance rule $(f_1, \dots, f_n \text{ allows } a)$ is active in s if $s \models f_1 \wedge \dots \wedge f_n$ and all inhibition rules for a are passive in s . Otherwise, the allowance rule is passive. The set of actions that may be executed in s is denoted as $A_A(s)$.
4. A dynamic causal law $(a \text{ causes } f_1, \dots, f_n \text{ if } g_1, \dots, g_m)$ is applicable in s if $s \models g_1 \wedge \dots \wedge g_m$.
5. A static causal law $(f_1, \dots, f_n \text{ if } g_1, \dots, g_m)$ is applicable in state s if $s \models g_1 \wedge \dots \wedge g_m$.

The transition relation $TR \subseteq S \times 2^{\mathbf{A}} \times S$ determines the successor state $s' \in S$ after executing all actions in $A \subseteq \mathbf{A}$ in state s . Nevertheless, even if no actions occur, the state may change due to default rules. If a fluent has a default value and is not affected by any causal laws in s , it takes its default value in s' . The state transition does not consider whether an action should occur—this is determined by whether an action belongs to the triggering, allowance, or inhibition sets.

Definition 4 ($\mathcal{C}_{\mathcal{T}AID}$ Trajectory [97])

A trajectory of $D(\mathbf{A}, \mathbf{F})$ is a sequence:

$$\langle s_0, A_1, s_1, A_2, \dots, A_n, s_n \rangle$$

such that for all $0 \leq i < n$:

- $(s_i, A_{i+1}, s_{i+1}) \in TR$,
- $A_T(s_i) \subseteq A_{i+1}$ (all triggered actions must occur),
- $A_T(s_i) \cap A_{i+1} \neq \emptyset$ (triggered actions are always included),
- $A_A(s_i) \cap A_{i+1} \neq \emptyset$ (allowed actions may occur),
- $A_I(s_i) \cap A_{i+1} = \emptyset$ (inhibited actions cannot occur).

A trajectory ensures that actions occur or do not occur according to the rules defined in $D(\mathbf{A}, \mathbf{F})$.

Action observation language

An action observation language allows to specify observations concerning the current state of fluents and the execution of actions,

Definition 5 (Action Observation Language [97]) *The action observation language of \mathcal{C}_{TAID} consisting of expressions of the following form:*

$$(f \text{ at } t) \text{ (} a \text{ occurs_at } t)$$

where $f \in \mathbf{F}$, a is an action and $t \in \mathbb{N}$ is a point in time.

The integration of observations O and the domain description D is referred to as an action theory (D, O) . An action theory forms the basis for constructing trajectory models, trajectories where all observations are satisfied, providing a structured representation of the system's dynamics over time.

Definition 6 (Action Theory [97]) *Let D be a domain description and O be a set of observations. The pair (D, O) is called an action theory. \mathcal{A}_T denotes all the possible action theories.*

A trajectory model is a structured sequence of states and actions that satisfies all observations O in an action theory (D, O) , ensuring that the described system evolution adheres to observed facts. Thus, trajectory models instantiate the dynamics encoded in action theories by constructing possible evolutions of the system that are observation-consistent.

Definition 7 (Trajectory Model [97]) *Let (D, O) be an action theory. A trajectory $\langle s_0, A_1, s_1, A_2, \dots, A_n, s_n \rangle$ of D is a trajectory model of (D, O) , if it satisfies all observations of O in the following way:*

1. if $(f \text{ at } t) \in O$, then $f \in s_t$
2. if $(a \text{ occurs_at } t) \in O$, then $a \in A_{t+1}$.

$\mathcal{M}_{(D,O)}$ denotes all the possible trajectories of (D, O) .

Action query language

The Action Query Language provides a means to inquire about specific sequences of actions and their impact. This is achieved by specifying subsets of the action set and their corresponding occurrences in time.

Definition 8 (Action Query Language [97]) *An action query language regards assertions about executing sequences of actions with expressions that constitute trajectories. A query is of the following form: $(f_1, \dots, f_n \text{ after } A_i \text{ occurs_at } t_i, \dots, A_m \text{ occurs_at } t_m)$ where f_1, \dots, f_n are fluent literals $\in \mathbf{F}$, A_i, \dots, A_m are subsets of \mathbf{A} , and t_i, \dots, t_m are points in time.*

By formulating queries in the action language, causal relationships between actions and their effects can be investigated, contributing to a deeper understanding of system dynamics, informed decision-making and controlled methods for automated planning. By combining observations with the causal laws of the domain description, we can generate plans, explanations, and predictions regarding the behavior of a system.

In later chapters, we demonstrate how action languages can be leveraged and extended to reason about the behavior of an "emotional system", enabling inferences about mental state dynamics. An action may influence "emotional fluents" to trigger transitions between "emotional states". However, the principles governing these transitions must be elicited and defined.

2.3 Quantitative Argumentation

Formal argumentation is a method of nonmonotonic reasoning and decision-making that models arguments, their interactions, and their evaluation within a logical framework. At its core, most commonly defined in terms of Dung's Abstract Argumentation Framework (AF) [94] representing arguments as nodes in a graph, with attack relations, the edges of the graph, define conflicts between them. The acceptability of arguments is determined by argumentation semantics, which determine which arguments survive under different acceptance conditions, producing sets of arguments, so-called extensions, which are consistent in the presence of attacks. Bipolar Argumentation Frameworks (BAFs) [66] extend this model by incorporating support relations between arguments, in addition to attacks. Quantitative Bipolar Argumentation Frameworks (QBAFs) [29] further refine this approach by assigning numerical weights to arguments while maintaining attack and support relations. In QBAFs, arguments are assessed under gradual semantics, producing a ranking based on their relative strength.

As we in this thesis particularly apply QBAFs, we cover its basic elements in this section.

Definition 9 (QBAF) *A QBAF is a quadruple $\langle X, R^-, R^+, \tau \rangle$, where X denotes a finite set of arguments, the binary relation $R^- \subseteq X \times X$ captures attack relations between arguments, 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 $\tau(a)$.*

In a QBAF, the attack relation captures arguments that counter another argument, while the support relation represents arguments that reinforce another argument. The attackers and supporters of an argument are the sets of arguments that attack or support it, respectively.

Definition 10 (Attack and support relations) *Let $Q = \langle X, R^-, R^+, \tau \rangle$ be a QBAF. For any argument $a \in X$, the set of its attackers is denoted as*

$R^-(a) = \{b \mid b \in X, (b, a) \in R^-\}$, and the set of its supporters is denoted as $R^+(a) = \{b \mid b \in X, (b, a) \in R^+\}$.

Example 2 The scenario presented in Example 1 can be represented as the QBAF $Q = \langle X, R^-, R^+, \tau \rangle$, where

- $X = \{ pu, w, why_w, tr, wi, yc \}$,
- $R^- = \{ (why_w, w), (tr, why_w), (wi, tr), (yc, wi) \}$,
- $R^+ = \{ (w, pu) \}$,
- $\{ \tau(pu) = 0.3, \tau(w) = 0.3, \tau(why_w) = 0.3, \tau(tr) = 0.3, \tau(wi) = 0.3, \tau(yc) = 0.3 \}$.

The arguments presented during the dialogue are captured within the set X . The base score, 0.3, is an arbitrary constant chosen a priori, represents information known beforehand about arguments' strength.

The strength function δ assigns a strength value to each argument by combining its base score with the strengths of its supporters and attackers. This process propagates through the argumentation structure, with leaf nodes retaining their base scores.

Definition 11 (Strength function) Let $Q = \langle X, R^-, R^+, \tau \rangle$ be a QBAF. The strength of an argument $a \in X$ is determined by a total function $\delta(a)$, where $\delta : X \rightarrow [0, 1]$. For any set of arguments $A \subseteq X$, we refer to the multiset $\{\delta(b) \mid b \in A\}$ as $\delta(A)$.

Example 3 A strength function δ can be characterized in different ways [29]. For instance, given a QBAF $Q = \langle X, R^-, R^+, \tau \rangle$ and $x, y, z \in X$:

$$\delta(x) = \max \left(0, \min \left(1, \tau(x) + \sum_{y \in R^+(x)} \delta(y) - \sum_{z \in R^-(x)} \delta(z) \right) \right)$$

meaning that $\delta(x)$ is computed by incorporating the base score $\tau(x)$, influenced by the accumulated strengths of supporting and attacking arguments while ensuring the result remains within $[0, 1]$.

In later chapters, we use QBAFs to model mental states within interactions. A set of topics, represented as arguments with base scores, is considered. By eliciting an agent's beliefs and applying a method for inferring relationships between topics from beliefs, we determine their strengths. To analyze transitions between QBAF-based mental state models, we incorporate them into formal dialogue games, which are introduced next.

2.4 Formal Dialogue Games

We draw inspiration from previous work on formal dialogues, such as [196, 38, 230, 141], which describe the formal properties and common elements of dialogue systems specialized for different types of dialogues, such as inquiry dialogues [38] and persuasion dialogues [230]. Dialogue systems that include formal descriptions of the involved agents, rules for legal moves, turn-taking, and conditions for termination are often referred to as dialogue games. These formalisms typically utilize two languages: a *topic language* \mathcal{U}_a^T , which describes the topic under discussion, and a *communication language* \mathcal{U}_a^B , which includes utterances such as beliefs that may be private or shared by agents.

For a given agent a , $\mathcal{U}_a := \mathcal{U}_a^T \cup \mathcal{U}_a^B$ denotes all the possible utterances from an agent a , $\mathcal{U}_a^T := \{t_1, \dots, t_p\}$ denotes the known topics by the agent a , and $\mathcal{U}_a^B := \{b_1, \dots, b_q\}$ denotes the known beliefs by the agent a . Sometimes, we just use $\mathcal{U} := \mathcal{U}^T \cup \mathcal{U}^B$ to denote topics and beliefs in general terms. Utterances are delivered to a dialogue through the so-called dialogue moves of an agent. While there are a range of moves considered in the literature, including *challenge*, *question*, etc., the following definition covers some basic moves:

Definition 12 (Dialogue moves) *The possible moves of an agent a are:*

- $\langle a, \text{open}, t \rangle$ such that $t \in \mathcal{U}_a^T$, called an open move.
- $\langle a, \text{close}, t \rangle$ such that $t \in \mathcal{U}_a^T$, called a close move.
- $\langle a, \text{assert}, b \rangle$ such that $b \in \mathcal{U}_a^B$, called an assert move.

The set of all possible moves of an agent a is denoted by \mathcal{M}_a .

These moves include “open”, marked as $\langle a, \text{open}, t \rangle$, where the agent initiates discussion on a topic t ; “close”, denoted as $\langle a, \text{close}, t \rangle$, signaling the conclusion of discussion on topic t ; and “assert”, represented by $\langle a, \text{assert}, b \rangle$, where the agent asserts a statement b .

A move sequence is a sequence of moves and is defined as follows:

Definition 13 (Move sequence) *Given a finite set of agents \mathcal{I} , a move sequence D_r^n is an ordered sequence of moves $[m_r, \dots, m_n]$, where each move $m_i \in \mathcal{M}_a$ such that $a \in \mathcal{I}$, ($r \leq i \leq n$) and $r, n \in \mathbb{N}$. \mathcal{D} is the set of all possible dialogues, such that:*

Sender(m_i) $\in \mathcal{I}$, in which Sender : $\mathcal{M} \rightarrow \mathcal{I}$ is a function where \mathcal{M} is the set of all the possible moves of the agents of \mathcal{I} .

Dialogue games allow for the analysis of communication dynamics, reasoning processes, and strategic behaviors among participants. *Dialogue protocols* have been defined to establish formal rules for representing different types of dialogues, commonly following Walton’s dialogue taxonomy (1997) [299] (see Table 2.2), each specifying initial assumptions, participants and goals, including dialogue types, such as Persuasion, Inquiry, Deliberation, Negotiation, Information-Seeking, Eristic, Debate, and Pedagogical dialogues.

Dialogue Type	Initial Situation	Participant Goals	Dialogue Goal
Persuasion	Difference of Opinion	Persuade Other Party	Resolve Difference
Inquiry	Ignorance	Contribute Findings	Prove or Disprove Conjecture
Deliberation	Contemplation of Consequences	Promote Personal Goals	Formulate Priorities
Negotiation	Conflict of Interest	Maximize Gains (Self-Interest)	Settlement, Harmony
Information-Seeking	One Party Lacks Information	Obtain Information	Transfer of Knowledge
Eristic	Personal Conflict	Verbally Hit Out at Opponent	Reveal Deeper Conflict
Debate	Adversarial	Persuade Third Party	Air Strongest Arguments for Both Sides
Pedagogical	Ignorance of One Party	Teaching and Learning	Transfer of Knowledge

Table 2.2: Eight Types of Dialogue (Adapted from Walton (1997) [299])

Agents participate in the dialogue by selecting utterances and committing to information exchanged throughout the interaction. While there is a range of dialogue agent definitions in the literature, specialized for different settings, a general dialogue agent can be defined as follows.

Definition 14 (Dialogue Agent) *An agent a is a tuple of the form $a := \langle \mathcal{U}_a, S_a \rangle$, where:*

- \mathcal{U}_a is a finite set of utterances available to the agent.
- $S_a : \mathcal{D} \rightarrow \mathcal{U}_a$ is a selection function that determines the next utterance based on the dialogue history.

A dialogue, including its protocol, is then defined as a structured exchange of moves between two agents, adhering to specific constraints on initiation, turn-taking, and closure.

Definition 15 (Dialogue) *A dialogue is denoted $\langle \mathcal{I}, D_r^n \rangle$, where $D_r^n = [m_r, \dots, m_n]$ is a move sequence, and $\mathcal{I} = \{a, b\}$ is a set of agents, where $r, n \in \mathbf{N}$ and $r \leq n$, satisfying the following conditions:*

1. $m_r = \langle x, \text{open}, t \rangle$, where $x \in \mathcal{I}$ and $t \in \mathcal{U}_x$.
2. $\text{Sender}(m_s) \in \mathcal{I}$ for all $r \leq s \leq n$.
3. $\text{Sender}(m_s) \neq \text{Sender}(m_{s+1})$ for all $r \leq s < n$.
4. For all open moves $m_s = \langle x, \text{open}, t \rangle$ ($r \leq s < n$), there exist two corresponding close moves: $m_j = \langle y, \text{close}, t \rangle$, $m_k = \langle x, \text{close}, t \rangle$ such that $y \neq x$, $y, x \in \{a, b\}$, and $s < j < k \leq n$.

In this example, the first move in the dialogue is an open move initiated by any agent $x \in \mathcal{I}$. Each move is performed by one of the agents, with turns alternating throughout the dialogue. Every topic introduced must be closed by both agents, ensuring mutual participation in concluding the discussion. The dialogue maintains a structured exchange, where each topic is addressed and properly closed before the conversation ends.

Example 4 *In the context of Example 1, a dialogue adhering to the protocol in Definition 15 can be represented as follows: $\langle \mathcal{I}, D^{[0,7]} \rangle$, $\mathcal{I} = \{a, b\}$, $D^{[0,7]} = [\langle b, \text{open}, pu_b \rangle, \langle a, \text{assert}, wa \rangle, \langle b, \text{assert}, why_wb \rangle, \langle a, \text{assert}, ta \rangle, \langle b, \text{assert}, wb \rangle, \langle a, \text{assert}, yc_a \rangle, \langle b, \text{close}, pu_b \rangle, \langle a, \text{close}, pu_b \rangle]$.*

Additional conditions or specializations to the protocol apply depending on the dialogue type, such as specific turn-taking and termination rules, which shape the structure of interactions. As a result, formal dialogue games provide a foundation for analyzing long-term interactions within these structured protocols.

2.5 Answer Set Programming

Answer Set Programming (ASP) is a declarative programming paradigm based on nonmonotonic reasoning and the stable model semantics of logic programs [117]. It is used for solving difficult, typically NP-hard, combinatorial search problems, knowledge representation and reasoning tasks. ASP can express all NP-search problems that are solvable using a nondeterministic Turing machine in polynomial time, such that the solutions correspond to the stable models (or answer sets) of the logic program [49]. Efficient computation of stable models is performed using specialized solvers, such as `clingo`¹ [116] and `dlv2` [101]. In this section, we present a short overview of ASP, the syntax of logic programs, and answer set semantics.

The language of a propositional logic has an alphabet consisting of proposition symbols p_0, p_1, \dots , logical connectives $\vee, \wedge, \leftarrow, \neg, \text{not}, \perp, \top$, and auxiliary symbols $(,)$. A signature L is a finite set of propositional atoms. The negation sign \neg represents *strong negation*, which explicitly asserts the falsity of a proposition, while *not* denotes *negation as failure*, indicating the absence of evidence for a proposition. A literal is either an atom a or its negation *not a*.

A clause consists of a head and a body, written as:

$$a \leftarrow b_1, \dots, b_k, \text{not } c_1, \dots, \text{not } c_l$$

where a is the head of the clause, and the body contains a set of positive literals b_1, \dots, b_k and negated literals *not* $c_1, \dots, \text{not } c_l$.

¹<https://potassco.org/clingo/>

²<https://www.dlvsystem.it>

A clause where the body contains only positive and negated literals without a head is called a constraint:

$$\perp \leftarrow b_1, \dots, b_k, \text{ not } c_1, \dots, \text{ not } c_l$$

where \perp is an atom that always evaluates to false. A clause with an empty body is called a fact and is written as:

$$a \leftarrow \top$$

where \top is an atom that always evaluates to true. This is sometimes written simply as a .

A logic program Π is a finite set of clauses, where each clause is either a rule, a fact, or a constraint.

The ground instantiation of Π , $\text{gr}(\Pi)$, replaces variables with constants, yielding a propositional form. An interpretation I , a subset of ground atoms, satisfies a literal l if $l \in I$ and satisfies *not* l if $l \notin I$. The reduct Π^I of Π relative to I is obtained by removing rules that contain *not* c where $c \in I$, and eliminating all remaining negative literals. I is a stable model if it is exactly the set of literals derivable from Π^I and minimal with respect to set inclusion. Stable models, or answer sets, are minimal sets of literals that satisfy all rules of the program. A program Π is consistent if it has at least one stable model. This minimality condition ensures that answer sets represent stable, consistent interpretations of the program.

ASP has been widely used to represent interactions and dynamic processes over time, particularly in domains requiring structured reasoning about change, action effects, and constraints, including commonsense reasoning, argumentation frameworks, and formal dialogues. By computing stable models, ASP enables the exploration of valid state transitions, planning trajectories, and logical consistency in complex systems. For instance, the binary predicates `holds/2` and `occurs/2`, and their variants, are fundamental constructs in reasoning about actions and change [197, 26, 171]. By computing stable models, ASP provides a method to represent and reason about interactions and verify system properties.

Illustrating ASP with Blocks World

The Blocks World domain is a well-known example for reasoning about dynamic systems, where blocks can be stacked on top of each other or placed on a table according to predefined rules. This domain is often used to illustrate how ASP can model and reason about actions and constraints, allowing us to predict system evolution, verify constraints, and generate plans. These examples are adapted from Lifschitz (2019) [171].

Fluents and state representation

A system state is described using fluents, which are logical propositions representing properties of objects in an environment at a given time. In Blocks

World, fluents represent the location of blocks. Formally, let \mathcal{B} be the set of blocks, and let $\mathcal{L} = \mathcal{B} \cup \{\text{table}\}$ be the set of possible locations. A fluent $\text{loc}(B,L)$ expresses that block $B \in \mathcal{B}$ is located at $L \in \mathcal{L}$. The set of blocks in this example consists of $\mathcal{B} = \{a, b, c\}$. For example, an initial state where a is on b , b is on the table, and c is also on the table is encoded as:

```
holds(loc(a,b),0).
holds(loc(b,table),0).
holds(loc(c,table),0).
```

The predicate $\text{holds}(F,T)$ expresses that fluent F holds at time T .

Actions and effects

The predicate $\text{occurs}(A,T)$ specifies when an action A takes place. Discrete time steps T are represented by integers from 0 to a given horizon $h \in \mathbb{N}_0$, where \mathbb{N}_0 represents the set of nonnegative integers.

The predicate $\text{move}(B,L)$ represents moving a block B onto a location L , which can be another block or the table. For example, moving a block a on top of a block c at time 0 and then moving block b on top of a block a at time 1 is represented in ASP as:

```
#const h=2.
occurs(move(a,c),0).
occurs(move(b,a),1).
```

The effects of actions are specified using rules that describe how state variables change when actions occur. Moving a block B in time T to a new location L updates its position at the next time step $T+1$:

```
holds(loc(B,L),T+1) :- occurs(move(B,L),T).
```

Frame Problem and Inertia

A key challenge in dynamic reasoning is the frame problem—how to specify that fluents remain unchanged unless affected by actions. This is addressed using the inertia rule:

```
holds(loc(B,L),T+1) :- holds(loc(B,L),T),
                        not -holds(loc(B,L),T+1),
                        T = 0..h-1.
```

This rule ensures that a block retains its position unless an action explicitly changes it. The double negation `not -` represents the assumption that a fluent remains true unless evidence to the contrary exists. This principle, known as the commonsense law of inertia, is crucial for correctly modeling dynamic systems.

Constraints and Nonexecutable Actions

Constraints prevent physically impossible states. For example, ensuring that no two blocks can be on the same block at the same time:

```
:- holds(loc(B1,B),T), holds(loc(B2,B),T),  
B1 != B2, block(B), block(B1), block(B2).
```

Other constraints enforces that a block can only be moved if no other block is on top of it:

```
:- occurs(move(B,L),T), holds(loc(_,B),T).
```

Prediction

Prediction in ASP involves computing the consequences of a sequence of actions given an initial state. This allows verification of whether a specified sequence of actions leads to the expected system evolution. The following ASP program predicts how a Blocks World configuration evolves:

```
block(a; b; c).  
#const h=2.  
init(loc(a,b); loc(b,table); loc(c,table)).  
occurs(move(a,c),0; move(b,a),1).
```

Here, the predicate `block/1` defines the set of blocks \mathcal{B} , while `init/1` defines the initial fluents. By running this program in `clingo`, we obtain a stable model (answer set) representing how the system evolves over time.

A possible answer set for this prediction program is:

```
holds(loc(a,b),0).  
holds(loc(b,table),0).  
holds(loc(c,table),0).  
occurs(move(a,c),0).  
holds(loc(a,c),1).  
holds(loc(b,table),1).  
holds(loc(c,table),1).  
occurs(move(b,a),1).  
holds(loc(a,c),2).  
holds(loc(b,a),2).  
holds(loc(c,table),2).
```

Planning

Planning in ASP involves generating a sequence of actions that leads from an initial state to a desired goal state. Instead of defining a fixed sequence of actions, planning problems allow ASP solvers to search for a valid sequence of actions that satisfies given constraints. The following ASP rules encode a planning approach for Blocks World:

```

action(move(B1,B2)) :- block(B1), block(B2), B1 != B2.
action(move(B,table)) :- block(B).
{occurs(A,T) : action(A)} = 1 :- T = 0..h-1.

```

These rules define a choice over possible actions at each time step, allowing the ASP solver to explore different action sequences. The constraint ensuring that a solution must lead to the goal state can be encoded as follows, ensuring that block `b` must be on top of block `c` at the final time step `h`:

```

:- not holds(loc(b,c),h).

```

The ASP solver searches for a sequence of actions that satisfies this constraint. If a stable model is found, it represents a valid plan. If no stable model exists, no such plan of length `h` is possible.

A possible answer set for a valid plan that satisfies this goal can be:

```

holds(loc(a,b),0).
holds(loc(b,table),0).
holds(loc(c,table),0).
occurs(move(b,c),0).
holds(loc(a,b),1).
holds(loc(b,c),1).
holds(loc(c,table),1).

```

This approach treats planning as a constrained search process. By encoding actions as choices and enforcing conditions that must hold at each time step, ASP finds sequences of actions that transition the system from its initial state to a specified goal. These examples highlight ASP's role in formal verification, prediction, constraint enforcement, and planning.

2.6 Chapter Remarks

In the context of this thesis, ASP serves as a declarative framework for modeling and reasoning about dynamic interactions, making it well-suited for representing formal dialogues and mental state dynamics. Similar to classical planning domains like Blocks World, ASP encodings in this setting address constraint satisfaction, prediction, and planning. A fundamental challenge in applying logic-based methods to HAI lies in determining the appropriate representations of objects, fluents, actions, and constraints within the domain of human mental states.

The two pillars of this work: (1) psychological theories for understanding the dynamics of the human mind and (2) formal methods for non-monotonic reasoning. On one hand, psychological theories provide a scientifically grounded foundation for modeling the complexity of the human mind. On the other

hand, ASP’s non-monotonic reasoning capabilities complement these psychological insights by offering a computational mechanism to account for the inherent variability in mental state reasoning. Unlike deterministic methods, ASP allows for the generation of multiple answer sets, making it well-suited for modeling scenarios with incomplete information, or multiple valid interpretations. This capability is particularly useful in reasoning about state transitions where different sequences of actions or conflicting pieces of information may lead to distinct yet logically consistent outcomes. In the context of HAI, answer sets must be filtered and evaluated in terms of psychological principles.

Action languages provide an intuitive way to specify dynamic reasoning processes by offering a structured, declarative representation of actions, effects, and constraints. This makes them particularly effective for knowledge elicitation, as they help translate expert-driven theories—such as psychological theories—into executable logic-based systems. By integrating action languages with ASP, the elicited psychological theories can be computationally realizable, adaptable, and verifiable. Another method for knowledge elicitation regards our use of ontologies, providing a formal and structured way to represent and reason about context, supporting the integration of psychological theories into models of mental states. We particularly use ontologies for modeling empathy, a central concept in the context of interpersonal human interactions.

QBAFs play the role of capturing gradual evolution of beliefs and mental states. In the setting of manipulation, where beliefs may be gradually influenced over long-term interactions, quantitative measures become important. A verification system must be capable of defining boundaries for acceptable belief change, detecting subtle accumulative effects. QBAFs and their gradual semantics allow for verification approaches that move beyond binary classifications to continuous assessments of belief modification. Moreover, defining boundaries for belief change is essential in regulatory and ethical AI assessment. The European Union’s Artificial Intelligence Act (AI Act) [105] adopts a risk-based approach to regulating AI, categorizing systems into unacceptable, high, limited, and minimal risk levels. The AI Act explicitly addresses emotion recognition [95] and manipulative AI systems [162], aiming to mitigate risks associated with covert influence strategies. However, a major challenge in enforcing such regulations is the lack of concrete verification tools that can quantify and assess these risks. A logic-based verification approach, enhanced with quantitative measures, can provide a formal, explainable, and computationally grounded methodology for assessing whether a system operates within acceptable boundaries.

Chapter 3

Empathy Verification

In this Chapter, we study the notion of empathy in interactive software agents, such as chatbots, which are increasingly used in health and well-being applications. Understanding their empathic capabilities is essential, yet there are no established tools for assessing empathy in such agents. To address this gap, we develop a formal ontology of empathy based on a systematic literature review of chatbots in health and well-being applications and a qualitative analysis of the notions of empathy identified in these works. The ontology is evaluated in a controlled user study, assessing empathy in two health-focused chatbots, Replika and Wysa. The findings suggest that the ontology captures necessary conditions for assessing empathy and reveals trends in changing perceptions over time. Implemented in Web Ontology Language (OWL), it offers a structured approach for recognizing and evaluating empathy in interactive agents, supporting both self-assessment by agents and external evaluation by intelligent systems.

3.1 Introduction

In everyday interpersonal interaction, humans largely depend on various notions in order to communicate. A cornerstone is the notion of empathy, the ability to understand and share the feelings of another [102]. Humans rely on their, as well as others', empathy to develop social bonds, to build and maintain trust, to collaborate, and to reach individual and shared goals. Humans need, and expect, empathy from others to different degrees in various contexts, but in scenarios where an individual seeks understanding or help from others, empathy is of particular concern.

Interactive software agents, embodied as Conversational Agents (CAs), such as chatbots [180], are progressively being used in the area of health and well-

being [180], designed for, e.g., coaching, comforting and behavior-change applications [264]. Such systems, more commonly designed using various techniques from Natural Language Processing (NLP) [34, 92] and Machine Learning (ML) [169, 269, 270], build response generation models to process and respond to user inputs in a conversational manner. Over recent years, advancements in large-scale language models, such as GPT [287], T5 [211], and LaMDA [289], have significantly improved chatbots' ability to generate human-like text, yet they struggle with maintaining coherence over prolonged interactions. Despite these challenges, interest in chatbot applications continues to expand. CAs, designed to simulate human-human conversations, have strong potential in various contexts, particularly as social companions to support mental well-being, helping individuals with loneliness, depression, and stress management. To provide such support, systems must meet users' expectations for relevance, coherence, and personalization. This requires interactive agents that can perceive, understand, and act empathically. When applications involve intimate, in-person interactions, such as social companionship, achieving interpersonal connectivity and relatedness becomes even more critical, necessitating a higher degree of empathic capability. While this development can have positive outcomes to users, it can also come with great risks of overtrust and influence that is harmful. As seen in Example 1 (highlighted in the introduction of this thesis), users can be misled by chatbots as their human-like attributes increase, and more cases continue to arise¹.

A starting point is to investigate how current interactive agents handle the empathic aspects of interaction by assessing how their empathic capabilities are perceived. Another challenge is making these empathy assessments computational, allowing intelligent systems to evaluate and reason about empathic performance. Research in computational empathy [42, 220, 179, 311] has primarily focused on emotion recognition and empathic emotion expression, such as models enabling agents to express emotions through facial mimicry [42]. While emotion is a key component of empathy, it is insufficient for a comprehensive understanding of empathic capability. Various approaches for evaluating and assessing empathic capability in interactive agents have been proposed [220, 311], but these approaches tend to be abstract and do not specify necessary conditions for perceived empathy. Furthermore, they are not designed to be computational or to facilitate automated empathy reasoning. To understand computational empathy holistically in state-of-the-art software systems and advance automated empathic reasoning in future systems, we require computational models and tools for explaining and assessing perceptions of a system's empathic capabilities. For this, a formal definition and vocabulary are needed to capture the essential aspects of empathy in human-agent interaction. While the literature explores various (in)formal notions of human empathy [88], there is no agreed-upon formal definition of "computational empathy" to assess interactive agents' empathic capabilities.

¹Lawsuit on CharacterAI: <https://www.nytimes.com/2024/10/23/technology/characterai-lawsuit-teen-suicide.html>

By considering the lack of consensus on empathy in human-agent interactions, the following research questions arise:

RQ3.1 How can we understand empathy to assess it in human-agent interactions?

RQ3.2 What are the minimum necessary conditions for assessing computational empathy?

Through a systematic literature review of prior research on empathy in CAs, data was collected regarding the notion of empathy in these systems. Following data collection, a series of qualitative data modeling steps was conducted using a Grounded Theory (GT) methodology [72]. The collected data was analyzed and conceptualized, resulting in an ontology for computational empathy, implemented using Web Ontology Language (OWL)². This is the first computational semantic model introduced as an ontology that captures functional necessary conditions for computational empathy.

We demonstrate the potential of this approach by applying it as a qualitative tool for assessing empathy in interactive software agents in the area of health and well-being. The evaluation is based on a controlled user-study, where participants interacted with two of the most popular health and well-being chatbots [303], Replika³ and Wysa⁴, and completed an assessment protocol based on the proposed definition of empathy (see Figure 3.1). To the best of our knowledge, no previous effort has been made to assess computational empathy as introduced in this work. Findings show that our definition can distinguish empathy in CAs at different levels of understanding by analyzing various sub-dimensions of empathy.

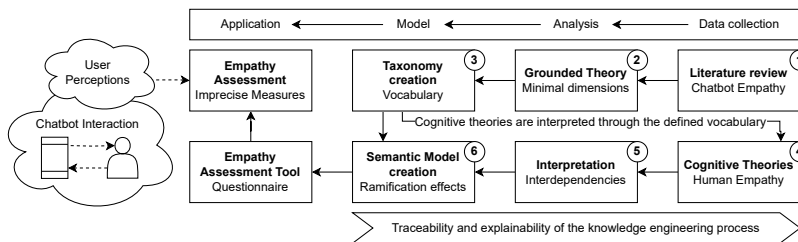


Figure 3.1: Research Design and Methodology.

The main contributions of this chapter are: (1) a formal multi-dimensional definition of computational empathy, (2) a qualitative methodology for evaluating computational empathy, including an assessment protocol and analysis method using precise and imprecise uncertainty measures, (3) an OWL ontology implementation, provided as open access material, and (4) a qualitative

²OWL ontology: <https://github.com/ComputationalEmpathy/empathy-ontology>

³Chatbot Replika: <https://replika.ai>

⁴Chatbot Wysa: <https://wysa.io>

assessment of computational empathy in two state-of-the-art chatbots, Replika and Wysa, using precise and imprecise uncertainty measures.

This chapter is organized as follows. Section 3.2 presents seminal and recent work defining and measuring empathy. Section 3.3 provides a literature review of prior work in CA-human interaction, which informs our definition of empathy. Section 3.4 outlines the data analysis and qualitative data modeling steps. Section 3.5 presents our formal definition of empathy, comprising six key dimensions. Section 3.6 evaluates the definition by applying it as a tool for measuring empathy in CAs. Section 3.7 concludes the chapter.

3.2 Related Work

Empathy has been understood as a multi-dimensional process that can be divided into two overall aspects, (1) affective/mirroring processes and (2) cognitive/reconstructive processes [122]. The affective processes concern arousal from, and (involuntary) response to, an agent’s expressed experiences, and the cognitive processes concern deliberation and interpretation of an agent’s experiences and delivering an inferred understanding back to the agent through interactions [103]. In this direction, empathy has been suggested as a complex interaction between affective and cognitive processes, in terms of inversely related influences between (1) theory of mind constructions, (2) awareness of affective responses, and (3) primitive affective resonance induction [304]. From a social cognition viewpoint [282], empathy has been considered in relation to interconnected and shared emotions. In this view, interpersonal social relations and groups with shared emotions are considered as important social constructs related to the definition of empathy. From a communication theory viewpoint, the Interpersonal Adaptation Theory (IAT) [54], understood to be closely related with empathic capability, addresses ways in which individuals adapt to one another in interactions to, for instance, mimic and synchronize behavior, social relation, and personality with another. This highlights an interpersonal aspect of empathic capability.

In the literature, different measures and models of human empathy have been developed. Let us look at some of them next.

Measures of Empathy

A variety of measures have been informed by previous definitions, used as tools for assessing empathy in humans. These measures have typically relied on self-reports, for instance, the Questionnaire Measure of Emotional Empathy (QMEE) [200] and Hogan Empathy Scale (HES) [138], mostly focusing on affective/emotional empathy, leaving out cognitive empathy, or vice versa. A more recent work by Jolliffe and Farrington (2006) [149] criticizes the one-sided approaches of QMEE and HES. Jolliffe and Farrington further criticize that QMEE and HES only capture hypothetical scenarios, with questions in the

format "I feel [hypothetical emotion] when I see [hypothetical scenario]." Instead, they propose a measure focusing on a participant's previous experiences and taking into account both cognitive and affective empathy, the so-called Basic Empathy Scale (BES) [149], with questions based on four 'basic emotions' (fear, sadness, anger, and happiness). Examples of questions in BES are "I get caught up in other people's feelings easily" (affective) and "I can often understand how people are feeling even before they tell me" (cognitive) [149].

While these previous measures of empathy have been focused on self-reports to measure empathy of the individual, we can observe limitations in these measures to capture empathy in interactive software agents because (1) they measure the individual's own empathy, in comparison to an interlocutor's perceived empathy, and (2) a software agent may not have the same criteria for empathy as humans. In order to measure computational empathy, we need a formal definition of what empathy means in a software agent, and if that measure should be derived from a human's perspective, then we need an assessment protocol with questions focused on an individual's perceptions of an agent's empathic capabilities in the interaction. These questions must be based on aspects that can be perceived and understood by the participant, and the formal model must be able to infer higher-level concepts based on low-level perceptions.

Models of Empathy

A variety of models of empathy have been developed to better understand empathy in humans in a broad sense. For instance, the Perception and Action Model (PAM) [233]. PAM is a model for empathy developed in cognitive neuroscience that defines empathy as a shared emotional experience occurring when an individual comes to feel a similar emotion to another as a result of perceiving, or imagining, the other's (emotional) state. PAM is informed by findings that there are shared representations (activation) in the human brain for perceiving and generating action. For instance, as an individual perceives a person crying, similar brain activation can be measured as if the perceiver was crying. PAM is further extended such that perception can arise from situations where the subject is not directly perceiving an object but imagines the state of the object (in the current work, we refer to these kinds of phenomena of an agent's mental state as its *manifestation*). Given the neuroscience perspective of PAM, we see limitations in using this theory alone for modeling computational empathy in interactive software agents, but it is an important theoretical base to be acknowledged; When modeling cues of empathy in an interlocutor agent, dependencies between the agent's perception and action, described by PAM, are critical to be captured by the model.

A recently defined model of empathy is the Nine Dimensions of the Empathic System (ES9) [127]. ES9 considers empathy as a complex series of multi-dimensional processes that comprise an "empathic system." In their taxonomy, empathy is defined in terms of nine high-level dimensions: (1) catalyst, (2) function, (3) process, (4) outcome, (5) affective state, (6) cognitive state,

(7) self and other, (8) behavior, and (9) context. This is relevant for getting an overarching perspective on human empathy. A limitation in ES9 is that the definition consists of abstract high-level concepts, not intended to be measurable. To enhance precision, abstract notions like “affective state” (in ES9) need concrete, observable, low-level counterparts for objective assessment. Furthermore, ES9 lacks computational applicability, requiring further exploration and modeling to be integrated into automated reasoning. This highlights the need for a formal definition and computational model of empathy.

The following section presents a literature study of prior work in CA-human interaction, analyzed through a Grounded Theory process [72], on which we inform our definition of computational empathy.

3.3 Systematic Literature Review

We conducted a systematic literature review to gather data on conversational agents (CAs) in the context of health and well-being, specifically focusing on the concept of empathy within these systems (see Figure 3.2). Adhering to PRISMA guidelines resulted in a number of search results based on our predefined criteria. After reviewing the titles and abstracts of the retrieved papers against the inclusion criteria, 29 papers were selected for data analysis (see Table 3.2). In the following sections, we outline the methodological steps undertaken in this review, following the PRISMA-S guidelines for conducting systematic literature reviews [244].

Search Strategy. The search strategy aimed to identify relevant literature for further assessment regarding eligibility and inclusion. Given the diverse nature of chatbot-related literature, encompassing both technical and conceptual perspectives, we opted for general research databases that included both aspects. Data were collected from two databases: Google Scholar and ACM Digital. Our approach did not involve utilizing cited or citing references of the included papers, nor did we incorporate additional studies from other sources, as we focused solely on papers retrieved based on our search criteria. The primary search term used was [“Conversational agent” OR “Chatbot”], given its prevalence in health and well-being applications. Additional search terms included AND [“Well-being” OR “Health”] AND [“Empathy”]. Figure 3.2 illustrates the systematic review process, guided by the following eligibility criteria: (1) Publication date between 1950–2021, (2) Papers must be in English, (3) Papers must be peer-reviewed, and (4) Papers must define empathy, evaluate empathy, and describe an interactive system in health and well-being.

Relevance Assessment Criteria. The papers that met the inclusion criteria were assessed based on a relevance evaluation framework (see Table 3.1), similar to the approach by Ahmad et al. [5]. Since our primary focus was on defining empathy, we prioritized studies that provided an

Quality Criteria	Score
Do authors clearly state the purpose of the chatbot(s)?	Yes = 1; No = 0
Do authors address psychological theories?	Yes = 1; No = 0
Do authors define empathy?	Yes = 2; No = 0
Do authors evaluate chatbot(s)?	Yes = 1; No = 0
Are there any considerations on trustworthiness?	Yes = 1; No = 0
Do authors clearly state limitations?	Yes = 1; No = 0

Table 3.1: Relevance assessment criteria.

empirical or theoretical definition of the concept. Additionally, we sought evaluations of chatbots that indirectly assessed empathy definitions. Our criteria were informed by extensive research on empathy from disciplines beyond computer science, such as psychology. Since trust and empathy are closely linked [115], we also considered author discussions on trustworthiness. Furthermore, we included limitations as a relevance criterion to ensure a rigorous selection process. In addition to the inclusion criteria, this quality assessment framework allowed for a structured evaluation of essential aspects of the papers. Papers needed to achieve a minimum score of '2' in the relevance assessment (Table 3.1) to be included in the final selection. A score of 2 was primarily assigned to papers that offered a clear and explicit definition of empathy, as this aspect was deemed most critical to the study. Other factors within the criteria were each assigned a score of 1, signifying their importance in contributing to the relevance of the research.

Data Extraction Process. The selected papers underwent thorough examination and were re-read by multiple authors. Using a structured approach and predefined criteria, we discussed and extracted the notion of empathy presented in each study. As shown in Table 3.2, these notions of empathy manifested in various forms, such as humor, relationships, or sociability. The selected papers aligned with our inclusion and relevance criteria by presenting different conceptualizations of empathy. We employed the five Ws framework [191], examining who, what, when, where, and why different notions of empathy play a role. Finally, we extracted pertinent sections from the reviewed material as a preparatory step before initiating the first phase of the Grounded Theory process.

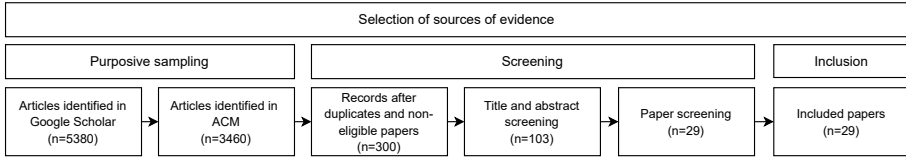


Figure 3.2: Systematic literature review, following guidelines provided by the PRISMA framework [244]

Paper	Year	Chatbot context	Empathy notion
[22]	2021	Social support	Emotional Support
[31]	2021	Supporting disorders	Therapeutic bonds; building rapport
[59]	2018	Self-assessment mental health	Error management; personality and understanding
[65]	2021	Empathic social chatbot	Detect emotions and respond accordingly
[67]	2021	Social well-being	Humor; self-defeating jokes evoke empathy
[70]	2020	Social support	Engaging; human-like
[73]	2021	Obstetric and mental health	Sympathetic responses; facilitate sharing knowledge
[82]	2021	Social companion	Goal-oriented; contextual; timing; good listener
[86]	2020	Social ostracism	Show understanding; empathic feedback
[111]	2020	Self-help mHealth	Reply emphatically; social awareness; interpersonal relationship
[121]	2019	Behavior change	Emotionally appropriate; acknowledge users emotional state
[125]	2019	Well-being support	Notice and acknowledge attainable goals
[126]	2021	Encouragement	Personality; adaptable to mood
[132]	2020	Build alliance	Affective empathy; supportive; coaching
[143]	2018	Mental resilience	Balanced engagement efficiency
[145]	2018	Social, chit-chat	Sensitive to context; personalized
[146]	2021	Increase concentration	Empathic responding; build relationship
[161]	2021	Trust in HCI	Understand statements; behaviors or feelings of another
[167]	2019	Self-care	Relate to suffering; Self-compassion
[165]	2020	Promote self-disclosure	Understand people; tone-aware
[170]	2021	Empathetic chatbot	Understand feelings and experiences; emotion causes
[174]	2019	Emotion detection; empathic responses	Express and perceive emotion; social bonding
[181]	2017	Self-reflection, practice behaviors	Tailored mood; tailored content
[183]	2020	Optimize preconception health	Humanness; acknowledge emotions
[199]	2019	Emotional support	Expressing empathy; supportive messages
[208]	2018	Empathic support	Perceive causes to emotions
[238]	2021	Health advice	Sentiment; emotion detection
[250]	2020	Depression support	Sociability; engaging in interactions
[273]	2021	Student's well-being	Show understanding; show affect; facilitate sharing

Table 3.2: Included papers

3.4 Data Analysis and Modeling

Proceeding the data collection, a series of qualitative data modeling steps was conducted following the methodology of Grounded Theory [72]. Grounded theory is a well-established method for knowledge engineering [172, 228, 247, 313], used to systematically code, categorize, and relate qualitative data. It provides a structured methodology for knowledge elicitation in the construction of qualitative knowledge bases. This iterative process ensures traceability and explainability in knowledge elicitation. Through this approach, we can systematically trace the process from raw qualitative data to the final computational model. Grounded theory is particularly relevant to our study due to its capability to identify and define concepts based on their interrelations, which aligns with our objective of establishing a structured model of computational empathy.

The grounded theory process involves the following steps: (1) initial coding, which involves identifying meaningful segments of information in the raw data extracted from the selected papers, (2) intermediate coding, which clusters the initial codes into categories, and (3) advanced coding, which defines overarching themes by grouping the categories from the previous step. This process identifies recurring concepts in the qualitative data that collectively contribute to forming more abstract, higher-level concepts.

Theoretical saturation. The goal of the data analysis is to determine a minimal set of themes (referred to as dimensions) that form a comprehensive framework for computational empathy. A critical benchmark for deriving a minimal set of concepts from qualitative data is the notion of *theoretical saturation* [261], which refers to the point in data analysis where no new insights or concepts emerge, and an overarching set of themes has been established, supported by a hierarchy of codes and categories derived from the data (see Table 3.3).

Although there are no strict rules for identifying the point at which theoretical saturation is reached, various methods help approximate it. Saturation is understood as a measure of completeness, where further data collection yields diminishing returns, meaning any newly identified code or category is either a synonym or can be subsumed under previously established concepts. Theoretical saturation is considered achieved when a “sufficient depth of understanding” [261] has been reached regarding the emerging categories and themes. In our study, “depth of understanding” is conceptualized as levels within a hierarchical structure. The hierarchy begins with the abstract root concept, which we define as “Computational empathy.” Each subsequent level contains less abstract concepts, where each node provides a partial explanation of its parent node. Theoretical saturation was determined when: (1) the lowest-level concepts could be objectively interpreted, (2) the parent nodes (categories and themes) could be explained based on their child nodes, and (3) the over-

all themes comprehensively captured the root concept on an abstract level. To evaluate the completeness of the model, we analyzed it by referencing empathy aspects from cognitive theories, including PAM [233], Social Cognition [22], and IAT [54], as previously discussed in Section 3.2. This process yielded a structure comprising three levels of understanding, culminating in a set of six overarching themes.

In what follows, we describe the three stages of our data analysis, which ultimately result in a hierarchical graph structure that forms the foundation of our proposed formal model of computational empathy.

Initial coding. The first step focused on segmenting the qualitative data into smaller, meaningful units, referred to as codes. This involved identifying relevant excerpts that aligned with the initial search queries. If a particular pattern was found across multiple papers, the corresponding codes were grouped into dominant codes. Initially, 80 codes were generated, each capturing key aspects related to the definition of empathy. Codes were extracted to facilitate qualitative data aggregation, as illustrated in the following examples: “[...] *expected the chatbot to retain context [CONTEXT-SENSITIVE] across chat sessions, thus providing users with personalized [PERSONALIZED] recommendations [...]*” (Jain et al., p. 901), “*sentiment [SENTIMENT] and emotion detection [EMOTION DETECTION] technique may be utilized to identify the corresponding [...]*” (Rahman et al., p. 13), “*An empathic human-like [HUMAN-LIKE] chatbot avatar promoting natural conversations can promote user engagement [ENGAGING] towards chatbot services.*” (Chen et al., p. 222). This aggregation step is further detailed in Table 3.4 and Table 3.5. The next step involved clustering these codes into broader categories while preserving their links to the qualitative data.

Intermediate coding. At this stage, 77 codes were grouped based on their conceptual relationships. Subsequently, a secondary clustering process was undertaken to assign appropriate labels to these groups. This resulted in the formation of 13 distinct categories: *Traits, Acts, Manifestations, Expressions, Moves, Activities, Interpersonal, Building trust, Functional performance, Context, Emotions, Mentalization, and Perception*, each consisting of a set of related codes (see Table 3.3). The next step was to consolidate these categories into a higher-level set of themes.

Advanced coding. Through an iterative process, the previously established categories were further analyzed, refined, and regrouped, with some codes being revised, added, or removed to fit the characterization of empathy dimensions. This process resulted in a refined set of six dimensions: *{Perceive, Act, Theory of Mind, Manifest, Ratification, and Interpersonal}*. These dimensions were then systematically linked to the underlying categories and initial codes, forming a hierarchical graph structure.

By defining themes based on the outcomes of the preceding steps, we ensured that their connections to the original qualitative data were preserved. Through achieving theoretical saturation at each stage, we established that these six dimensions represent the necessary conditions for computational empathy.

Initial coding: Codes (n=77)
Emotional self-awareness CAT[11], Perspective-taking CAT[7,8,12], Social Awareness CAT[7,11], Sense, hurt or pleasure as he senses it CAT[7,8,12], Human-like CAT[3,7,12], Curious CAT[7], Naturally understand CAT[7,8,12], Tone-aware CAT[9,10], Suffering CAT[2,4,5,12], Self-defeating CAT[1,2,4,5,8], Natural CAT[7], Savoring CAT[2,5], Lack of humanity CAT[3,7,12], Sadness/loneliness CAT[1,2], Appreciation CAT[3], Positive emotion CAT[11], Intelligence CAT[1,2,10], Personalized CAT[1,2,5,7,10], Humour CAT[1,2,5,6], Mood CAT[4,2,10], Inspiring CAT[1,2], Charismatic CAT[1,2], Affective CAT[4], Engagement CAT[7, 8], Conflict resolution CAT[6], Comforting CAT[5,8], Coaching CAT[6], Engaging CAT[4,6,7,8,10], Goal-oriented CAT[4,6], Tailored CAT[5,10], Positive reappraisal CAT[6], Good listener CAT[2,5,7,10], Acts of kindness CAT[2,4,5,7], Build rapport CAT[7,8,10], Curiosity CAT[7], Grateful CAT[2,5], Sociability CAT[1,2,7], Humanness CAT[3,7,12], Acknowledge CAT[3,4], Interpersonal relationship CAT[2,7,8], Emotionally appropriate CAT[4,8,9,11], Detect emotional states CAT[13], Context-sensitive CAT[5,10,13], Emotion detection CAT[13], Contextual CAT[5], Relationship CAT[7], Fun CAT[1,2], Friendly CAT[1,2,5,7,8], Adaptable to mood CAT[9,10], Show understanding of others emotional state CAT[3,4,11], Concern feelings and experience of others CAT[7,8,11,12], Show understanding CAT[3,4], Show affect CAT[3], “feel for” CAT[3,4], Acknowledge others feelings CAT[3,4], Emotionally expressive CAT[3,4,11], Supportive CAT[4,6,8], Sentiment CAT[3,4], Expressively emotionally supportive CAT[3,4,11], Express/Perceive emotion CAT[3,4,11], Support CAT[4,8], Notice CAT[3,4], Sympathy CAT[2,4,12], Understand/respond to emotions appropriately CAT[4,8,9,11], Understand feelings and experience of others CAT[7,8], Understand statements, behaviors and feelings of another CAT[7,8], Accurately perceive frame of reference of other CAT[4,8,9], Relate and respond accordingly to emotions CAT[4,8,9,11], Social bonding CAT[1,2,6,7], Understanding CAT[12,13], Humorous CAT[1,2], Intelligent CAT[1,2], Personality CAT[1,2], Persuasive CAT[2], Strategic CAT[3], Avoid negative words CAT[8]
Intermediate coding: Categories (CAT, n=13)
1:Empathic traits DIM[A], 2:Empathic acts DIM[A,M], 3:Empathic manifestation DIM[M], 4:Empathic expressions DIM[A,M], 5:Empathic moves/skills DIM[A], 6:Empathic activities DIM[A], 7:Interpersonal DIM[I], 8:Building trust DIM[R], 9:Functional performance DIM[P,T,A,M,R,I], 10:Context DIM[P], 11:Emotions DIM[P,T], 12:Mentalization DIM[T], 13:Perception DIM[P,M]
Advanced coding: Dimensions of Empathy (DIM, n=6)
P:Perceive, A:Act, T:ToM, M:Manifest, R:Ratification, I:Interpersonal

Table 3.3: Data aggregation process. This table presents how codes are clustered into categories and dimensions.

Grounded Theory Outcome

The grounded theory analysis process resulted in a structured hierarchy of codes, categories, and overarching themes. A final step, which was essential for

the computational modeling phase, involved refining the codes and categories. This resulted in a well-defined taxonomy for computational empathy. The highest level of the taxonomy consists of six main concepts: $\{Perceive, Theory\ of\ Mind, Act, Manifest, Ratification, Interpersonal\}$, each comprising a set of associated sub-components. Below, we provide an explanation of the intended meaning of each of these main concepts and their respective elements.

Perceive. *Perceive* = $\{Emotion, Behavior, Mood, Message = \{Body\ language, Facial\ expressions, Presence, Speech, Text\}, Interaction\ property = \{Response\ rate, Time\ between\ sessions, Duration, Response\ content\ level\}\}$. This dimension represents the agent’s ability to observe and interpret external stimuli, which is fundamental for establishing key aspects of empathy and Theory of Mind. The relevance of perceptual abilities varies depending on the nature of the agent and the interaction setting. In certain human-agent interactions, it is crucial to recognize mood, emotion, behavior, and communicative messages. Additionally, in longer-term interactions, factors such as response rate, time between interactions, interaction duration, and the level of content provided gain significance in shaping the agent’s perception.

Theory of Mind. *Theory of Mind* = $\{Agreements, Emotions, Goals, Mood, Needs, Personality, Rapport, Social\ bond, Trust, Values, Errors = \{Deception, Misunderstanding, Misperception\}\}$. This dimension refers to the agent’s ability to infer and understand the mental states of other agents based on perceived information. Theory of Mind encompasses various aspects of another agent’s internal state, including emotions, goals, personality traits, and relational dynamics, such as rapport and trust. Some aspects, such as recognizing values and needs, require more advanced reasoning. Furthermore, this dimension includes the ability to detect misunderstandings, misperceptions, and deceptive behaviors. This capacity relates to cognitive and reconstructive processes in empathy theories [122].

Act. *Act* = $\{Emotion, Behavior, Mood, Context, Message = \{Body\ language, Facial\ expressions, Presence, Speech, Text\}, Interaction\ property = \{Response\ rate, Time\ between\ sessions, Duration, Response\ content\ level\}\}$. A crucial aspect of empathy is an agent’s ability to generate responses that are perceived by others, shaping social bonds. The agent’s actions should align with its understanding of the other agent’s mental state (Theory of Mind). Demonstrating understanding and emotional responsiveness is central to effective empathic interaction. In human-agent interactions, expressions of mood, emotion, and communicative messages are essential. Additionally, as interactions extend over time, aspects such as response rate, session frequency, interaction duration, and content complexity become relevant factors in shaping engagement and interaction quality.

Manifest. *Manifest* = {*Agreements, Emotions, Goals, Mood, Needs, Personality, Rapport, Social bond, Trust, Values, Errors* = {*Deception, Misunderstanding, Misperception*}}. Manifestation refers to how an agent’s internal state is influenced by the situation of other agents. This internal adaptation is a core aspect of mentalizing and acquiring a Theory of Mind. The agent’s internal states are reflected in its actions, forming a feedback loop in human-agent interactions. Through this process, an agent develops a consistent behavioral pattern, akin to a personality, which evolves based on its interaction history. The ability to internalize and adapt to another agent’s state allows for deeper, ongoing interpersonal interactions where agents mutually influence each other’s mental and emotional states. This dimension is closely associated with affective and mirroring processes in empathy theories [122].

Ratification. *Ratification* = {*Acceptance, Agreements, Equality, Mutuality*}. This dimension pertains to an agent’s ability to foster consensus and mutual acceptance in interactions. Ratification is underpinned by two key aspects: *equality* and *mutuality*. Equality ensures that both agents have sufficient opportunities to communicate and interpret each other’s responses, promoting a balanced interaction. Mutuality emphasizes positive communication, fostering trust and understanding while minimizing errors such as misunderstandings, misperceptions, and deception. The interplay between these components enables agents to reach shared agreements and establish tolerance toward each other. Ratification supports trust-building, increasing the robustness of long-term interactions.

Interpersonal. *Interpersonal* = {*Social bonding, Tolerance, Trust building, Interconnection* = {*Connected emotion, Connected mood, Connected behavior, Connected message*}}. A fundamental aspect of empathy is interpersonal interaction, which contributes to trust-building. Throughout an ongoing interaction, agents influence each other’s behaviors and internal states. In human-agent interaction scenarios, fostering interpersonal connections is essential for enhancing perceived empathy. The ability to engage in mutual exchanges of emotions, moods, behaviors, and information supports the development of trust, tolerance, and social bonds. Trust-building relies on key principles such as consistency, honesty, transparency, and commitment to agreed-upon behaviors.

This section has outlined how our methodological approach supports a multi-dimensional understanding of empathy. The empirical definition of these six dimensions informs our formal framework for computational empathy, which is presented in the following section.

No.	Coded definitions	Source
1	"A high performing ML model would become a necessity when conversation volumes increase to ensure high user engagement [ENGAGING] and retention [RELATIONAL]."	Inkster et al., p. 9
2	"understanding how our feelings and emotions work. [EMOTIONAL SELF-AWARENESS]"	Gabrielli et al., p. 3
3	"[...] understand the statements, behaviors or feelings of another person, from the counterpart's perspective or preconditions. [EXPRESS/PERCEIVE EMOTIONS]"	Kraus et al., p. 358
4	"providing an emotionally appropriate response [...], similar to what happens in a successful human-human interaction [...] [EMOTIONALLY APPROPRIATE]"	Ghandeharioun et al., p. 13
5	"the empathizing agent communicates his or her understanding of the other individual's emotional state [SHOW UNDERSTANDING OF OTHERS EMOTIONAL STATE]"	de Genaro et al., p. 3
6	"[...] users' perception of Abot's human-like qualities and affective abilities and acceptability as a chat companion [SHOW UNDERSTANDING]"	Sia et al., p. 39
7	"Each message sequence begins with a warm greeting, in which the chatbot enquires about the participant's mood and replies in an empathic way [SHOW AFFECT]"	Hauser-Ulrich et al., p. 5
8	"However, agents with sophisticated empathic abilities (ie, agents that seem to truly understand the user's emotional experience) [ACCURATELY PERCEIVE FRAME OF REFERENCE OF OTHER]"	Morris et al., p. 2
9	"[...] expected the chatbot to retain context [CONTEXT-SENSITIVE] across chat sessions, thus providing users with personalized [PERSONALIZED] recommendations [...]"	Jain et al., p. 901
10	"sentiment [SENTIMENT] and emotion detection [EMOTION DETECTION] technique may be utilized to identify the corresponding [...]"	Rahman et al., p. 13
11	"An empathic human-like [HUMAN-LIKE] chatbot avatar promoting natural conversations can promote user engagement [ENGAGING] towards chatbot services."	Chen et al., p. 222
12	"empathy is the ability to understand[UNDERSTAND] and concern the feelings and experience of others [CONCERN FEELINGS AND EXPERIENCES OF OTHERS]"	Li et al., p. 2041
13	"Emotional support - Expressions of empathy [EMOTIONALLY EXPRESSIVE], love, trust, and caring [CURIOUS]"	Brandtzaeg et al., p. 6
14	"older users dealing with anxiety could potentially benefit from spare-time interactions with technologies that provide engaging interactions. [SOCIAL BONDING] [SOCIAL BONDING]"	Ryu et al., p. 20
15	"Compassion and empathy are associated, but are not the same ["FEEL FOR"]. Empathy allows people to relate to others suffering [SUFFERING] cognitively and affectively"	Lee et al., p. 8
16	"[...] giving compassion to Vincent (or another being) than towards oneself may be more natural [NATURAL, NATURALLY UNDERSTAND, TONE-AWARE] in conversational contexts"	Lee et al., p. 8

Table 3.4: Part I: Extracted coding definitions.

No.	Coded definitions	Source
17	"Self-defeating [SELF-DEFEATING] humour [HUMOUR] is characterized by an excessive use of self disparaging humour, by which the user attempts to amuse others at their own expense."	Ceha et al., p. 4
18	"the ability to detect another person's emotional state. [EMOTION DETECTION] [...] relate to these emotions and respond according to how the other person must be feeling [RESPOND AND RELATE ACCORDINGLY]"	Casas et al., p. 2
19	"asked participants to what extent they agreed with these statements: "Mitsuku said the right thing to make me feel better," [CONTEXTUAL] "Mitsuku responded appropriately to my feelings and emotions," "Mitsuku came across as empathic," "Mitsuku said the right thing at the right time," [TIMING] and "Mitsuku was a good listener." [GOOD LISTENER] [GOAL-ORIENTED]"	Croes et al., 287
20	"Empathetic chatbots are conversational agents that can understand user emotions and respond appropriately, which is an essential step toward human-like conversation. " [UNDERSTAND/RESPOND TO EMOTIONS APPROPRIATELY] "[...] humans express and perceive emotion [EXPRESS/PERCEIVE EMOTION in natural language to increase their sense of social bonding. [SOCIAL BONDING]"	Lin et al., p. 1
21	"as empathic responses based on user's mood [MOOD], tailored [TAILORED] content based on user's previous inputs, daily check-ins to create a sense of accountability, and weekly summaries in the end of each week containing [...]"	Ly et al., p. 42
22	"express empathy to humans in a way that they perceive [PERCEP] it is natural. [NATURAL] [NATURALLY UNDERSTAND]"	Medeiros et al., p. 234
23	"Because as weird as it is talking to a robot [SAVORING], it's nice to vent and be able to see [NOTICE] others with cancer talking and speaking out [ACKNOWLEDGE] about how they coped or felt during their treatment. Seeing that I'm not alone and having someone guide me to find the positives [ACTS OF KINDNESS] in my life now is really helpful. [GRATEFUL]"	Greer et al., p. 6
24	"Ash is the guide, it is inspiring and charismatic [CHARISMATIC], fun [FUN], friendly [FRIENDLY], empathic, humorous [HUMOUR] [PERSONALITY] [MOOD] [ADAPTABLE TO MOOD]"	Grové et al., p. 7
25	"The 7 categories are: Answering, Error management, Intelligence, Navigation, Onboarding, Personality [PERSONALITY] and Understanding [UNDERSTANDING] [INTELLIGENCE]"	Cameron et al., p. 126
26	"A lack of humanity [LACK OF HUMANITY] or empathy, e.g. 'robotism', 'coldness' [COLDNESS] or 'one-way interaction' was mentioned"	Maeda et al., p. 1139
27	"the use of a pictorial character within conversation text appeared to be a useful "shortcut" for building rapport [BUILD RAPPORT] with users."	Beilharz et al., p. 9
28	"Todaki uses dialogue skills such as empathic responding and reflection [APPRECIATION] [CURIOSITY] to create a therapeutic relationship with the users. [RELATIONSHIP] [INTERPERSONAL RELATION]"	Jang et al., p. 4
29	"explained the need for partner support in a friendly tone [AVOID NEGATIVE WORDS][SUPPORTIVE] and delivered practical strategies with relevant images in which a man actively supported [SUPPORT] his partner, showing empathic concerns and sympathetic responses"	Chung et al., p. 6

Table 3.5: Part II: Extracted coding definitions.

3.5 Formal Definition of Empathy

Following the grounded theory results, the identified empathy dimensions are formalized in terms of an ontology, defining the concept we call "computational empathy". The computational empathy ontology is a representation, and we expect perceptions of an agent's empathy to be represented there. By aggregating

gating the perceptions through the ontology, we get *performance measures* of an agent in terms of computational empathy dimensions.

The proposed ontology is expressed following the syntax of the description logic \mathcal{ALC} (Attribute Language with general Complement) [20] (refer to Table 2.1). The ontology is organized into a taxonomy, where general and abstract concepts are positioned at a higher level, complemented by more specialized and less abstract concepts at a lower level. This structure creates a hierarchy of classes, outlining conditions for empathy across different levels of understanding. At the first level, six minimal necessary conditions for empathy are defined. Progressing to the second level, there are application-specific dimensions, each holding varying relevance in specific settings. The third level further refines these dimensions, incorporating specialized and more directly or indirectly measurable concepts. These concepts can be nuanced based on specific theories, such as cognitive theories. In this proposed ontology, while we consider the first level to outline the necessary conditions, the second and third levels of the taxonomy introduce sub-concepts. It is important to note that these sub-concepts are not universally necessary; their relevance may vary across different contexts and for different agents.

In order to enable interoperability with a wide range of knowledge-based systems, such as on the semantic web, the ontology is implemented in Web Ontology Language (OWL) (see Figure 3.3). This enables integration of external ontologies to define sub-concepts of computational empathy. This is a motivation for maintaining a hierarchical structure in the ontology, making intended meanings of concepts less abstract on a low level, supporting conceptual alignment [278, 316, 159] with other ontologies.

Taxonomy of Computational Empathy

Given the overall concept of *Computational_Empathy* (referred to as Level 0), comprised of parts, so-called *Empathy_dimension* \sqsubseteq *Computational_Empathy*, the model states that the necessary conditions for computational empathy is a set of 6 main dimensions (referred to as Level 1), formally specified as:

Perceive \sqsubseteq *Empathy_dimension*
Act \sqsubseteq *Empathy_dimension*
Manifest \sqsubseteq *Empathy_dimension*
Theory_of_mind \sqsubseteq *Empathy_dimension*
Ratification \sqsubseteq *Empathy_dimension*
Interpersonal \sqsubseteq *Empathy_dimension*

The main dimensions are comprised of sets of sub-dimensions (Level 2).

Perceive_Behavior \sqsubseteq *Perceive*
Perceive_Context \sqsubseteq *Perceive*
Perceive_Emotion \sqsubseteq *Perceive*

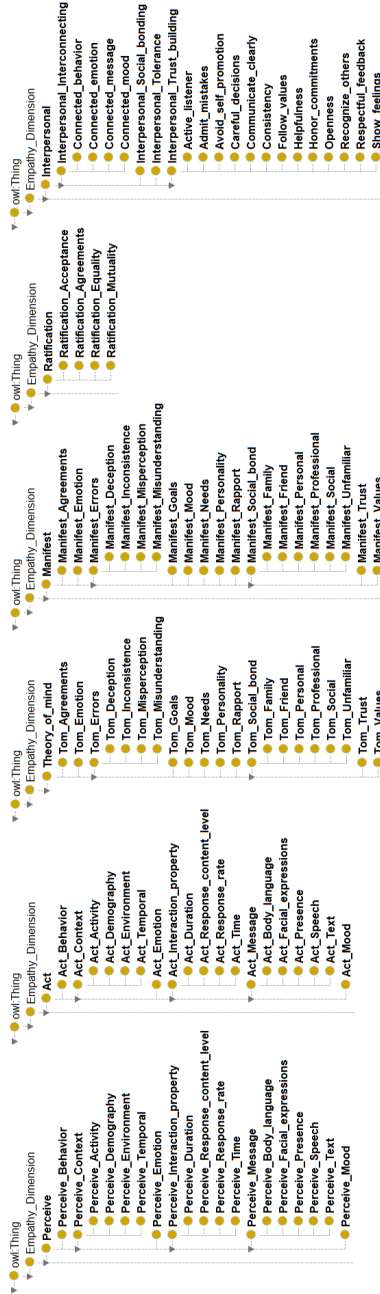


Figure 3.3: Empathy ontology implemented in OWL format.

Perceive_Interaction_property \sqsubseteq *Perceive*
Perceive_Message \sqsubseteq *Perceive*
Perceive_Mood \sqsubseteq *Perceive*

Act_Behavior \sqsubseteq *Act*
Act_Context \sqsubseteq *Act*
Act_Emotion \sqsubseteq *Act*
Act_Interaction_property \sqsubseteq *Act*
Act_Message \sqsubseteq *Act*
Act_Mood \sqsubseteq *Act*

Manifest_Agreements \sqsubseteq *Manifest*
Manifest_Emotion \sqsubseteq *Manifest*
Manifest_Errors \sqsubseteq *Manifest*
Manifest_Goals \sqsubseteq *Manifest*
Manifest_Mood \sqsubseteq *Manifest*
Manifest_Needs \sqsubseteq *Manifest*
Manifest_Personality \sqsubseteq *Manifest*
Manifest_Rapport \sqsubseteq *Manifest*
Manifest_Social_bond \sqsubseteq *Manifest*
Manifest_Trust \sqsubseteq *Manifest*
Manifest_Values \sqsubseteq *Manifest*

Tom_Agreements \sqsubseteq *Theory_of_mind*
Tom_Emotion \sqsubseteq *Theory_of_mind*
Tom_Errors \sqsubseteq *Theory_of_mind*
Tom_Goals \sqsubseteq *Theory_of_mind*
Tom_Mood \sqsubseteq *Theory_of_mind*
Tom_Needs \sqsubseteq *Theory_of_mind*
Tom_Personality \sqsubseteq *Theory_of_mind*
Tom_Rapport \sqsubseteq *Theory_of_mind*
Tom_Social_bond \sqsubseteq *Theory_of_mind*
Tom_Trust \sqsubseteq *Theory_of_mind*
Tom_Values \sqsubseteq *Theory_of_mind*

Ratification_Acceptance \sqsubseteq *Ratification*
Ratification_Agreements \sqsubseteq *Ratification*
Ratification_Equality \sqsubseteq *Ratification*
Ratification_Mutuality \sqsubseteq *Ratification*

Interpersonal_Interconnection \sqsubseteq *Interpersonal*
Interpersonal_Social_Bonding \sqsubseteq *Interpersonal*
Interpersonal_Tolerance \sqsubseteq *Interpersonal*
Interpersonal_Trust_building \sqsubseteq *Interpersonal*

These sub-dimensions are further defined by their sub-dimensions (referred to as Level 3). For instance, examples of sub-dimensions of *Perceive_Message* \sqsubseteq *Perceive* are:

Perceive_Body_language \sqsubseteq *Perceive_Message*
Perceive_Facial_Expressions \sqsubseteq *Perceive_Message*
Perceive_Presence \sqsubseteq *Perceive_Message*
Perceive_Speech \sqsubseteq *Perceive_Message*
Perceive_Text \sqsubseteq *Perceive_Message*

The ontology’s hierarchical structure provides explicit definitions of concepts at lower levels, enhancing assessments by enabling more objectively understood concepts. This promotes interoperability and conceptual alignment between systems that require a shared understanding of empathy concepts. Furthermore, the hierarchical structure aids in human-system conceptual alignment. When gathering a human’s perceptions about empathy concepts, it is crucial that the human’s understanding aligns with the intended meaning of the ontology. Therefore, the ontology must offer precise and unambiguous means for comprehending observations, benefiting both systems and humans. Through the definition of concepts in this hierarchical structure, the ontology facilitates both system-to-system and *human-to-system conceptual alignment*.

Given a set of subjective perceptions from different sources, we assume that the dimensions of computational empathy regard imprecise perceptions, which define intervals between an upper bound and a lower bound. In order to capture imprecise perceptions, for each dimension and sub-dimension, we assign properties for imprecise uncertainty intervals. In this way, the ontology allows us to represent and reason about *empathic performance* in terms of imprecise measures.

Imprecise Uncertainty Intervals

Imprecise perceptions are captured in the model in terms of imprecise uncertainty intervals [93], which values can be obtained from sets of subjective perceptions of an interactive agent, consisting of an upper bound (most optimistic perception) and a lower bound (most skeptical perception). Hence, for each empathy dimension, given a scale $S \in [0, 1]$, there is an imprecise uncertainty interval I between an upper bound $t \in [0, 1]$ and a lower bound $b \in [0, 1]$, where $t \geq b$. This provides an imprecise measure of performance based on the interval I , where the width of I represents the level of agreement or accuracy in the measurements. An interval with both upper and lower bounds closer to 1.0 is interpreted as a higher performance measure, such as an accurately perceived emotion. This representation allows for the analysis of sub-dimensions in terms of *performance* and *accuracy*. Furthermore, through the ontology’s taxonomy, intervals of sub-dimensions can be aggregated into higher-level dimensions, enabling reasoning about performance across the six

main dimensions.

We implement the intervals by defining a set of classes for managing weighted concepts:

- Class *Uncertainty_interval* - A concept for describing imprecise uncertainty for empathy dimensions.
Properties: *hasLowerBound*, *hasUpperBound*
Sub class of: *EmPAT:Concept*
- Data Property *hasUpperBound* - An imprecise uncertainty interval has an upper bound.
Domain: *Uncertainty_interval*
Range: xsd:decimal
- Data Property: *hasLowerBound* - An imprecise uncertainty interval has a lower bound.
Domain: *Uncertainty_interval*
Range: xsd:decimal
- Object Property: *hasInterval* - An empathy dimension has an imprecise uncertainty interval.
Domain: *Empathy_dimension*
Range: *Uncertainty_interval*

With the ontology’s taxonomy as a platform, it is essential to capture interdependencies among concepts (semantic relations) to facilitate advanced reasoning on empathy observations. This necessity arises from the inherent challenge of being unable to directly observe an agent’s internal empathic processes. As an example, the model must capture relations between “observable concepts,” such as *Act*, and “internal concepts,” such as *Manifest* and *Theory_of_mind*.

Semantic Interdependencies

By considering theories about empathy, e.g., social cognition [282], the Perception and Action Model (PAM) [233], and the Interpersonal Adaptation Theory (IAT) [54], we can define semantic relations among empathy concepts.

According to the Perception and Action Model (PAM), there is a chain of dependencies between perception, affective and cognitive processes, and action (see a summary in Section 3.2). Let us specify an abstraction of PAM in terms of the proposed six main dimensions of the current work. Perceptions lead to affection of an agent’s mental/affective state (*Manifest*) and are reflected in its understanding of the other/cognitive state (*Theory_of_mind*). An agent’s acting (*Act*) is in turn influenced by the affective processes and tailored by cognitive processes (*Manifest* and *Theory_of_mind*).

In order to capture interrelations between mental and social processes, we further look at principles from social cognition [282], where empathy has

been considered an important construct for interconnected and shared emotion (see a summary in Section 3.2). Social cognition, in terms of the proposed model, is an abstraction where *Manifest enables Interpersonal* (e.g., *Interconnected Emotions*), and *Theory_of_mind* is required for *Ratification*, where an agent's *Act* is *forming Ratification* (e.g., reaching social *Acceptance* and *Agreements*). An agent's behavior has further relevance for adapting the *Interpersonal* relation, explained by principles from the Interpersonal Adaptation Theory (IAT) [54]. In terms of the proposed model, we specify an abstraction such that *Act* is *adapting Interpersonal*.

Before formalizing these relations, let us begin by specifying some relevant object (semantic) properties:

- Object Property: *isAffecting* - Perceive is affecting manifest.
Domain: *Perceive*
Range: *Manifest*
- Object Property: *isReflectedIn* - Perceive is reflected in theory of mind.
Domain: *Perceive*
Range: *Theory_of_mind*
- Object Property: *isEnabling* - Manifest is enabling interpersonal capability.
Domain: *Manifest*
Range: *Interpersonal*
- Object Property: *isRequiredFor* - Theory of mind is required for ratification.
Domain: *Theory_of_mind*
Range: *Ratification*
- Object Property: *isInfluencing* - Manifest is influencing an agent's acts.
Domain: *Manifest*
Range: *Act*
- Object Property: *isTailoring* - Theory of mind is tailoring an agent's acts.
Domain: *Theory_of_mind*
Range: *Act*
- Object Property: *isForming* - Act is forming ratification.
Domain: *Act*
Range: *Ratification*
- Object Property: *isAdapting* - Act is adapting Interpersonal.
Domain: *Act*
Range: *Interpersonal*

The chain of dependencies can be formalized as follows.

- (1) **Perceive *isAffecting* Manifest.** An agent's perception *is affecting* the agent's manifest. For instance, a perceived emotion will affect the agent's emotion model. Hence, we say that a concept c that is perceived ($c \sqsubseteq$

Perceive) will also have a representation in manifest ($c \sqsubseteq \text{Manifest}$). This is specified by the relation *isAffecting*, formally defined as:

$$(1a) \exists \text{isAffecting}.\text{Manifest} \sqsubseteq \text{Perceive} \sqcap \text{Manifest}$$

By considering this dependency between *Perceive* and *Manifest*, we say that an agent has an *Empathic Manifest* if *isAffecting.Manifest* holds for all perceptions:

$$(1b) \text{EmpathicManifest} \sqsubseteq \text{Perceive} \sqcap \text{Manifest} \sqcap \forall \text{isAffecting}.\text{Manifest}$$

- (2) **Perceive *isReflectedIn* Theory_of_mind.** An agent's perception *is reflected in* the agent's theory of mind. For instance, a perceived emotion will be reflected on by the agent. Hence, we say that a concept c that is perceived ($c \sqsubseteq \text{Perceive}$) will also have a representation in theory of mind ($c \sqsubseteq \text{Theory_of_mind}$). This is specified by the relation *isReflectedIn*, formally defined as:

$$(2a) \exists \text{isReflectedIn}.\text{Theory_of_mind} \sqsubseteq \text{Perceive} \sqcap \text{Theory_of_mind}$$

By considering this dependency between *Perceive* and *Theory of mind*, we say that an agent has an *Empathic ToM* if *isReflectedIn.Theory_of_mind* holds for all perceptions:

$$(2b) \text{EmpathicToM} \sqsubseteq \text{Perceive} \sqcap \text{Theory_of_mind} \sqcap \forall \text{isReflectedIn}.\text{Theory_of_mind}$$

- (3) **Manifest *isEnabling* Interpersonal.** An agent's manifest *is enabling* the agent's capability for interpersonal relation, where, e.g., an agent's manifested emotions may affect the other. Hence, we say that if an agent has manifested a concept c ($c \sqsubseteq \text{Manifest}$) such that *Empathic Manifest* holds, it will also have a representation in *Interpersonal* ($c \sqsubseteq \text{Interpersonal}$), e.g., in the sub-dimension *Interconnected.Emotion*. This is specified by the relation *isEnabling*, formally defined as:

$$(3a) \exists \text{isEnabling}.\text{Interpersonal} \sqsubseteq \text{EmpathicManifest} \sqcap \text{Interpersonal}$$

By considering this dependency between *Manifest* and *Interpersonal*, we say that an agent has an *Empathic Interpersonal* if *isEnabling.Interpersonal* holds for all *EmpathicManifest*:

$$(3b) \text{EmpathicInterpersonal} \sqsubseteq \text{EmpathicManifest} \sqcap \text{Interpersonal} \sqcap \forall \text{isAffecting}.\text{Interpersonal}$$

- (4) **Theory_of_mind *isRequiredFor* Ratification.** An agent's theory of mind *is required for* the agent's capability for ratification, where, e.g., an

agent's theory of trust is central for reaching agreements. Hence, we say that if an agent has theory of mind of a concept c ($c \sqsubseteq Theory_of_mind$) such that *EmpathicToM* holds, it will also have a representation in Ratification ($c \sqsubseteq Ratification$), e.g., in the sub-dimensions of *Trust* and *Agreement*, respectively. This is specified by the relation *isRequiredFor*, formally defined as:

$$(4a) \exists isRequiredFor.Ratification \sqsubseteq Theory_of_mind \sqcap Ratification$$

By considering this dependency between Theory of mind and Ratification, we say that an agent has an *Empathic Ratification* if the relation *isRequiredFor.Interpersonal* holds for all *EmpathicToM*:

$$(4b) EmpathicRatification \sqsubseteq EmpathicToM \sqcap Ratification \sqcap \forall isRequiredFor.Ratification$$

- (5) **Manifest *isInfluencing* Act.** An agent's manifest *is influencing* the agent's act. For instance, a manifested emotion will influence the agent's emotional acting. Hence, we say that a concept c that is manifested ($c \sqsubseteq Manifest$) will also have a representation in act ($c \sqsubseteq Act$). This is specified by the relation *isInfluencing*, formally defined as:

$$(5a) \exists isAffecting.Act \sqsubseteq Manifest \sqcap Act$$

By considering this dependency between Manifest and Act, we say that an agent has an *Empathic Act* if the relation *isInfluencing.Act* holds for all *EmpathicManifest*:

$$(5b) EmpathicAct \sqsubseteq EmpathicManifest \sqcap Act \sqcap \forall isInfluencing.Act$$

- (6) **Theory_of_mind *isTailoring* Act.** An agent's Theory of mind *is tailoring* the agent's act. For instance, a theory of another agent's emotion will tailor the agent's emotional acting. Hence, we say that a concept c that is in theory of mind ($c \sqsubseteq Theory_of_mind$) will also have a representation in act ($c \sqsubseteq Act$). This is specified by the relation *isTailoring*, formally defined as:

$$(6a) \exists isTailoring.Act \sqsubseteq EmpathicToM \sqcap Act$$

By considering this dependency between Theory of mind and Act, we say that an agent has an *Empathic Act* if the relation *isTailoring.Act* holds for all *EmpathicToM*:

$$(6b) EmpathicAct \sqsubseteq EmpathicToM \sqcap Act \sqcap \forall isTailoring.Act$$

- (7) **Act *isForming* Ratification.** An agent's Act *is forming* the agent's ratification with other agents. For instance, the agent's emotional acting will form emotional acceptance. Hence, we say that a concept c that is in an empathic act ($c \sqsubseteq EmpathicAct$) will also have a representation in ratification ($c \sqsubseteq Ratification$). This is specified by the relation *isForming*, formally defined as:

(7a) $\exists \text{isForming.Ratification} \sqsubseteq \text{EmpathicAct} \sqcap \text{Ratification}$

By considering this dependency between *EmpathicAct* and *Ratification*, we say that an agent has an *Empathic Ratification* if the relation *isForming.Ratification* holds for all *EmpathicAct*:

(7b) $\text{EmpathicRatification} \sqsubseteq \text{EmpathicAct} \sqcap \text{Ratification} \sqcap \forall \text{isForming.Ratification}$

- (7) **Act *isAdapting* Interpersonal.** An agent’s Act *is adapting* the agent’s interpersonal relation with other agents. For instance, the agent’s emotional acting will adapt the emotional connection between agents. Hence, we say that a concept *c* that is in an empathic act ($c \sqsubseteq \text{EmpathicAct}$) will also have a representation in interpersonal ($c \sqsubseteq \text{Interpersonal}$). This is specified by the relation *isAdapting*, formally defined as:

(8a) $\exists \text{isAdapting.Interpersonal} \sqsubseteq \text{EmpathicAct} \sqcap \text{Interpersonal}$

By considering this dependency between *EmpathicAct* and *Interpersonal*, we say that an agent has an *Empathic Interpersonal* if the relation *isAdapting.Interpersonal* holds for all *EmpathicAct*:

(8b) $\text{EmpathicInterpersonal} \sqsubseteq \text{EmpathicAct} \sqcap \text{Interpersonal} \sqcap \forall \text{isAdapting.Interpersonal}$

The above specifications allow a semantic understanding of the defined concepts of computational empathy. The semantic interdependencies are important for reasoning about concepts that may not be directly observed nor inferred through the taxonomy. The semantic interdependencies go beyond the taxonomy to reason about an agent’s perceived empathic capability. Such perceptions can be transformed into DL queries, provided as input to the ontology to give further explanations of the perceptions.

Examples of Reasoning Queries

These semantic interdependencies enable the ontology to be applied for reasoning about abstract empathy concepts. A collection of low-level perceptions of the interaction allows conclusions to be drawn about how an agent’s empathic capabilities are perceived. For instance, given the perception: “an agent seems to be sensing and acting by considering emotions,” the following DL query can be created: “Perceive.Emotion \sqcap Act.Emotion \sqcap isAffecting.Act”. Given this query, we can, by considering the above chain of semantic interdependencies, infer “EmpathicManifest \sqcap EmpathicToM,” estimating that the agent was perceived to be mentally affected by, and to reason about, emotions. Given this assertion, we can advance with further queries, such as: “EmpathicManifest \sqcap EmpathicToM” to infer “EmpathicAct \sqcap (isInfluencing.Act \sqcup isTailoring.Act),” estimating that an agent was perceived to act empathically.

By aggregating the associated uncertainty intervals of asserted concepts, explicit performance measures can be estimated. For instance, measured uncertainty intervals of observed acts, e.g., Act.Emotion and Act.Context, with

upper bounds $\in [0, 1]$ and lower bounds $\in [0, 1]$, are aggregated to corresponding uncertainty intervals of higher-level concepts, e.g., Act, estimating the performance of abstract computational empathy concepts. The interdependencies further allow for the inference of aggregated uncertainty intervals of semantically related concepts, e.g., interpersonal and ratification.

In this section, we have formally defined a model for computational empathy. A taxonomy is informed by a grounded theory study based on data collected in a literature review of empathy. The interdependencies between concepts, for reasoning about ramification effects, are informed by prior empathy theories (e.g., PAM [233] and social cognition [282]). In the following section, we present the process and results of a user study where we apply the model as a tool for reasoning about perceptions of empathy in two state-of-the-art chatbots: Replika and Wysa.

3.6 Perceptions of Chatbot Empathy

Our primary goal is to establish a broad, versatile definition of computational empathy that can be tailored to specific applications as needed. In the following example, applying the proposed model of computational empathy as a tool for assessing empathy in chatbots, we illustrate how we can uncover trends in changing perceptions of computational empathy over time. By correlating these trends with low-level concepts within the ontology, detailed insights into observed changes can be provided.

In order to enable data collection for assessing empathy in chatbots, an assessment protocol (in the form of a questionnaire) is developed, consisting of Likert scale (1–5) questions designed to collect information corresponding to the dimensions of computational empathy. A user study was conducted using the tool, in which the participants interacted with two state-of-the-art chatbots in the area of health and well-being (Replika and Wysa). The user study collected the users’ perceptions of the chatbots’ interaction capabilities in terms of the lowest-level dimensions (Level 3; the most trivial concepts regarding understanding and measuring) of the empathy ontology. The answers were normalized ($[0,1]$) and clustered on different levels in the ontology to derive qualitative *empathy measures*. The measures were analyzed on an aggregated level, in terms of the six main dimensions, as well as on a detailed level, considering each sub-dimension’s contribution.

Recall that the ontology is structured according to a taxonomy where general, more abstract, concepts are at a high level and more specialized, less abstract, concepts are at a low level. By designing the questionnaire questions based on low-level concepts in the ontology, the questionnaire acts as an interface between the ontology and the users’ perceptions. The six high-level dimensions of the ontology (Level 1), such as “Theory of Mind” and “Perceive”, are abstract concepts that are difficult to intuitively understand for a participant. An answer to such a high-level question would be based on the

participant’s subjective meaning of the concept, which may differ from what is intended in the questionnaire. A level down the taxonomy (Level 2), more understandable, intermediately abstract, concepts are reached, such as “Emotion” and “Message”, which still may be subjectively understood. By continuing further down the taxonomy (Level 3), we reach concepts such as “Response rate”, “Facial expressions”, and “Body language”, which are more explicit and more intuitively understood by a participant. This makes it easier to map the intended meaning of the concepts to questions in the questionnaire. As a result, the answers are less subjective, making them more trivial to measure and analyze. This method of subdividing concepts to reach more trivial concepts can be applied at any number of levels until a sufficient formality is reached. This study stopped at Level 3 to balance formality and detail.

User Study Design

The goal of the study is to illustrate the capability of the model to describe users’ perceptions of computational empathy in different chatbots (which have different interaction capabilities). Preferably, when conducting such a study, we would have a configurable chatbot that could enable/disable different empathic features (e.g., actuators and sensors) to create different experimental conditions. However, let us mention that there are no models in the state of the art to configure an empathic chatbot. Alternatively, by designing the study using two different chatbots, Wysa and Replika (see Figure 3.4)^{5,6}, with a diverse set of features, we create different experimental conditions, whose empathetic features can be traced back to design decisions of the chatbots (see Table 3.7).

The chatbot features are selected based on two classes: *Actuators* (social actions) and *Sensors* (social perceptions). These are aspects that we can objectively identify in the chatbots, which otherwise can be understood as “black-box” systems. Given these features, we assess *performance measures* in terms of the dimensions of computational empathy. We can understand the chatbot features as objective truths about the chatbots. For instance, Replika has an actuator in terms of facial expressions, whereas Wysa does not. Hence, some performance measures for sub-dimensions, such as *Act_Facial_expressions*, are conceptually and functionally linked to this feature. This allows us to explain the participants’ answers more objectively as well as trace aspects of empathy to explicit design features of the chatbots.

We distinguish our experimental conditions (the two chatbots) by conducting human–chatbot interaction “in the wild”, providing a realistic setting for assessing users’ perceptions of chatbot empathy.

Given the relations between chatbot features and empathy dimensions (Table 3.7), by considering actuators and sensors for Wysa in comparison to Replika, we can specify a set of hypotheses for the chatbots’ performances, and

⁵Press kit Wysa: <https://www.wysa.io/media>

⁶Press kit Replika: <https://replika.com/about/press>

by considering the interdependencies defined in the semantic model, we can specify an argument chain for each hypothesis:

- Hypothesis 1 (H1): Measures of **Act** on day 3 and 7 will be lower for Wysa than Replika
 - * Argument for H1: Wysa lacks actuator A1, A3, A4, A5, which limit Act
 - * Argument for H1: Wysa lacks sensor S2, S3, S4, which limit Perceive
 - * Argument for H1: Perceive is affecting Manifest
 - * Argument for H1: Perceive is reflected in Theory of mind
 - * Argument for H1: Manifest is influencing Act
 - * Argument for H1: Theory of mind is tailoring Act

- Hypothesis 2 (H2): Measures of **Manifest** on day 3 and 7 will be lower for Wysa than Replika
 - * Argument for H2: Wysa lacks sensor S2, S3, S4, which limit Perceive
 - * Argument for H2: Perceive is affecting Manifest

- Hypothesis 3 (H3): Measures of **Perceive** on day 3 and 7 will be lower for Wysa than Replika
 - * Argument for H3: Wysa lacks sensor S2, S3, S4, which limit Perceive

- Hypothesis 4 (H4): Measures of **Theory of mind** on day 3 and 7 will be lower for Wysa than Replika
 - * Argument for H4: Wysa lacks sensor S2, S3, S4, which limit Perceive
 - * Argument for H4: Perceive is reflected in Theory of mind

- Hypothesis 5 (H5): Measures of **Interpersonal** between day 3 and 7 will be lower for Wysa than Replika
 - * Argument for H5: Wysa lacks actuator A1, A3, A4, A5, which limit Act
 - * Argument for H5: Wysa lacks sensor S2, S3, S4, which limit Perceive
 - * Argument for H5: Perceive is affecting Manifest
 - * Argument for H5: Manifest is enabling Interpersonal
 - * Argument for H5: Act is adapting Interpersonal

- Hypothesis 6 (H6): Measures of **Ratification** on day 3 and 7 will be lower for Wysa than Replika
 - * Argument for H6: Wysa lacks actuator A1, A3, A4, A5, which limit Act
 - * Argument for H6: Wysa lacks sensor S2, S3, S4, which limit Perceive
 - * Argument for H6: Perceive is reflected in Theory of mind
 - * Argument for H6: Theory of mind is required for Ratification
 - * Argument for H6: Act is forming Ratification

Participant	Age	Gender	Study Area	Chatbot
W21	35	M	Computing Science	Wysa
W25	25	F	Cognitive science	Wysa
W27	33	F	Cognitive science	Wysa
W29	29	M	Cognitive science	Wysa
W30	26	M	Computing Science	Wysa
W36	29	M	Philosophy	Wysa
R22	41	M	Computing Science	Replika
R23	29	F	Digital Design	Replika
R24	32	F	Cognitive Science	Replika
R31	36	M	Law	Replika
R33	28	M	Healthcare	Replika
R35	33	M	Psychology	Replika
R37	60+	M	Engineering	Replika

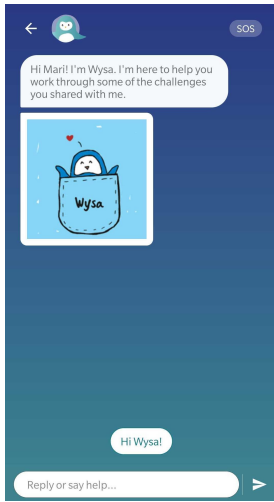
Table 3.6: Participants of the user study.

Selection of participants. The selection aimed for university students, below the age of 45 (young adults) and with minimal prior chatbot experience. In total, 13 participants were selected. The participant group consisted mostly of current or prior university students in a variety of study areas (digital design, philosophy, cognitive science, computing science, law, healthcare, and engineering). Given the similarity in nature between social media chats and chatbot interactions, we collected the participants’ social media chat experience in addition to prior chatbot experience. This was done in order to estimate a sufficient capability among participants for handling a chat interface. The participants generally described their prior experience with chatbot interactions as low to medium and social media interactions as medium to high (see Table 3.6).

Procedure. We clustered the 13 participants into two groups, one to interact with Replika and the other to interact with Wysa. The participants were asked to interact with the chatbot for about 10 minutes, once per day, for a week. On two occasions (on day three, $N = 13$, and day seven, $N = 12$), the participants were asked to answer an online questionnaire about their interaction experience. The two rounds of the questionnaire were conducted in order to collect a measure of how a long-term interaction affects the empathy scores.

We clarified that the participants would decide what to share with the chatbot and that we would not ask to see the chat log. Making the interaction private was important in order to promote participants’ comfort in choosing discussion topics with the chatbot.

Assessment protocol. The protocol consists of Likert scale questions in a 1–5 range (see Table 3.8 and Table 3.9), covering the lower-level dimensions of empathy. The 1–5 alternatives on each question represent levels



(a) Wysa



(b) Replika

Figure 3.4: Chatbots constitute different experimental conditions (figures retrieved from press kits; footnote 5 and 6).

of performance regarding a particular empathic capability, where 1 represents low performance and 5 represents high performance.

Limitations. This user study aims to illustrate trends in changing perceptions of computational empathy over time, presenting a potential application of the ontology. Considering the constraints of our sample size, our intent is to provide a starting point for understanding differences in empathy dimensions. For these particular chatbots, the presented trends should be viewed as preliminary and subject to further investigation.

ID	Feature	Wysa	Replika	Empathy
A1	Actuator: Human appearance		X	Act
A2	Actuator: Text messages	X	X	Act
A3	Actuator: Facial expressions		X	Act
A4	Actuator: Body language		X	Act
A5	Actuator: Affective/emotional expressions		X	Act, Manifest
S1	Sensor: Natural language text messages	X	X	Perceive, ToM
S2	Sensor: Emojis; User writes, e.g., :) :/ :(X	Perceive, ToM
S3	Sensor: Roleplay; e.g., User: *Looks at you*		X	Perceive, ToM
S4	Sensor: Message count; Unlocking traits		X	Perceive, Interpersonal

Table 3.7: Experimental conditions in terms of chatbot features. Features are related to empathy dimensions.

Initial analysis (average)

The 1-5 data range is normalized to values between 0 and 1. In order to get an overview of the data to find overall patterns, an initial analysis was done by calculating average values (arithmetic mean) for each question (sub-dimension), separately for each chatbot. The calculated averages for each sub-dimension

Empathy Dimension	Question (1-5 Likert scale)
Perceive Mood	1. How much did the chatbot perceive your current mood?
Perceive Emotion	2. How much did the chatbot perceive emotions you expressed?
Perceive Behavior	3. How much did the chatbot perceive your behaviour?
Perceive Message	4. How much did the chatbot perceive relevant information?
Perceive Demography	5. How much did the chatbot perceive your age, gender and culture?
Perceive Activity	6. How much did the chatbot perceive your purpose of the dialogue?
Perceive Resonse Rate	7. How much did the chatbot perceive your response rate?
Perceive Time	8. How much did the chatbot perceive time between sessions?
Perceive Duration	9. How much did the chatbot perceive duration of interaction?
Perceive Content	10. How much did the Chatbot perceive the amount of text of your response?
ToM Agreements	11. How much did the chatbot understand established agreements?
ToM Personality	12. How much did the chatbot understand your personality?
ToM Emotion	13. How much did the chatbot understand your emotions?
ToM Rapport	14. How much did the chatbot understand your rapport?
ToM Trust	15. How much did the chatbot understand your trust?
ToM Errors	16. How much did the chatbot understand your errors?
ToM Deception	17. How much did the chatbot understand deception?
ToM Social Bond	18. How much did the chatbot understand what type of relation you had?
ToM Values	24. How much did the chatbot understand your values?
ToM Goals	25. How much did the chatbot understand your goals?
ToM Needs	26. How much did the chatbot understand your needs?
Act Mood	27. How much did the Chatbot express appropriate mood?
Act Emotion	28. How much did the Chatbot express appropriate emotions?
Act Behavior	29. How much did the Chatbot express appropriate behavior?
Act Message	30. How much did the Chatbot express relevant information?
Act Demography	31. How much did the Chatbot act appropriately according to your demography?
Act Activity	32. How much did the Chatbot act in relation to your purpose of the dialogue?
Act Response Rate	33. How much did the Chatbot act in relation to your response rate?
Act Time	34. How much did the Chatbot act in relation to time between sessions?
Act Duration	35. How much did the Chatbot act in relation to the duration of interaction?
Act Response Content	36. How much did the Chatbot act in relation to the size of your responses?
Manifest Emotion	37. How much was the Chatbot's emotions affected by your emotions?
Manifest Personality	38. How much was the Chatbot's personality appropriately affected?
Manifest Trust	40. How much did the Chatbot appear respectful?
Manifest Trust	41. How much did the Chatbot appear responsible?
Manifest Trust	42. How much did the Chatbot appear fair?
Manifest Trust	43. How much did the Chatbot appear honest?
Manifest Behavior	44. How much was the Chatbot's behavior appropriately affected?
Manifest Values	45. How much did the Chatbot show signs of behaving in line with certain values?
Manifest Goals	46. How much did the Chatbot show signs of behaving in line with certain goals?
Manifest Needs	47. How much did the Chatbot show signs of behaving in line with certain needs?

Table 3.8: Part I: Questionnaire questions. Questions are mapped to sub-dimensions of the ontology.

was then further aggregated to calculate an average for each of the 6 main dimensions. Finally, a total average value for each chatbot is calculated. These averages give a rough estimate about the common values in the data set and an approximation of the performance measures (empathy scores) for each chatbot. We now present the initial analysis of average performance measures w.r.t. the dimensions on Day 3 and Day 7. In each stage of the analysis, we trace consistencies between the chatbot features (Table 3.7) and the participants' perceptions of the chatbot interaction.

Empathy Dimension	Question (1-5 Likert scale)
Ratification Agreements	48. How much did the Chatbot and you reach agreements?
Ratification Equality	49. How much did the Chatbot and you have equality?
Ratification Mutuality	50. How much did the Chatbot and you have mutuality?
Ratification Acceptance	51. How much did the Chatbot and you accept each other?
Interpersonal Connected Mood	52. How much did the Chatbot and you share mood?
Interpersonal Connected Emotion	53. How much did the Chatbot and you share emotions?
Interpersonal Connected Behavior	54. How much did the Chatbot and you share behavior?
Interpersonal Connected Message	55. How much did the Chatbot and you share knowledge?
Interpersonal Trust Building	56. How much did the Chatbot and you build trust?
Interpersonal Trust Building	57. The Chatbot honoured commitments
Interpersonal Trust Building	58. The Chatbot communicated effectively and clear
Interpersonal Trust Building	59. The Chatbot made decisions in a careful way
Interpersonal Trust Building	60. The Chatbot behaved in a consistent way
Interpersonal Trust Building	61. The Chatbot is an active listener
Interpersonal Trust Building	62. The Chatbot provided respectful feedback
Interpersonal Trust Building	63. The Chatbot was open for your thoughts and feelings
Interpersonal Trust Building	64. The Chatbot showed feelings
Interpersonal Trust Building	65. The Chatbot was honest
Interpersonal Trust Building	66. The Chatbot was helpful
Interpersonal Trust Building	67. The Chatbot was kind in an genuine way
Interpersonal Trust Building	68. The Chatbot did not promote itself
Interpersonal Trust Building	69. The Chatbot recognised you as a person
Interpersonal Trust Building	70. The Chatbot behaved in line with certain values
Interpersonal Trust Building	71. The Chatbot admitted when it made a mistake
Interpersonal Trust Building	72. How much did you accept the Chatbot despite its flaws?
Interpersonal Trust Building	73. How much did the Chatbot and you build a relation?
Interpersonal Trust Building	74. How much did the Chatbot and you build a prof relation, such as how it would be between doctor and patient?
Interpersonal Social Bond	75. How much did the Chatbot and you build a family relation, such as how it would be between mother and daughter?
Interpersonal Social Bond	76. How much did the Chatbot and you build a friend relation, such as how it would be between two close friends?
Interpersonal Social Bond	77. How much did the Chatbot and you build a personal relation, such as how it would be between two co-workers?
Interpersonal Social Bond	78. How much did the Chatbot and you build a social relation, such as how it would be between individuals at a party?

Table 3.9: Part II: Questionnaire questions. Questions are mapped to sub-dimensions of the ontology.

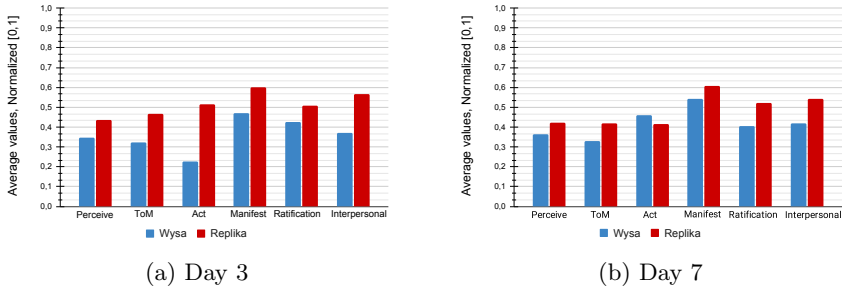


Figure 3.5: Average dimension measures

Average dimension measures on Day 3

On day three (see Figure 3.5a), we can observe overall higher scores for Replika total average 0.52 than for Wysa total average 0.36. Looking at averages for the 6 main dimensions, we can observe the following average differences (comparing Replika/Wysa): Perceive (0.44/0.35), Act (0.52/0.23), ToM (0.47/0.32), Manifest (0.60/0.47) Ratification (0.51/0.43), Interpersonal (0.57/0.37), showing that each of the 6 main dimensions were higher for Replika. This rough estimate is analyzed in detail in the previous analysis on Explaining change.

Average dimension measures on Day 7

On day seven (see Figure 3.5b), we can observe overall higher scores for Replika total average 0.49 than for Wysa total average 0.42. Looking at averages for the 6 main dimensions, we can observe the following average differences (comparing Replika/Wysa): Perceive (0.42/0.36), Act (0.42/0.46), ToM (0.42/0.33), Manifest (0.61/0.54), Ratification (0.52/0.41), Interpersonal (0.54/0.42), showing that most of the 6 main dimensions were still higher for Replika, except for Act that on day 7 is higher for Wysa. Act notably increased for Wysa, from 0.23 (day 3) to 0.46 (day 7), and Act notably decreased for Replika, from 0.52 (day 3) to 0.42 (day 7). This approximation of change shows that the long-term interaction had an impact on the measures. These rough estimates of change between day 3 and day 7 can be explained by analyzing the individual sub-dimensions of e.g., Act between Day 3 and Day 7, finding dominant sub-dimensions that gave rise to the change. The sub-dimensions suggest the participants' initial perceptions (expectations) of the various chatbots (measures on Day 3) which were either met or rejected with further interaction (measures on Day 7). These causes of change are analyzed in detail in the previous analysis on Explaining change.

Analysis of possibilistic intervals

An analysis is done looking at intervals between maximum and minimum values of each dimension. This is done through an imprecise uncertainty measure considering intervals of possibilistic distributions [93]. Possibilistic distributions are a class of fuzzy sets [314], which can be calculated for each dimension to capture qualitative data, making it a practical method to deal with uncertain empathy measures. In order to preprocess the collected data from the questionnaire to create possibilistic distributions, we collected, for each dimension, the most optimistic (highest) answers and calculated the average. Similarly, for each dimension, we collected the most skeptical (lowest) answers and calculated the average, in this way defining an interval for each dimension. Through this approach, we can understand and visualize the data in a more nuanced way, finding a skeptical lowest point and an optimistic highest point for each observation.

Looking at the intervals for Replika and Wysa (see Figure 3.6), we can see that some intervals are more compact in terms of minimum and maximum (a higher agreement/accuracy of the participants' perceptions). We can further observe that the intervals generally became more compact between Day 3 and Day 7, showing that the long-term interaction streamlined the users' conceptions. Hence, the data on Day 7 can be regarded as less subjective.

Average dimension intervals on Day 3

On day three (see Figure 3.6a), we can observe a generally more diverse perspective from the participants, with less compact intervals, and with averages more toward the min/max endpoints. For instance, Wysa's Act (average 0.23) is close to the minimum (0.15) and far from the maximum (0.80). An interesting observation is that most of the dimensions had a higher maximum on day three, compared to day seven. This suggests an initially higher optimism from participants on the chatbots' interaction capabilities.

Average dimension intervals on Day 7

On day seven (see Figure 3.6b), we can observe more compact intervals in most of the dimensions, compared to day three. There are some exceptions, e.g., a higher spread on Replika's Ratification interval on day seven (max: 1.0, min: 0.06), compared to the same dimension on day three (max: 0.88, min: 0.13). The generally more compact intervals at day seven, compared to day three, suggest that the participants had a more common view on the interaction.

This change can be explained in different levels by looking at different sets of dimensions. In the next subsection, we look at each sub-dimension for a detailed analysis and explanation.

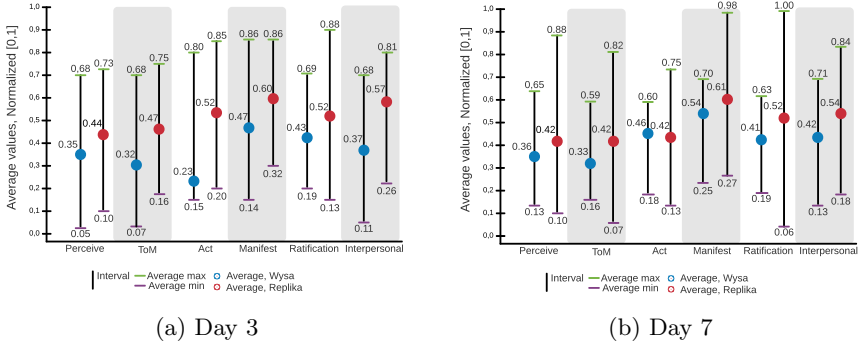


Figure 3.6: Average dimension intervals

Explaining change between Day 3 and Day 7

We now analyze the data from a bottom-up perspective. Generally, the bottom-up approach focuses its analysis on the lower concepts in the ontology descending the 6 main dimensions, to provide detailed explanations. Each sub-dimension in the ontology is mapped to questions in the questionnaire. Hence, by analyzing the collected responses through the ontology’s taxonomy, we can get an aggregated understanding of the six main dimensions of empathy. Furthermore, by analyzing each sub-dimension, we can explain in detail why these changes, between day 3 and day 7, occurred. More dominant sub-dimensions which give rise to these changes can be recognized to ground the base for an explanation. Dominance of sub-dimensions can be calculated using different methods (e.g., in terms of difference, interval compactness, invariance, etc). Here, we define dominance as the difference of maximum, and minimum, between day 3 and day 7. Let us look at the changed intervals to highlight dominant sub-dimensions in each chatbot study.

Wysa sub-dimension intervals

In the Wysa user study, by only considering the most skeptical (minimum) measures on each dimension (see Figure 3.7), we can observe that minimum values of *Act Mood*, *Act Purpose* and *Act Message* notably increased between day 3 and day 7 while minimum values of *Act Time* and *Act Duration* notably decreased. The sub-dimensions which made the most dominant change, w.r.t. their minimum measure was *Manifest Goals* which increased from 0.0 to 0.5 and *Ratification Acceptance* which decreased from 0.5 to 0.0.

Continuing our analysis of the Wysa user study, by only considering the most optimistic (maximum) measures on each dimension (see Figure 3.8) we can notice dominant changes between day 3 and day 7. Particularly notable changes are for sub-dimensions under *Act*, *Manifest* and *Theory of Mind*, where most of the of the maximum values increased, and sub-dimensions below *Inter-*

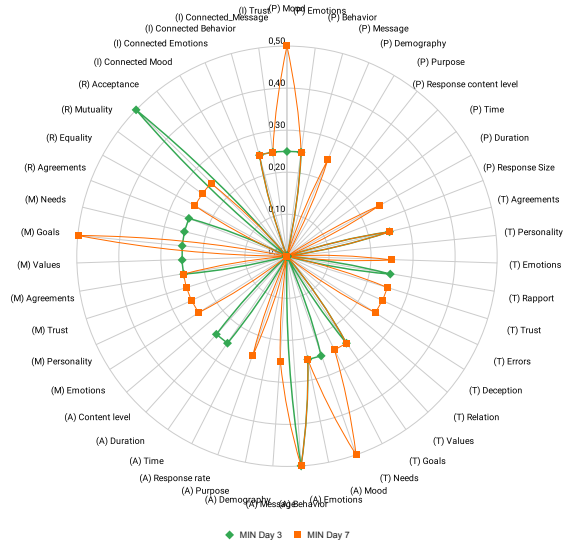


Figure 3.7: Wysa: Minimums Day 3; Day 7

personal where a majority of the maximum values decreased. For instance, notable increase of maximum values can be observed in *Act Mood*, *Act Emotions*, *Act Behavior*, *Act Content level*, *Manifest Content level*, *Manifest Emotions*, *Manifest Personality*, *Manifest Trust*, *Manifest Values*, *Manifest Needs* and *Manifest Agreements*. On the other hand, particular decreases can be observed in *Interpersonal Connected Mood*, *Interpersonal Connected Emotions*, *Interpersonal Connected Behavior*, *Perceive Purpose* and *Perceive Response Content level*. Hence, we can identify subsets of sub-dimensions that more dominantly changed between day 3 and day 7, explaining how the participants' perspectives changed w.r.t. the dimensions of computational empathy.

Replika sub-dimension intervals

In the Replika user study, by only considering the most skeptical (minimum) measures on each dimension (see Figure 3.9), we can observe that minimum values of *Manifest Emotions*, *Manifest Trust*, and *Theory of mind Goals* and *Perceive Mood* notably increased between day 3 and day 7 while minimum values of *Manifest Agreements*, *Theory of mind Values*, *Act Mood*, *Act Emotions* and *Act Purpose* notably decreased. The sub-dimensions which made the most dominant change w.r.t. their minimum measure were *Manifest Emotions*, increasing from 0.25 to 0.5, *Act Mood*, decreasing from 0.5 to 0.25, *Act Emotions*, decreasing from 0.5 to 0.25, and *Act Purpose*, decreasing from 0.5 to 0.0. Hence, we can make detailed estimates about which perceptions changed be-

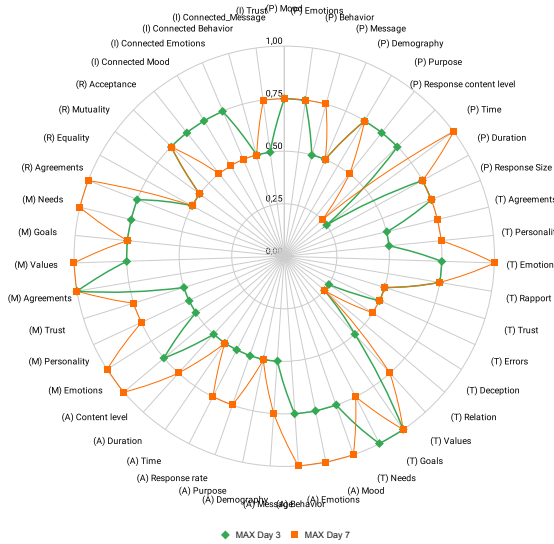


Figure 3.8: WYSA: Maximums Day 3; Day 7

tween day 3 and day 7. For instance, we can estimate that the participants had high expectations of Replika’s purposeful acting at day 3, which was rejected through the long-term interaction (measured at day 7), a dominant factor of the greatly decreased measure of *Act*.

Continuing our analysis of the Replika user study, by only considering the most optimistic (maximum) measures on each dimension (see Figure 3.10), we can observe dominant changes between day 3 and day 7. Particularly notable changes are for sub-dimensions under *Manifest*, *Theory of Mind* and *Perceive*, where most of the of the maximum values increased, and sub-dimensions below *Act* and *Interpersonal* where a majority of the maximum values decreased. For instance, notable increase of maximum values can be observed in *Manifest Emotions*, *Manifest Values*, *Manifest Goals*, *Manifest Agreements*, *Perceive Mood*, *Perceive Emotions*, *Perceive Behavior*, *Perceive Response content level*, *Theory of mind Personality*, *Theory of mind Feelings*, *Theory of mind Trust*, *Theory of mind Goals*, *Theory of mind Needs* and *Interpersonal Trust building*, while notable decrease of maximum values can be observed in *Act Behavior*, *Act Message*, *Act Time*, *Interpersonal Connected mood*, *Perceive Duration* and *Theory of mind Relation*. By considering these observations, we can provide explanations about the chatbot’s computational empathy. For instance, Replika was perceived to be emotionally affected by the interaction (given by *Manifest Emotions*) and had an increased understanding of the user’s trust (given by *Theory of mind Trust*). Concurrently, Replika was perceived to have increased

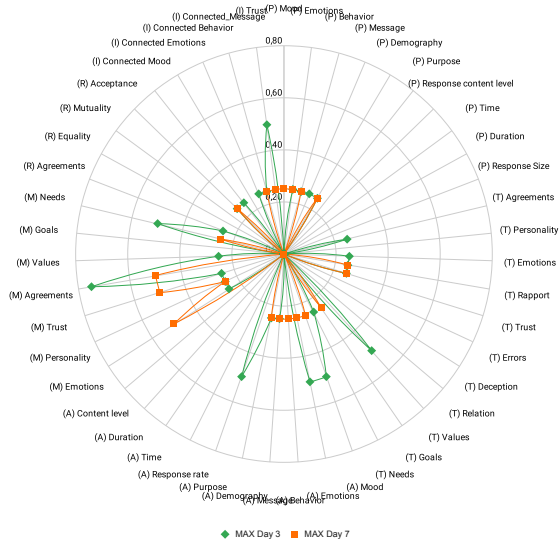


Figure 3.9: Replika: Minimums Day 3; Day 7

capability for trust building (given by *Interpersonal Trust building*). These observations give rise to aggregated as well as detailed explanations of the user’s perceptions.

Empathy measures and chatbot features

We now reconnect to the chatbot features presented in Table 3.7. Recall that two experimental conditions were defined as a configuration of features for each chatbot, Replika and Wysa. Each chatbot feature was further related to one or more dimensions of computational empathy. By considering these relations, six hypotheses (H1-H6) were specified. In the previous analyses, we processed the participants’ perceptions of the chatbots’ capabilities to identify dominant empathy dimensions. By mapping the perceived dimensions to chatbot features (according to Table 3.7) we can further analyze the data to make estimates about how these features had an impact on the perceptions.

By considering the hypotheses H1-H6 and the measured perceptions of empathy dimensions of Replika and Wysa, we can make the following observations:

- H1 does not hold on Day 7. Wysa’s Act was measured as higher than Replika’s.
- H2 holds. Replika’s Manifest was measured as higher than Wysa’s.
- H3 holds. Replika’s Perceive was measured as higher than Wysa’s.
- H4 holds. Replika’s Theory of mind was measured as higher than Wysa’s.
- H5 holds. Replika’s Interpersonal was measured as higher than Wysa’s.
- H6 holds. Replika’s Ratification was measured as higher than Wysa’s.

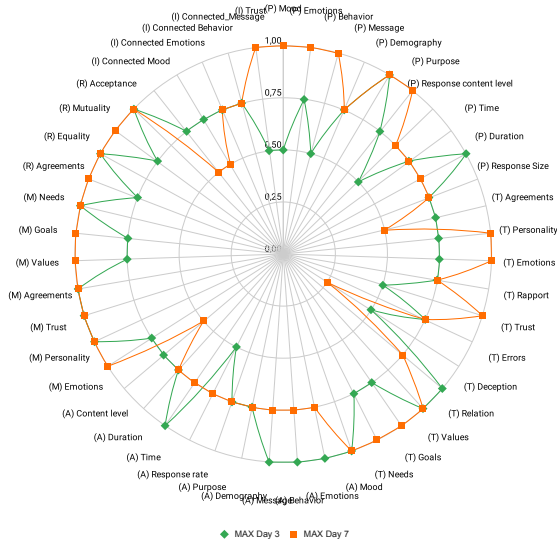


Figure 3.10: Replika: Maximums Day 3; Day 7

By considering the specified chatbot features, we can observe that actuators and sensors, w.r.t. the defined dimensions, in most cases followed the hypotheses regarding the chatbots’ perceived computational empathy. Such observations can provide directions for how a system can be designed to increase this performance measure. An interesting observation is that Wysa’s *Act* was perceived as higher than Replika’s *Act* on day 7, although Wysa has fewer sensors and actuators than Replika. In our previous analyses, this is explained by specific sub-dimensions, where, e.g., *Act Emotions*, *Act Mood*, *Act Behavior* notably increased for Wysa, while sub-dimensions such as *Act Time*, *Act Emotions*, *Act Mood*, and *Act Behavior* notably decreased for Replika.

Let us look at how these observations can be further analyzed computationally. The ontology allows for alternative explanations of users’ perceptions by considering higher-level empathy concepts indirectly derived from their responses. By identifying dominant dimensions in the measures, we can formulate queries to gain further insights into perceptions of computational empathy.

Reasoning queries on observed dominant dimensions

The semantic relations of the ontology can be utilized for reasoning about perceptions to generate theories about high level empathy concepts. Given a set of observations, we can create corresponding queries to the ontology. For instance, in the Replika user-study, we observed the dominant sub-dimensions *Perceive Emotions* and *Manifest Emotions*. These observations can be trans-

formed into the following query: “*Manifest_Emotion* \sqcap *Perceive_Emotion* \sqcap *isAffecting.Manifest*”. Then, given the defined formula

(1b) $EmpathicManifest \sqsubseteq Perceive \sqcap Manifest \sqcap$

$\forall isAffecting.Manifest$, we can infer “*EmpathicManifest*”, asserting a theory stating that the chatbot was perceived to have a mental state affected by its perceptions. Furthermore, given the previous assertion, we can advance with additional queries, such as: “*EmpathicManifest* \sqcap *Manifest_Emotion* \sqcap *Act*” to infer “*EmpathicAct* \sqcap *isInfluencing.Act* \sqcap *Act_Emotion*”, asserting theories about the chatbot’s perceived capability of acting emphatically w.r.t. emotions. In this way, by considering the holistic semantic model of computational empathy, perceptions can be analyzed on a low level to reach conclusions about high level interdependent concepts, e.g., *EmpathicAct*, which may not be reached by looking at isolated dimensions, e.g., *Act*, alone. Hence, the ontology can provide different explanations of the users’ perceptions by considering ramification effects captured in the model.

In this section, we have presented an evaluation of the developed model of computational empathy, applied as a tool to measure perceptions of empathy in two interactive agents, Wysa and Replika. We show how we can understand and explain the perceptions of empathy with (1) an analysis considering average measures on the main dimensions, (2) a detailed bottom-up analysis considering possibilistic intervals on each sub-dimension, and (3) a semantic analysis considering ramification effects between interdependent concepts.

3.7 Chapter Remarks

Through an examination of prior research on empathy and conversational agents in the domain of health and well-being, along with a subsequent user study exploring human-agent interaction, we have identified a set of necessary conditions for computational empathy. These conditions can be encapsulated in a formal, multi-dimensional definition and semantic model—an ontology—consisting of six overarching dimensions, expanded in more explicit underlying concepts. The hierarchical structure of the formal definition aids in understanding empathy at different levels of abstraction, allowing to incorporate perspectives from various human empathy theories. This includes the two routes to empathy, affective and cognitive, from a psychological standpoint [122], the observed relations between perception and action in neuroscience [233], and concepts such as shared emotions [282] and social bonds [304] from a social cognition perspective.

Conceptual Alignment

We have highlighted the hierarchical structure of the ontology to provide a bottom-up approach for defining abstract empathy concepts. Moreover, we see the potential of integrating external ontologies to define sub-concepts of com-

putational empathy. For instance, there are ontologies developed to define emotions [175], personality [8], honesty [305], trust [294], and other factors that may be important for empathy in particular settings. Nevertheless, when integrating external ontologies, careful attention must be paid to ensure alignment of the concepts. The issue of “alignment” is a well-recognized challenge in ontology integration [159, 278, 316] and must be addressed to maintain semantic consistency and meaningful integration. The potential misalignment of concepts highlights the importance of maintaining the hierarchical structure within our proposed ontology, as it facilitates the clear definition of concepts in lower levels. The hierarchical structure supports achieving conceptual alignment from multiple perspectives, whether it is between systems through integrated ontologies or in relation to human comprehension. When eliciting perceptions from human participants, it is crucial for our observations to resonate with the participants’ subjective understanding of the concepts. By employing explicit low-level concepts, a human-system conceptual alignment can be achieved, ensuring that our inquiries connect with the participants’ understanding while maintaining the intended meaning of ontological concepts.

Data Interoperability versus Specialization

The primary aim of this work has been to present a high-level general definition of computational empathy that, through adaptable underlying sub-concepts, serves as a versatile tool applicable to various systems and applications. A significant advantage of the presented ontology is that it promotes data interoperability [222] and standardization [113]. By establishing a shared concept of empathy among various systems, we enable a common ground for communication and understanding. In the rapidly evolving landscape of communicative systems, having standardized concepts becomes relevant to ensure effective interaction and knowledge sharing. However, the identified high-level dimensions of computational empathy can be understood in different ways for different user groups, contexts, and agents. Thus, the model must be specialized for particular use-cases. For example, empathy dimensions relevant in chatbots may be different from the dimensions relevant in autonomous cars or smart homes. Moreover, the model can be analyzed for understanding which dimensions are relevant for perceived empathy in humans. Through extended user studies analyzed through qualitative and quantitative methods, we can identify dominant dimensions for various populations and interactive agents.

In applying the model to state-of-the-art mental health chatbots (Replika and Wysa), we presented an example of the ontology’s diagnostic utility and evaluative granularity, suggesting not only how different empathy dimensions manifest in interaction, but also how user perceptions shift over time. Our possibilistic interval analysis indicated a temporal convergence of empathy perceptions, motivating to further explore formal models of mental change and causal dynamics in empathic interactions.

Chapter 4

Emotion Verification

In this chapter, we introduce the action language \mathcal{C}_{MT} (Mind Transition Language), built on top of Answer Set Programs and transition systems to model the dynamics of the human mind. Drawing inspiration from well-established psychological theories, such as Appraisal theory of Emotion (AE), we establish formalizations of mental states, such as emotions, in terms of multi-dimensional configurations, such as a human’s appraisals of a situation. With the primary objective to address the need for controlled agent behaviors to minimize unwanted mental side-effects, the framework encompasses the modeling of valid transitions between mental states by employing principles of mental change, suggested by psychological theories, such as Hedonic Emotion Regulation (HER) and Utilitarian Emotion Regulation (UER). These principles of mental change are translated into transition constraints, and properties of invariance, which are rigorously evaluated using formal methods. This enables controlled reasoning about the dynamic evolution of human mental states.

4.1 Introduction

Interactive and intelligent systems are increasingly being designed to display human-like mental capabilities [3, 205, 208, 188, 53]. For instance, in the area of health and wellbeing, software assistants are being developed to display complex human traits, such as empathy and sympathy [208], e.g., to deliver emotionally charged actions [3] or to provoke empathic responses from users [205]. Some of these systems are deployed in society for, e.g., depression support, therapy and behavior-change interventions [188, 53]. In such applications, that in various ways deal with manipulation of human mental states, such as emotions, ensuring reliable system behavior is crucial [210], where a system

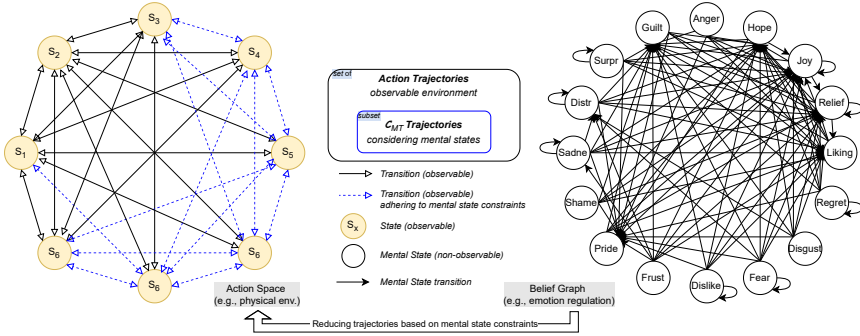


Figure 4.1: In \mathcal{C}_{MT} , available sequences of actions and states (trajectories) in the “physical” state space are constrained by the actions’ influence on the “mental” state space.

may need to constrain and plan its interactions to anticipate and reduce unwanted side-effects (e.g., unwanted mental states) as a result of its behavior. However, a key challenge in modeling human mental attributes is their constant change [219]. Hence, a system’s model of the human mind must be dynamic and capture the causality of mental states.

In this chapter, we introduce the action language \mathcal{C}_{MT} (Mind Transition Language) to model the dynamics of the human mind. Inspired by well-established psychological theories, such as Roseman’s Appraisal Theory of Emotion (AE) [248], we formalize mental states—such as emotions—as multi-dimensional configurations, representing a human’s appraisals of a given situation. The primary objective of \mathcal{C}_{MT} is to enable controlled agent behavior that minimizes unwanted mental side effects. To achieve this, our framework formalizes mental state transitions using principles derived from psychological research, including Hedonic Emotion Regulation (HER) [315] and Utilitarian Emotion Regulation (UER) [284]. These principles are encoded in ASP as transition constraints within a structured representation called the *Belief Graph*.

The action language \mathcal{C}_{MT} enables the verification of key system properties such as *invariance*, ensuring the system avoids deliberately entering undesirable mental states. These properties can be rigorously analyzed using formal verification techniques [130]. Ultimately, our framework facilitates reasoning about the evolution of human mental states while constraining actions in the “physical” environment (see Figure 4.1).

While action languages serve as intuitive specifications for dynamic reasoning processes, they can be characterized as Answer Set Programs. This allows users who are not familiar with logic programming to specify dynamic reasoning using the vocabulary of an action language. These specifications can then be translated into logic programming to implement the dynamics. Consequently, a multidisciplinary team can collaborate effectively to capture mental

state dynamics, working as a method for knowledge elicitation and engineering. Emotional reasoning is inherently complex due to the exponential space of emotion states, especially when accounting for changes over time. This complexity poses a computational challenge, demanding solutions to combinatorial problems. ASP stands out as a powerful logic programming tool in this regard. It is capable of expressing NP-search problems that can be solved using a non-deterministic Turing machine in polynomial time, with solutions encoded as answer sets [49]. Consequently, logic programming, particularly ASP, is a suitable method for managing and implementing emotional reasoning. With this motivation, we introduce the framework’s syntax and semantics, and establish a link to ASP. Additionally, effective emotional reasoning requires a structured understanding of the dynamics governing emotion states, which can be elicited from psychological theories of emotion. Logic programming facilitates the integration of domain knowledge pertaining to these dynamics, thereby enhancing aspects such as controllability and transparency of the approach.

Given this background and approach, this chapter makes the following key contributions:

- We introduce the action language \mathcal{C}_{MT} (Mind Transition Language), serving as a foundational platform, capturing domain-independent properties and constraints relevant to mental state reasoning.
- We further present characterizations of \mathcal{C}_{MT} in the setting of emotional reasoning, capturing different psychological principles from a set of emotion theories; Appraisal theory of Emotion [248], Hedonic Emotion Regulation [315] and Utilitarian Emotion Regulation [284].
- We present the framework’s syntax and semantics, and its practical implementation in Answer Set Programming, evaluated formally and empirically.
- We present how the introduced framework can be applied as a formal method for representing, analyzing and comparing psychological theories in terms of action trajectories.

The rest of this chapter is organized as follows. In Section 4.2, the related work is presented. In Section 4.3, we present the proposed formal framework. In Section 4.4, we present a case study in the setting of emotional reasoning. In Section 4.6, we conduct an empirical evaluation. In Section 4.5, we conduct a formal analysis. Finally, in Section 4.7, conclude the chapter.

4.2 Related Work

There is a diverse body of research related to the formal framework presented in the current work [239, 272, 271, 240, 156]. Plan recognition as planning, originally introduced by Ramirez and Geffner [239], uses planning algorithms to enable an agent to recognize the goals and plans of other agents. Empathetic Planning [272] defines empathy as the ability to understand and share

the thoughts and feelings of another. Following this definition, an assistive empathetic agent is formalized as being able to reason about the preferences of an empathizee [272]. In Active Goal Recognition (AGR) [271], an AGR agent actively senses and acts as part of the goal recognition process. While pursuing its goal, the agent executes sensing and world altering actions. This relates to the notion of *influence* in the current work, where actions aim to influence aspects of a human agent’s perception (e.g., appraisal) to change mental fluents. In contrast to AGR, the current work additionally models a mental state-space.

A variety of Emotion BDI (Belief, Desire, Intention) frameworks [227, 148, 150, 259, 258] have been introduced. These approaches have aimed, e.g., to model behaviors which are expected from agents under the influence of emotions [227], or to provide modular generic interfaces for emotional agents [148] to enable emotion theory-based models as filters for emotional reasoning. While recent works have considered affective states, such as emotion, and aim to integrate affective states throughout the BDI model, challenges persist in implementing principles of mental change, such as emotion regulation [259].

In the setting of mental state representation and reasoning, two related areas of research regard: 1) logics of mental attitudes and emotion [176, 2, 177, 84, 280], and 2) epistemic planning [40] extended to cognitive planning [178, 85]. Logics of mental attitudes and emotion aim to formalize the relationships between epistemic and motivational attitudes of human and artificial agents, as well as the influence of mental attitudes on emotions. Some related works in this line of research include the logical formalization of OCC theory of emotions [2], the formalization of counterfactual emotions [177], the representation of emotion intensity and coping strategies [84], the modeling of emotion triggers [280], and the logical theory of epistemic and motivational attitudes and their dynamics [176]. Nevertheless, the principles constraining this causality and its potential side effects are not considered. In contrast to previous approaches to modeling and reasoning about mental states, the proposed \mathcal{C}_{MT} action language deals with a multi-dimensional representation of mental states and the constraints for modeling principle-based transitions between them.

In the following sections, we introduce the formal framework that integrates these concepts, providing a structured approach to modeling and managing the dynamics of human mental states.

4.3 Formal Framework

This section introduces the syntax and semantics of the proposed formal framework built on top of ASP-based action reasoning. The framework offers flexibility to specialize for specific mental state domains, such as emotions, aligning with established psychological theories, such as the Appraisal theory of Emotion by Roseman (1996) [248]. By incorporating principles of mental change, sets of transition constraints are formalized and implemented in terms of integrity constraints in answer set programs.

In the proposed action language, similar to previous action languages, *fluents* represent properties that can change over time. These can describe various aspects of human-agent interactions, including non-observable aspects such as psychological attributes and observable aspects of agents or the environment. The value of a fluent at any given time depends on how it is affected by so-called *actions* or indirectly by other fluents.

We build on the action language \mathcal{C}_{TAID} [97], originally designed for modeling biological systems. \mathcal{C}_{TAID} includes features such as *allowance*, *inhibition*, and *triggers*, which are also relevant for modeling the dynamics of mental states. The *triggers* causal law accounts for interactions based on reactions, making it particularly useful for capturing how mental states may change as indirect effects of actions or external events. *Allowance* rules specify that an action can occur under certain conditions but is not mandatory, while *inhibition* rules prevent an action from occurring in certain contexts. These mechanisms are especially important in modeling mental states, where dependencies between cognitive and emotional factors may be partially known or not explicitly modeled. In such cases, allowance and inhibition rules provide a flexible way to account for uncertainties and exceptions in mental state reasoning. These rules are relevant in the context of mental states where we have partial knowledge about the dependencies and reasons behind interactions in the mind. In situations where dependencies are partially known or not explicitly modeled, such as some conditions of the environment, allowance and inhibition rules provide a flexible way to handle uncertainties and exceptions.

We introduce the action language \mathcal{C}_{MT} (Mind Transition Language), serving as a foundational platform, capturing domain-independent properties and constraints relevant to mental state reasoning. Along with the action language, we introduce some “syntactic sugar” to facilitate expressions about mental state dynamics. This includes “influences mental fluent” (similar to “causes” in \mathcal{C}_{TAID}), handling actions or events that may influence a change in a mental state, and the rules “facilitates” and “contravenes” (similar to “allowance” and “inhibition” rules in \mathcal{C}_{TAID}), which regulate actions’ execution in an initial state. However, we specialize these rules to particularly concern *human actions* regulated by *mental fluents*. This is motivated by emotion theories, suggesting that “emotions have distinctive goals and action tendencies” [249].

Moreover, \mathcal{C}_{MT} introduces abstractions that we call **mental states**, which are defined by sets of fluents, along with mental state transition constraints, called **forbids to cause**, that specify relationships among mental fluents from one state to the next. Notably, such expressions enabling direct restriction of fluents between states, independent of actions, have not been explicitly incorporated in previous action languages like \mathcal{C}_{TAID} .

Consequently, the proposed framework for mental-state reasoning is comprised of two components: 1) The action language \mathcal{C}_{MT} , which defines actions in the environment/interaction that trigger changes in mental states, and 2) a set of constraints that precisely define valid transitions between mental states, derived from psychological principles. These constraints characterize a so-called

Belief Graph (BG). In ASP, a BG is encoded as sets of integrity constraints, restricting particular fluent changes in transitions between mental states. Mental state dynamics are linked to actions in the environment by considering their effects on fluents in the mental state abstractions. In this way, a BG filters the potential trajectories resulting from the action language, based on their effects on mental states (see Figure 4.2 for a conceptualization of the framework).

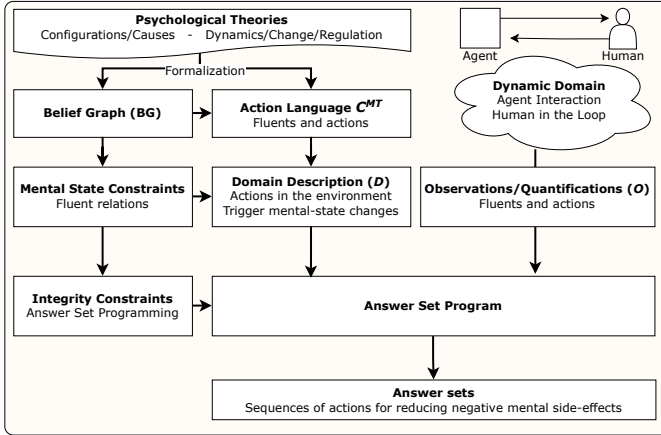


Figure 4.2: Conceptual Framework

Belief Graph (BG)

A specialized BG is a set of propositional atoms for valid transitions between mental states. The mental states within the BG represent configurations of factors that contribute to specific mental states. These factors, known as mental fluents, are determined by psychological theories and encompass a range of possible values, thus defining the potential states within the BG. The mental fluents that constitute the states of the BG are defined as follows:

Definition 16 (Mental fluent) Let $C = \{c_1, \dots, c_n\}$ be a set of symbols denoting psychological classes, and let $V = \{V_{c_1}, \dots, V_{c_n}\}$ be a set of sets of constants denoting psychological values for each class of C . A mental fluent is a ground atom $f(c, v)$ of arity 2 such that $c \in C$, $v \in V_c$.

A mental state space S is defined as the set of all possible combinations of mental fluents.

Definition 17 (Mental state space) Let $C = \{c_1, \dots, c_n\}$ be a set of psychological classes, and let $V = \{V_{c_1}, \dots, V_{c_n}\}$ be a set where each V_{c_i} ($1 \leq i \leq n$)

is the set of possible values for the class c_i . The mental state space S is defined as

$$S = \{ \{ f(c_1, v_1), f(c_2, v_2), \dots, f(c_k, v_k) \} \mid 1 \leq i \leq k, c_i \in C, v_i \in V_{c_i} \},$$

Each set $\{ f(c_1, v_1), f(c_2, v_2), \dots, f(c_k, v_k) \}$ represents a unique combination of mental fluents, called a mental state.

We now proceed by defining a BG with states S and directed edges $E \subseteq S \times S$. These edges between states capture principles of mental change suggested by psychological theories, providing rules for allowed mental state transitions.

Definition 18 (Belief Graph) *A Belief Graph is a directed graph $BG = (S, E)$ where S is a finite set of mental states and $E \subseteq S \times S$ is a set of directed edges representing valid transitions between mental states.*

A BG needs to be characterized for each specific application since different approaches to mental influence may be applicable. Consequently, to support controlled system behavior and principle-based assessment, it is essential to establish reasoning principles of “monotonicity” governing valid transitions between mental states. These principles can be motivated from various sources, such as psychological theories or insights from human experts. In the setting of reasoning about transitions between emotions, there are different principles and theories that have contrasting views on emotional change. For instance, principles of *hedonic emotion regulation* [315] aim to increase positive emotions and decrease negative emotions. Another set of principles regards *utilitarian emotion regulation* [284], which aim to increase emotions that provide utility, such as control. Hence, the BG is agnostic to the choice of psychological theories. We show how this flexibility enables the framework to model, and compare, psychological theories through the principle-based assessment of generated trajectories.

\mathcal{C}_{MT} action language

\mathcal{C}_{MT} consists of a set of symbols representing actions and fluents, forming the alphabet of the action language. Given that the purpose of an action language is to provide a higher-level framework that makes modeling dynamic systems more natural and modular, several aspects of the language serve as “syntactic sugar” to simplify the modeling of mental states. The primary addition, beyond a specialized alphabet, is the incorporation of constraints, **forbids to cause**, that given a set of mental fluents in the current state forbids a set of mental fluents to hold in the next state. By capturing elements of both the external environment and the human mind, the language facilitates the description and analysis of how environmental events influence human mental states and behavior, and through the new forbidding constraints, *actions* as well as *indirect effects* on mental fluents can be constrained. We require this addition to the language in order to model principles of mental change.

Definition 19 (\mathcal{C}_{MT} alphabet) *Let \mathbf{A} be a non-empty set of actions and \mathbf{F} be a non-empty set of fluents.*

- $\mathbf{F} = \mathbf{F}^E \cup \mathbf{F}^H$ such that \mathbf{F}^E is a non-empty set of fluent literals describing observable items in an environment and \mathbf{F}^H is a non-empty set of fluent literals describing quantified non-observable features of mental-states of humans. \mathbf{F}^E and \mathbf{F}^H are pairwise disjoint.
- $\mathbf{A} = \mathbf{A}^E \cup \mathbf{A}^H$ such that \mathbf{A}^E is a non-empty set of actions that can be performed by a software agent and \mathbf{A}^H is non-empty set of actions that can be performed by a human agent. \mathbf{A}^E and \mathbf{A}^H are pairwise disjoint.

Within \mathcal{C}_{MT} , a domain description defines static and dynamic causal laws for actions. These laws precisely express the expected influences exerted on mental fluents, either as direct effects of actions or as indirect causal effects. The laws governing mental change operate by modulating the variables of a given mental state while adhering to the constraints outlined by the BG.

Definition 20 (\mathcal{C}_{MT} domain description language) *The \mathcal{C}_{MT} domain description language $D^{MT}(\mathbf{A}, \mathbf{F})$ consists of static and dynamic causal laws of the following form:*

\mathcal{C}_{MT} domain description language (extending \mathcal{C}_{TAID}):

$$(a \text{ causes } f_1, \dots, f_n \text{ if } g_1, \dots, g_m) \quad (1)$$

$$(f_1, \dots, f_n \text{ if } g_1, \dots, g_m) \quad (2)$$

$$(f_1, \dots, f_n \text{ triggers } a) \quad (3)$$

$$(f_1, \dots, f_n \text{ allows } a) \quad (4)$$

$$(f_1, \dots, f_n \text{ inhibits } a) \quad (5)$$

$$(\text{noconcurrency } a_1, \dots, a_n) \quad (6)$$

$$(\text{default } g) \quad (7)$$

Mental-state domain description language extension:

$$(a \text{ influences } f_1^h, \dots, f_n^h \text{ if } g_1, \dots, g_m) \quad (8)$$

$$(g_1, \dots, g_m \text{ influences } f_1^h, \dots, f_n^h) \quad (9)$$

$$(g_1^h, \dots, g_m^h \text{ facilitates } a^h) \quad (10)$$

$$(g_1^h, \dots, g_m^h \text{ contravenes } a^h) \quad (11)$$

$$(g_1^h, \dots, g_m^h \text{ forbids to cause } f_1^h, \dots, f_n^h) \quad (12)$$

where $a \in \mathbf{A}$, $a^h \in \mathbf{A}^H$, and $a_i \in \mathbf{A}$ ($0 \leq i \leq n$) and $f_j \in \mathbf{F}$, ($0 \leq j \leq n$) and $g_j \in \mathbf{F}$, ($0 \leq j \leq m$), and $f_1^h, \dots, f_n^h \in \mathbf{F}^H$, $g_1^h, \dots, g_m^h \in \mathbf{F}^H$ are mental fluents of the form $f(c, v)$, where c is a psychological class and v is a psychological value.

BGs are embedded in \mathcal{C}_{MT} logic programs, i.e., finite sets of dynamic and causal laws on mental states. This characterization allows us to ensure controlled mental change by restricting the states and state-transitions. The mental state, within the context of the domain description, represents an interpretation of the current state of the system.

Definition 21 (Mental state interpretation) A state $s \in S$ of the domain description $D^{MT}(\mathbf{A}, \mathbf{F})$ is an interpretation over \mathbf{F} such that

1. for every static causal law $(f_1, \dots, f_n \text{ if } g_1, \dots, g_m) \in D^{MT}(\mathbf{A}, \mathbf{F})$, we have $\{f_1, \dots, f_n\} \subseteq s$ whenever $\{g_1, \dots, g_m\} \subseteq s$.
2. for every static causal law $(g_1, \dots, g_m \text{ influences } f_1^h, \dots, f_n^h) \in D^{MT}(\mathbf{A}, \mathbf{F})$, we have $\{f_1^h, \dots, f_n^h\} \subseteq s$ whenever $\{g_1, \dots, g_m\} \subseteq s$, and $\{f_1^h, \dots, f_n^h\} \subseteq \mathbf{F}^H$.

S denotes all the possible states of $D^{MT}(\mathbf{A}, \mathbf{F})$.

In this definition, a state is determined by the satisfaction of static causal laws within the domain description $D^{MT}(\mathbf{A}, \mathbf{F})$. The first condition ensures that if the prerequisite mental fluents for an action are true, then the consequent mental fluents will also be true in the state. The second condition specifies that if certain mental fluents influence a particular mental fluent, then the influenced mental fluent must be true when all the influencing mental fluents are true.

Let us define the laws of the domain description more precisely.

Definition 22 (Domain description) By considering the domain description $D^{MT}(\mathbf{A}, \mathbf{F})$ and a state s , the following rules and laws apply:

1. An inhibition rule $(f_1, \dots, f_n \text{ inhibits } a)$ is active in s , if $f_1, \dots, f_n \in s$, otherwise, passive. The set $A_I(s)$ is the set of actions for which there exists at least one active inhibition rule in s (as in \mathcal{C}_{TAID} [97]).
2. A triggering rule $(f_1, \dots, f_n \text{ triggers } a)$ is active in s , if $f_1, \dots, f_n \in s$ and all inhibition rules of action a are passive in s , otherwise, the triggering rule is passive in s . The set $A_T(s)$ is the set of actions for which there exists at least one active triggering rule in s . The set $\bar{A}_T(s)$ is the set of actions for which there exists at least one triggering rule and all triggering rules are passive in s (as in \mathcal{C}_{TAID} [97]).
3. An allowance rule $(f_1, \dots, f_n \text{ allows } a)$ is active in s , if $f_1, \dots, f_n \in s$ and all inhibition rules of action a are passive in s , otherwise, the allowance rule is passive in s . The set $A_A(s)$ is the set of actions for which there exists at least one active allowance rule in s . The set $\bar{A}_A(s)$ is the set of actions for which there exists at least one allowance rule and all allowance rules are passive in s (as in \mathcal{C}_{TAID} [97]).
4. A facilitating rule $(g_1^h, \dots, g_m^h \text{ facilitates } a^h)$ is active in s , if $a^h \in \mathbf{A}^H$ and $g_1^h, \dots, g_m^h \in s$ and all inhibition rules and contravening rules of action a are passive in s , otherwise, the facilitating rule is passive in s . The set $A_{FAC}(s)$ is the set of actions for which there exists at least one active facilitating rule in s . The set $\bar{A}_{FAC}(s)$ is the set of actions for which there exists at least one facilitating rule and all facilitating rules are passive in s .

5. An contravening rule (g_1^h, \dots, g_m^h **contravenes** a^h) is active in s , if $a^h \in \mathbf{A}^H$ and $g_1^h, \dots, g_m^h \in s$ and all inhibition rules and facilitating rules of action a are passive in s , otherwise, the contravening rule is passive in s . The set $A_{INT}(s)$ is the set of actions for which there exists at least one active contravening rule in s .
6. A dynamic causal law (a causes f_1, \dots, f_n if g_1, \dots, g_m) is applicable in s , if $g_1, \dots, g_m \in s$.
7. A static causal law (f_1, \dots, f_n if g_1, \dots, g_m) is applicable in s , if $g_1, \dots, g_m \in s$.

A dynamic causal law (a **influences** f_1^h, \dots, f_n^h if g_1, \dots, g_m) is applicable in s , if $g_1, \dots, g_m \in s$, and $f_1^h, \dots, f_n^h \in F^H$, and $f_i \in F (1 \leq i \leq n)$.

8. A static causal law (g_1, \dots, g_m **influences** f_1^h, \dots, f_n^h) is applicable in s , if $g_1, \dots, g_m \in s$, and $f_1^h, \dots, f_n^h \in F^H$, and $f_i \in F (1 \leq i \leq n)$.
9. A forbidding rule (g_1^h, \dots, g_m^h **forbids to cause** f_1^h, \dots, f_n^h) is active in s if $\{g_1^h, \dots, g_m^h\} \subseteq s$, where $f_1^h, \dots, f_n^h, g_1^h, \dots, g_m^h \in F^H$. The set $F(s)$ is the set of fluents that are forbidden in $s + 1$ when at least one active forbidding rule exists in s .

The output of the action language is in terms of trajectories. A mental-state trajectory consists of a sequence of valid transitions, represented as $\langle s_0, A_1, s_1, A_2, \dots, A_n, s_n \rangle$, with alternating sets of mind-altering actions $A \subseteq \mathbf{A}$ and mental-states $s \in S$, following the constraints of the BG.

Definition 23 (Trajectory) Let $D^{MT}(\mathbf{A}, \mathbf{F})$ be a domain description. A trajectory $\langle s_0, A_1, s_1, A_2, \dots, A_n, s_n \rangle$ of $D^{MT}(\mathbf{A}, \mathbf{F})$ is a sequence of sets of actions $A_i \subseteq A$ and states s_i of $D^{MT}(\mathbf{A}, \mathbf{F})$ satisfying the following conditions for $0 \leq i < n$:

1. $(s_i, A, s_{i+1}) \in S \times 2^A \setminus \{\emptyset\} \times S$
2. $A_T(s_i) \subseteq A_{i+1}$
3. $A_{FAC}(s_i) \subseteq A_{i+1}$
4. $\overline{A}_T(s_i) \cap A_{i+1} = \emptyset$
5. $\overline{A}_A(s_i) \cap A_{i+1} = \emptyset$
6. $A_I(s_i) \cap A_{i+1} = \emptyset$
7. $\overline{A}_{FAC}(s_i) \cap A_{i+1} = \emptyset$
8. $A_{INT}(s_i) \cap A_{i+1} = \emptyset$
9. $|A_i \cap B| \leq 1$ for all (noconcurrency B) $\in D^{MT}(\mathbf{A}, \mathbf{F})$.
10. $F(s_i) \cap s_{i+1} = \emptyset$ (no forbidden fluents in s_{i+1})

The set B in condition (9) represents a subset of actions restricted by noconcurrency constraints in $D^{MT}(\mathbf{A}, \mathbf{F})$, ensuring that actions in B cannot execute simultaneously. This prevents conflicts where multiple actions attempt to modify the same fluent at the same time. While actions affecting distinct fluents can occur concurrently, those altering the same fluent are mutually exclusive, maintaining consistency in fluent updates and ensuring well-defined state transitions.

Definition 24 (Action Observation Language) *The action observation language of \mathcal{C}_{MT} (similar to \mathcal{C}_{TAID}) consists of expressions of the following form:*

$$(f \text{ at } t) \quad (a \text{ occurs.at } t) \quad (8)$$

where $f \in \mathbf{F}$, a is an action and $t \in \mathbb{N}$ is a point in time.

The action observation language allows us to specify observations concerning the current state of mental fluents and the execution of actions that influence mental states. By combining observations with the causal laws of the domain description, we can generate plans, explanations, and predictions regarding the behavior of the system. This integration of observations and the domain description is referred to as an action theory.

Definition 25 (Action Theory) *Let D be a domain description and O be a set of observations. The pair (D, O) is called an action theory.*

The action theory forms the basis for constructing trajectory models, trajectories where all observations are satisfied, providing a structured representation of the system's dynamics over time. Trajectory models enable us to analyze and reason about the evolution of mental states and actions, allowing for a deeper understanding how the mental state domain operates and how it responds to different observations and changes.

Definition 26 (Trajectory Model) *Let (D, O) be an action theory. A trajectory $\langle s_0, A_1, s_1, A_2, \dots, A_n, s_n \rangle$ of D is a trajectory model of (D, O) , if it satisfies all observations of O in the following way:*

1. if $(f \text{ at } t) \in O$, then $f \in s_t$
2. if $(a \text{ occurs.at } t) \in O$, then $a \in A_{t+1}$.

We can observe that actions in a trajectory model can be actions executed by a rational software agent to influence mental fluents, or action tendencies estimated to be executed by the human agent. To further enhance the formal understanding and analysis, we need a mechanism to write queries about state dynamics. The Action Query Language provides a means to inquire about specific sequences of actions and their impact on the mental-state system. This is achieved by specifying subsets of the action set and their corresponding occurrences in time.

Definition 27 (Action Query Language) *The action query language of \mathcal{C}_{MT} regards assertions about executing sequences of actions with expressions that constitute trajectories. A query is of the following form: $(f_1, \dots, f_n \text{ after } A_i \text{ occurs.at } t_i, \dots, A_m \text{ occurs.at } t_m)$ where f_1, \dots, f_n are fluent literals $\in \mathbf{F}$, A_i, \dots, A_m are subsets of \mathbf{A} , and t_i, \dots, t_m are points in time.*

By formulating queries in the action language, causal relationships between actions and their effects can be investigated, contributing to a deeper understanding of system dynamics, informed decision-making and controlled methods for automated planning.

In the upcoming section, we will explore a case study that applies a specialization of \mathcal{C}_{MT} for emotional reasoning. The case study models emotion states and dynamics based on well-established psychological theories. By examining this specialized application, we aim to demonstrate the effectiveness and versatility of the \mathcal{C}_{MT} framework for modeling different kinds of mental state dynamics.

Action Language in Answer Set Semantics

In this section, we provide operational semantics for \mathcal{C}_{MT} . This operational semantics is based in translations of expressions from the \mathcal{C}_{MT} domain description (20) into answer set programs. The descriptions of expressions 1-7 follow the definitions in [97]. In order to keep the chapter self-contained with a complete translation, we cover the necessary translations from [97] building up to the proposed mental-state specifications. We organize the translations into distinct subsections, addressing the Action Description Language, the Action Observation Language, and the Action Query Language.

Encoding of the Action Description Language

We consider the encoding of \mathcal{C}_{TAID} action description language, which we extend with components of \mathcal{C}_{MT} .

Define symbols for a fluent $f \in \mathbf{F}^E$ and an action $a \in \mathbf{A}^E$ (extending [97]).

```
fluent_e(f), action_e(a).
fluent(f) :- fluent_e(f).
action(a) :- action_e(a).
```

Define symbols for a fluent $e \in \mathbf{F}^H$ and an action $u \in \mathbf{A}^H$ (extending [97]).

```
mental_fluent(e), human_action(u).
fluent(e) :- mental_fluent(e).
action(u) :- human_action(u).
```

Define a range of time points $0 \leq t \leq t_{max}$, $t \in \mathbb{N}$ (as in [97]).

```
time(0..t_max). #const t_max = N.
```

Contradiction Constraint (as in [97]): A fluent $f \in \mathbf{F}$ and its negation $\neg f$ cannot hold simultaneously at the same time step $T \in \mathbb{N}$. In ASP, this is enforced as a constraint, ensuring that $\text{holds}(f, T)$ and $\text{holds}(\text{neg}(f), T)$ cannot both be true. The predicates $\text{fluent}(f)$ and $\text{time}(T)$ declare fluents and time steps.

```
:- holds(f, T), holds(neg(f), T), fluent(f), time(T).
```

Inertial Fluents (as in [97]): An inertial fluent persists across time steps unless modified by an action or a static causal law. In ASP, this is encoded by ensuring that if $\text{holds}(f, T)$ is true and not overridden, then $\text{holds}(f, T + 1)$ holds by default. The predicates $\text{fluent}(f)$ and $\text{time}(T)$ track fluents over time.

```
holds(f,T+1) :- holds(f,T), not holds(neg(f,T+1)), not default(f),
fluent(f), time(T), time(T+1).
```

Non-Inertial Fluents (as in [97]): A non-inertial fluent resets to a default value unless explicitly updated. In ASP, this is encoded by inferring $\text{holds}(f, T)$ whenever $\text{default}(f)$ holds and $\text{holds}(\text{neg}(f), T)$ is not inferred. The predicates $\text{fluent}(f)$, $\text{default}(f)$, and $\text{time}(T)$ define non-inertial fluents and their default behavior.

```
holds(f,T) :- not holds(neg(f),T), default(f), fluent(f), time(T).
```

Dynamic Causal Law (as in [97]): A dynamic causal law (*a causes f if g_1, \dots, g_n*) states that if an action $a \in \mathbf{A}$ occurs at time $T \in \mathbb{N}$ and fluents $g_1, \dots, g_n \in \mathbf{F}$ hold, then a fluent $f \in \mathbf{F}$ must hold at time $T + 1$. In ASP, one rule is generated for each dynamic causal law, ensuring that if $\text{holds}(\text{occurs}(a, T))$ and $\text{holds}(g_1, T), \dots, \text{holds}(g_n, T)$ hold, then $\text{holds}(f, T + 1)$ is inferred. The predicates $\text{fluent}(f)$ and $\text{fluent}(g_1), \dots, \text{fluent}(g_n)$ declare fluents, $\text{action}(a)$ declares the action, and $\text{time}(T)$ ensures valid time progression.

```
holds(f,T+1) :- holds(occurs(a),T),
holds(g_1,T), ..., holds(g_n,T),
fluent(g_1), ..., fluent(g_n),
fluent(f), action(a), time(T), time(T+1).
```

Static Causal Law (as in [97]): A static causal law (*f if g_1, \dots, g_n*) states that a fluent $f \in \mathbf{F}$ holds at time $T \in \mathbb{N}$ if fluents $g_1, \dots, g_n \in \mathbf{F}$ hold at the same time. In ASP, one rule is generated for each static causal law, ensuring that if $\text{holds}(g_1, T), \dots, \text{holds}(g_n, T)$ hold, then $\text{holds}(f, T)$ is inferred. The predicates $\text{fluent}(f)$ and $\text{fluent}(g_1), \dots, \text{fluent}(g_n)$ declare fluents, and $\text{time}(T)$ ensures valid time progression.

```
holds(f,T) :- holds(g_1,T), ..., holds(g_n,T),
fluent(g_1), ..., fluent(g_n), fluent(f), time(T).
```

Inhibition Rule (as in [97]): An inhibition rule (*f_1, \dots, f_n inhibits a*) states that an action $a \in \mathbf{A}$ is prevented from occurring at time $T \in \mathbb{N}$ if fluents $f_1, \dots, f_n \in \mathbf{F}$ hold. In ASP, one rule is generated for each inhibition rule, ensuring that if $\text{holds}(f_1, T), \dots, \text{holds}(f_n, T)$ hold, then $\text{holds}(\text{ab}(\text{occurs}(a), T))$ is inferred, marking the action as inhibited. The predicates $\text{fluent}(f_1), \dots, \text{fluent}(f_n)$ declare fluents, $\text{action}(a)$ declares the action, and $\text{time}(T)$ ensures valid time progression.

`holds(ab(occurs(a)),T) :- holds(f_1,T), ..., holds(f_n,T),
action(a), fluent(f_1), ..., fluent(f_n), time(T).`

As in \mathcal{C}_{TAID} , unlike standard ASP planning encodings, we do not use a choice rule for action generation, such as

`{ holds(occurs(A), T): action(A) } = 1 :- T = 1..t_max.` Instead, actions are determined by logical conditions: they occur if they are explicitly triggered, allowed, and not inhibited. This ensures that only valid and necessary actions are selected, avoiding unnecessary non-determinism.

Triggering Rule (as in [97]): A triggering rule (f_1, \dots, f_n **triggers** a) states that an action $a \in \mathbf{A}$ occurs at time $T \in \mathbb{N}$ if fluents $f_1, \dots, f_n \in \mathbf{F}$ hold and no inhibition rule is active. In ASP, one rule is generated for each triggering rule, ensuring that if `holds(f_1,T), ..., holds(f_n,T)` hold and `not holds(ab(occurs(a,T)))`, then `holds(occurs(a,T))` is inferred. The predicates `fluent(f_1), ..., fluent(f_n)` declare fluents, `action(a)` declares the action, and `time(T)` ensures valid time progression.

`holds(occurs(a),T) :- not holds(ab(occurs(a)),T),
holds(f_1,T), ..., holds(f_n,T),
fluent(f_1), ..., fluent(f_n), action(a), time(T).`

Allowance Rule (as in [97]): An allowance rule (f_1, \dots, f_n **allows** a) states that an action $a \in \mathbf{A}$ is permitted to occur at time $T \in \mathbb{N}$ if fluents $f_1, \dots, f_n \in \mathbf{F}$ hold and no inhibition rule is active. In ASP, one rule is generated for each allowance rule, ensuring that if `holds(f_1,T), ..., holds(f_n,T)` hold and `not holds(ab(occurs(a,T)))`, then `holds(allow(occurs(a),T))` is inferred. The predicates `fluent(f_1), ..., fluent(f_n)` declare fluents, `action(a)` declares the action, and `time(T)` ensures valid time progression.

`holds(allow(occurs(a)),T) :- not holds(ab(occurs(a)),T),
holds(f_1,T), ..., holds(f_n,T), fluent(f_1), ..., fluent(f_n),
action(a), time(T).`

Ensure Exogenous Actions Can Always Occur (as in [97]): This rule ensures that an action $a \in \mathbf{A}$ is always permitted to occur at any time step $T \in \mathbb{N}$, regardless of specific preconditions. In ASP, this is encoded by inferring `holds(allow(occurs(a),T))` for every declared action `action(a)` and time point `time(T)`. This guarantees that exogenous actions remain available throughout the execution.

`holds(allow(occurs(a)),T) :- action(a), time(T).`

No-Concurrency Constraint (as in [97]): The actions $a_1, \dots, a_n \in \mathbf{A}$ cannot occur simultaneously at time $T \in \mathbb{N}$. In ASP, this is encoded as a constraint that ensures selecting two or more actions at T leads to inconsistency. The rule applies over declared actions `action(a_1), ..., action(a_n)` and time points `time(T)`.

```

:- time(T), 2 {holds(occurs(a_1),T) :
action(a_1), ..., holds(occurs(a_n),T) : action(a_n)}.

```

Encoding of the Action Observation Language

We consider the encoding of \mathcal{C}_{TAID} action observation language, which we extend with components of \mathcal{C}_{MT} . The action observation language allows for the specification of observations about fluents and actions over time. Observations provide constraints on the initial state and subsequent state transitions, ensuring that execution traces align with known facts.

Initial State Fluent Observations (as in [97]): A fluent $f \in \mathbf{F}$ that holds in the initial state at time $T = 0$ is directly asserted. In ASP, this is represented by `holds(f,0)`, ensuring that observed fluents are set at the start of execution, forming the basis for reasoning about subsequent state transitions.

```
holds(f,0).
```

Fluent Observations for Other States (as in [97]): If a fluent $f \in \mathbf{F}$ is observed to hold at time $T \in \mathbb{N}$, it must be enforced. In ASP, this is encoded as a constraint ensuring that `holds(f,T)` must be true whenever f is observed. The rule applies over declared fluents `fluent(f)` and time points `time(T)`.

```
:- not holds(f,T), fluent(f), time(T).
```

Generate Possible Completions of the Initial State (as in [97]): A fluent $f \in \mathbf{F}$ in the initial state at $T = 0$ must either hold or its negation $\neg f$ must hold, but not both. In ASP, this is encoded using a default completion principle, ensuring that `holds(f,0)` is inferred unless `holds(neg(f),0)` is explicitly stated, and vice versa. The rule applies over declared fluents `fluent(f)` at time step `time(0)`.

```
holds(f,0) :- not holds(neg(f),0).
holds(neg(f),0) :- not holds(f,0).
```

Exogenous Action Observations (as in [97]): An exogenous action $a \in \mathbf{A}$ that occurs at time $T \in \mathbb{N}$ is recorded as a fact. This corresponds to an external execution of a , independent of triggering, allowance, or inhibition conditions. In ASP, this is represented by asserting `holds(occurs(a),T)`, ensuring that the action is registered as having taken place. The rule applies over declared actions `action(a)` and time points `time(T)`.

```
holds(occurs(a),T).
```

Observed Non-Occurrence of Actions (as in [97]): If an action $a \in \mathbf{A}$ does not occur at time $T \in \mathbb{N}$, this must be explicitly recorded. This corresponds to cases where no triggering or allowance rule is active, or an inhibition

rule prevents execution. In ASP, this is enforced as a constraint, ensuring that `holds(neg(occurs(a),T))` is inferred when the action does not occur. The rule applies over declared actions `action(a)` and time points `time(T)`.

```
:- not holds(neg(occurs(a)),T), action(a), time(T).
```

Action Execution Conditions (as in [97]): An action $a \in \mathbf{A}$ occurs at time $T \in \mathbb{N}$ if at least one allowance rule (f_1, \dots, f_n **allows** a) is active, no inhibition rule (f_1, \dots, f_n **inhibits** a) applies, and no external constraint negates its occurrence. In ASP, this is encoded by ensuring that `holds(occurs(a),T)` is inferred under these conditions. Additionally, if an action does not occur, its negation `holds(neg(occurs(a),T))` is inferred. The rule applies over declared actions `action(a)` and time points `time(T)`, where $T < t_{\max}$.

```
holds(occurs(a),T) :- holds(allow(occurs(a)),T),
not holds(ab(occurs(a)),T), not holds(neg(occurs(a)),T),
action(a), time(T), T < t_max.
```

```
holds(neg(occurs(a)),T) :- not holds(occurs(a),T),
action(a), time(T), T < t_max.
```

Encoding of the Action Query Language

We consider the encoding of \mathcal{C}_{TAID} action query language, which we extend with components of \mathcal{C}_{MT} .

Goal Achievement Constraint (as in [97]): The goal must be achieved in every valid plan. This corresponds to ensuring that at least one fluent configuration satisfying the goal holds. In ASP, this is enforced as a constraint ensuring that `achieved` is inferred.

```
:- not achieved.
```

Initial Goal Satisfaction (as in [97]): The goal is considered achieved if it already holds at the initial time step $T = 0$.

```
achieved :- achieved(0).
```

Persistence of Goal Achievement (as in [97]): If the goal is achieved at time $T + 1$ but was not achieved at T , then it remains achieved for all subsequent time steps. The predicates `time(T)` ensure valid time progression.

```
achieved :- achieved(T+1), not achieved(T), time(T), time(T+1).
```

Fluent-Based Goal Satisfaction (as in [97]): The goal is achieved at time T if a set of fluents $f_1, \dots, f_n \in \mathbf{F}$ required for goal satisfaction hold. This corresponds to a static causal law of the form (f_1, \dots, f_n **if** `goal_conditions`). In ASP, this is encoded by inferring `achieved(T)` when `holds(f1,T), ..., holds(fn,T)` hold.

```

achieved(T) :- holds(f_1, T), ..., holds(f_n, T),
achieved(T+1), fluent(f_1), ..., fluent(f_n), time(T), time(T+1).

```

Final Goal Satisfaction (as in [97]): The goal must hold at the maximum time step t_{\max} . This ensures that the trajectory satisfies the required final state.

```

achieved(t_max) :- holds(f_1, t_max), ..., holds(f_n, t_max),
fluent(f_1), ..., fluent(f_n).

```

Execution of Allowed Actions (as in [97]): An action $a \in \mathbf{A}$ occurs at time $T \in \mathbb{N}$ if it is allowed by an active allowance rule (f_1, \dots, f_n **allows** a), is not inhibited by an inhibition rule (f_1, \dots, f_n **inhibits** a), and the goal has not yet been achieved. In ASP, this ensures that `holds(occurs(a), T)` is inferred only when these conditions hold. The predicates `fluent(f1), ..., fluent(fn)` declare fluents, `action(a)` declares actions, and `time(T)` defines time steps.

```

holds(occurs(a), T) :- holds(allow(occurs(a)), T), not achieved(T),
not holds(ab(occurs(a)), T), not holds(neg(occurs(a)), T),
action(a), time(T).

```

Explicit Non-Occurrence of Actions (as in [97]): If an action $a \in \mathbf{A}$ does not occur at time $T \in \mathbb{N}$, this must be explicitly recorded. This corresponds to enforcing that an action not chosen in the trajectory is negated.

```

holds(neg(occurs(a)), T) :- not holds(occurs(a), T),
action(a), time(T).

```

Encoding of the Mental State Language Extension

We now proceed by presenting the translations for static and dynamic causal laws for mental state specifications introduced in \mathcal{C}_{MT} .

Dynamic causal law for Mental Fluents (introduced in \mathcal{C}_{MT}): A dynamic causal law (a **influences** f_1^h, \dots, f_n^h **if** g_1, \dots, g_m) specifies that an action $a \in \mathbf{A}$ causes mental fluents $f_1^h, \dots, f_n^h \in \mathbf{F}^H$ to hold at $T + 1$ if fluents $g_1, \dots, g_m \in \mathbf{F}$ hold at $T \in \mathbb{N}$. The ASP encoding generates a rule for each such statement.

```

holds(f, T+1) :- holds(occurs(a), T),
holds(g_1, T), ..., holds(g_m, T),
fluent(g_1), ..., fluent(g_m), mental_fluent(f),
action(a), time(T).

```

Static causal law for Mental Fluents (introduced in \mathcal{C}_{MT}): A static causal law (g_1, \dots, g_m **influences** f_1^h, \dots, f_n^h) states that if fluents $g_1, \dots, g_m \in \mathbf{F}$ hold at time T , then mental fluents $f_1^h, \dots, f_n^h \in \mathbf{F}^H$ also hold at $T \in \mathbb{N}$. The ASP encoding generates a rule for each such statement.

```

holds(f, T) :- holds(g_1, T), ..., holds(g_m, T),
               fluent(g_1), ..., fluent(g_m), mental_fluent(f),
               time(T).

```

The following translations regard contravening and facilitating rules (introduced in \mathcal{C}_{MT}). Unlike the ‘inhibition’ and ‘allowance’ rules outlined in \mathcal{C}_{TAID} [97], which impact actions in the current time step, ‘contravenes’ and ‘facilitates’ relate particularly about a human action $a^h \in \mathbf{A}^H$ and mental fluents $f^h \in \mathbf{F}^H$. As we have previously discussed; appraisal theories have suggested that emotional states have distinctive “action tendencies” [249]. Capturing these semantics, while seemingly redundant due to the general rules in the action language, supports knowledge representation.

Facilitation Rule (introduced in \mathcal{C}_{MT}): A facilitation rule (g_1^h, \dots, g_m^h **facilitates** a^h) states that if mental fluents $g_1^h, \dots, g_m^h \in \mathbf{F}^H$ hold at time $T \in \mathbb{N}$, and the action $a^h \in \mathbf{A}^H$ is not inhibited, then a^h occurs. The ASP encoding ensures that `holds(occurs(a^h), T)` is inferred if `not holds(ab(occurs(a^h), T))` holds.

```

holds(occurs(a), T) :- not holds(ab(occurs(a)), T),
                      holds(g_1, T), ..., holds(g_m, T),
                      mental_fluent(g_1), ..., mental_fluent(g_m),
                      human_action(a), time(T).

```

Contravening Rule (introduced in \mathcal{C}_{MT}): A contravening rule (g_1^h, \dots, g_m^h **contravenes** a^h) states that if mental fluents $g_1^h, \dots, g_m^h \in \mathbf{F}^H$ hold at time $T \in \mathbb{N}$, then the human action $a^h \in \mathbf{A}^H$ is inhibited from occurring. The ASP encoding infers `holds(ab(occurs(a^h), T))` under these conditions.

```

holds(ab(occurs(a)), T) :- holds(g_1, T), ..., holds(g_m, T),
                          mental_fluent(g_1), ..., mental_fluent(g_m),
                          human_action(a), time(T).

```

Forbidding Rule (introduced in \mathcal{C}_{MT}): A forbidding rule (g_1^h, \dots, g_m^h **forbids to cause** f_1^h, \dots, f_n^h) enforces constraints on mental state transitions in the belief graph. If mental fluents $g_1^h, \dots, g_m^h \in \mathbf{F}^H$ hold at time $T \in \mathbb{N}$, then fluents $f_1^h, \dots, f_n^h \in \mathbf{F}^H$ are forbidden to hold at $T + 1$. This ensures that the belief graph’s constraints on state transitions are respected. The ASP encoding introduces an integrity constraint to prohibit transitions that violate these restrictions.

```

:- holds(f_1, T+1), ..., holds(f_n, T+1),
   holds(g_1, T), ..., holds(g_m, T),
   mental_fluent(f_1), ..., mental_fluent(f_n),
   mental_fluent(g_1), ..., mental_fluent(g_m),
   time(T).

```

4.4 Case Study: Emotional Reasoning

The section presents a characterization of emotion dynamics in \mathcal{C}_{MT} with specialized subsets of fluents and mental state constraints adapted for emotional reasoning. This is achieved by formalizing different emotion theories; the Appraisal theory of Emotion (AE) by Roseman (1996) [248], Hedonic Emotion Regulation (HER) [315], and Utilitarian Emotion Regulation (UER) [284], capturing links between human emotions and their underlying causes in the environment. To this end, subsets of mental state fluents, called emotion fluents, together with a set of action rules and sets of mental state constraints are specified, capturing principles from emotion theories. By following the psychological theory of AE, an emotion state-space is defined with 108 emotional configurations (states), through which a set of 16 basic human emotions, according to AE, can be represented and explained. By following the theories of HER and UER, different constraints for emotional-change are defined. HER focuses on augmenting positive emotions while diminishing negative ones. Conversely, UER seeks to promote particular emotions, including potentially negative ones, that enhance specific attributes, such as motivation or control, serving a utilitarian purpose in the long run. Consequently, HER and UER adopt contrasting principles for effecting emotional change. By examining the trajectories produced through the application of either HER or UER within the state-space of AE, we can compare and evaluate their respective behaviors.

Emotion Theories: AE, HER and UER

The **Appraisal theory of Emotion (AE)** by Roseman (1996) [248] proposes that emotions are caused by an appraisal of a situation in terms of 1) being consistent or inconsistent with needs, importance of the situation, 2) being consistent or inconsistent with goals, the attainability/potential to achieve goals, 3) who/what is accountable/caused the situation, which can be the environment, others, or oneself, and 4) as being easy or difficult to control. According to AE, the difference between goal consistency/attainability and need consistency/importance determines negative, stable and positive emotions. More intense negative emotions (e.g., Anger or Fear) arise when the need consistency is greater than the goal consistency, while less intense negative emotions can arise when both the need consistency and goal consistency are low. On the other hand, positive emotions (e.g., Joy or Liking) arise when the goal consistency is greater than the need consistency, or when both are high. By ranking consistency values as *Low* < *Undecided* < *High* and by looking at the difference between need and goal consistency, positive and negative emotions can be distinguished.

Hedonic Emotion Regulation (HER) [315] is a theory for regulating emotions, guided by the goals to increase positive emotion and decrease negative emotion. According to HER, both of these emotion regulation goals are associated with improved well-being, where decreasing of negative emo-

tion has been most effective [216]. The principles of HER can be applied to reason about emotional change. For instance, the relation between goal attainability and goal importance has been empirically explored, showing that goal attainability, rather than goal importance, was positively linked to well-being [51]. Another empirical study analyzed self-responsibility and emotions, and showed that accountability does not by itself affect whether the emotion is negative or positive. Nevertheless, when something else than when an individual (self/other) is perceived as accountable, less negative emotions appear [226]. Another empirical study examined the relationship between control potential and emotions [248]. The study found that the perception of high or low control potential influenced the experience of accommodating emotions or contending emotions, respectively. However, the study did not find a direct effect of control potential on the experience of positive or negative emotions.

By analyzing the four dimensions of emotion as proposed by AE (Need consistency, Goal consistency, Accountability, and Control potential), we can derive meaningful interpretations about the relationship between HER and AE in the context of hedonic emotional change. These interpretations are summarized in Table 4.1.

Utilitarian Emotion Regulation (UER) [284] is a theory emotion regulation that, in contrast to HER, is guided by the goals to experience emotions in the short-term for a long-term utilitarian purpose. This may involve accepting a temporary negative emotion to increase their capability of dealing with a situation. For instance, valuing long-term goals (i.e., focusing on need, importance and motivation) [96] over immediate pleasure makes individuals more likely to engage in UER. Furthermore, a study on goal-setting [265] found that when individuals perceive a discrepancy between their present performance and a desired goal, it can lead to dissatisfaction. This dissatisfaction, in turn, can serve as a motivator for increased effort and striving towards the goal. Another study, exploring accountability and emotion [18], suggests that self-accountability for a situation has utilitarian gains by having effects on the individual’s “locus of control”. Moreover, a study exploring control and emotion [286] found that increasing control potential, such as through an increased level of anger, can have utilitarian gains. Another study on the emotion of fear [285] proposes the concept of “fear as function for goal pursuit”, highlighting its role in motivating goal-directed behavior. Yet another study [224] suggests that “negative” emotions can modify an individual’s “readiness to think and act, and their potential functional utility”.

By analyzing the four dimensions of emotion as proposed by AE (Need consistency, Goal consistency, Accountability, and Control potential), we can derive meaningful interpretations about the relationship between UER and AE in the context of utilitarian emotional change. These interpretations are summarized in Table 4.2.

Following the interpretations of HER and UER, we can define transition constraints in an emotion specialized BG. This will allow certain sets of trajectories depending on which emotional-change theory we model. In both cases,

Table 4.1: Hedonic Emotion Regulation as basis for Principles of Change

Hedonic Emotion Regulation	Principle of Change
Need Consistency/Importance: A high need consistency is associated with positive emotions only when the goal consistency is high, but negative if the goal consistency is low (supported by [51]).	High need consistency maximizes positive emotion only when goal consistency is high.
Goal Consistency/Attainability: A high goal consistency is associated with positive emotions. In particular when the need consistency is high (supported by [51]).	High goal consistency maximizes positive emotion.
Accountability: Accountability does not affect whether the emotion is negative or positive. Nevertheless, when the environment is perceived as primarily accountable for the situation, less negative emotions appear than when an individual (self/other) is perceived as accountable (supported by [226]).	Accountability to environment maximizes positive emotions.
Control Potential: Control Potential does not affect whether the emotion is negative or positive. Nevertheless, a control potential above low is often associated with accommodating emotions, closer to positive emotions (supported by [248]).	High control potential maximizes positive emotion when there is a non-negative balance between goal consistency and need consistency, i.e., when goal consistency is equal or higher than need consistency.

we use the state space defined by AE.

In the following section, we present a formalization of these theories in terms of \mathcal{C}_{MT} . We first define an emotion state in terms of the dimensions of AE. We then define a set of transition constraints that restrict particular configurations (states) of AE-appraisals to arise. Two sets of transition constraints are defined, constituting two different sub-graphs of valid transitions in the AE state-space, a so-called *hedonic emotion state* and an *utilitarian emotion state*, based on HER and UER principles, respectively.

Formalizing Emotional Reasoning

Components of AE are formalized as a specialized Belief Graph (see Definition 18), called an emotion graph (EG), to reason about emotion states and

(hedonic contra utilitarian) emotional change to reduce unintended emotional side-effects. The constraints of the EG serve as safety restrictions for emotion-influencing actions, applicable for different types of emotional reasoning settings.

Recall that AE defines emotions as a composition of an individual’s appraisal of a situation, in terms of need consistency (importance), goal consistency (attainability), accountability (who/what) and control potential. By following this definition of emotional causes, we can define *emotion fluent*, a changeable emotion variable.

Definition 28 (Emotion fluent) *An emotion fluent is a mental fluent $f(c, v)$, a ground atom of arity 2, where $c \in C$, and $v \in V_c$, such that $C = \{ne, go, ac, co\}$ is a set of constants denoting psychological classes and $V = \{V_{ne}, V_{go}, V_{ac}, V_{co}\}$ a set of sets of constants denoting psychological values for each class of C , where:*

$$\begin{aligned} V_{ne} &= \{high, low, undecided\} \\ V_{go} &= \{high, low, undecided\} \\ V_{ac} &= \{self, other, environment, undecided\} \\ V_{co} &= \{high, low, undecided\}. \end{aligned}$$

where *ne* denotes need consistency, *go* denotes goal consistency, *ac* denotes accountability and *co* denotes control potential.

By defining a set of emotions following AE in this way, and by utilizing different principles of emotion regulation, we can specify preferred (e.g., hedonic or utilitarian) transitions between emotion states. In the following subsection, we specify an EG to reason about emotional transitions.

Emotion Graph (EG)

Following the AE-theory, 16 emotion states are specified, corresponding to each basic emotion: {Anger, Dislike, Disgust, Sadness, Hope, Frustration, Fear, Distress, Joy, Liking, Pride, Surprise, Relief, Regret, Shame, Guilt}. It is important to note that various other emotion states can be defined (in total 108) by different combinations of the four AE dimensions, and it is not necessary, nor always preferred, to assign a specific label, such as “Joy”, to each of them. While labels aid in expressing these states in a human readable way, they do not inherently contribute to the functionality of a system’s reasoning. It is worth noting that emotional expressions vary across individuals, cultures, languages, and other factors. Therefore, it is crucial for the system to analyze emotion states using a multidimensional format. We can model the states and transitions as an EG that represents valid emotional change given a recognized emotion state configuration. In Fig. 4.3, examples of emotion states are presented.

Anger	Hope	Joy	Relief
ne:h go:l ac:o co:h	ne:u go:h ac:e co:l	ne:h go:h ac:e co:u	ne:h go:h ac:e co:u
Dislike	Frustration	Liking	Regret
ne:u go:l ac:o co:l	ne:h go:l ac:e co:h	ne:u go:h ac:o co:u	ne:u go:l ac:s co:l
Disgust	Fear	Pride	Shame
ne:l go:l ac:e co:h	ne:u go:l ac:e co:l	ne:u go:h ac:s co:u	ne:l go:l ac:s co:h
Sadness	Distress	Surprise	Guilt
ne:h go:l ac:e co:l	ne:l go:l ac:e co:l	ne:u go:u ac:e co:u	ne:h go:h ac:s co:h

Figure 4.3: Emotion states following the Appraisal theory of Emotion by Roseman (1996) [248]. Each emotion state is here expressed by an intuitive “emotion” label on top, and an appraisal configuration consisting of a set of variable:value pairs below. The variables are: ne = need consistency, go = goal consistency, ac = accountable, co = control potential. The values are: l = low, h = high, u = undecided, o = other, s = self, e = environment.

Definition 29 (emotion graph) *An emotion graph (EG) is a specialized Belief Graph $EG = \langle S, E \rangle$, where S is a set of nodes representing emotion states and $E \subseteq S \times S$ is a set of directed edges representing transitions between emotion states. Each emotion state $s \in S$ is a set of emotion fluents.*

The emotion graph is captured by a given program specified by the semantics of the action language \mathcal{C}_{MT} , with constraints in terms of HER or UER, serving as restrictions for hedonic contra utilitarian emotional change, presented in the following section.

Action language emotion specification

The \mathcal{C}_{MT} alphabet is specialized with an emotion related vocabulary and causal laws to specify fluents and actions according to AE-theory.

The emotional reasoning semantics is characterized by the constraints of the EDG, specified through a set of static causal laws and mental state constraints. In this way, we can restrict states and state-transitions to comply with principles of emotional change.

For any particular application, we need to define an EG that, based on application specific interaction goals and relevant theories for emotion regulation, avoids unintended mental states. This specifies an EG with a subset of transitions (in the fully connected graph) that is considered valid. We define a hedonic theory specification that follows principles of HER, aiming to increase positive emotion and decrease negative emotion (see more details of HER in Section 4.4). In an abstraction of HER, the hedonic emotion specification is defined to not allow entering a negative emotion state and to preserve a positive balance in emotion.

Definition 30 (Hedonic emotion theory specification) *A hedonic theory specification is a collection of causal laws as follows:*

1. $f(c_{ne}, high)$ forbids to cause $f(c_{go}, low)$.
2. $f(c_{ne}, high)$ forbids to cause $f(c_{go}, undecided)$.
3. $f(c_{ne}, high)$ forbids to cause $f(c_{go}, high)$.
4. $f(c_{ne}, undecided)$ forbids to cause $f(c_{go}, low)$.
5. $f(c_{ne}, undecided)$ forbids to cause $f(c_{go}, undecided)$.
6. $f(c_{ne}, undecided)$ forbids to cause $f(c_{go}, high)$.
7. $f(c_{go}, high)$ forbids to cause $f(c_{go}, low)$.
8. $f(c_{go}, high)$ forbids to cause $f(c_{go}, undecided)$.
9. $f(c_{go}, undecided)$ forbids to cause $f(c_{go}, low)$.
10. $f(c_{go}, low)$ forbids to cause $f(c_{co}, high)$.
11. $\{f(c_{ne}, high), f(c_{go}, low), f(c_{ac}, other)\}$ forbids to cause $f(c_{co}, high)$.
12. $\{f(c_{ne}, high), f(c_{go}, undecided), f(c_{ac}, other)\}$ forbids to cause $f(c_{co}, high)$.
13. $\{f(c_{ne}, high), f(c_{go}, low), f(c_{ac}, self)\}$ forbids to cause $f(c_{co}, high)$.
14. $\{f(c_{ne}, high), f(c_{go}, undecided), f(c_{ac}, self)\}$ forbids to cause $f(c_{co}, high)$.
15. $\{f(c_{ne}, high), f(c_{go}, low), f(c_{ac}, undecided)\}$ forbids to cause $f(c_{co}, high)$.
16. $\{f(c_{ne}, high), f(c_{go}, undecided), f(c_{ac}, undecided)\}$ forbids to cause $f(c_{co}, high)$.

We further define an utilitarian emotion theory specification that follows principles of UER, aiming to prioritize particular emotional dimensions (e.g., need_consistency) that are associated with increased utilitarian gain (see more details of UER in Section 4.4). In an abstraction of UER, the utilitarian emotion theory specification is defined to allow entering a negative emotion state in the short-term if it may result in a long-term utilitarian gain.

Definition 31 (Utilitarian emotion theory specification) *A utilitarian theory specification is a collection of causal laws as follows:*

1. $f(c_{ne}, low)$ forbids to cause $f(c_{ne}, undecided)$.
2. $f(c_{ne}, low)$ forbids to cause $f(c_{ne}, high)$.
3. $f(c_{ne}, undecided)$ forbids to cause $f(c_{ne}, low)$.
4. $f(c_{ne}, undecided)$ forbids to cause $f(c_{ne}, high)$.
5. $f(c_{go}, low)$ forbids to cause $f(c_{ne}, low)$.
6. $f(c_{go}, low)$ forbids to cause $f(c_{ne}, undecided)$.
7. $f(c_{go}, undecided)$ forbids to cause $f(c_{ne}, low)$.
8. $f(c_{go}, undecided)$ forbids to cause $f(c_{ne}, undecided)$.
9. $f(c_{go}, high)$ forbids to cause $f(c_{ne}, low)$.
10. $f(c_{go}, high)$ forbids to cause $f(c_{ne}, undecided)$.
11. $f(c_{go}, high)$ forbids to cause $f(c_{ne}, high)$.
12. $f(c_{ac}, undecided)$ forbids to cause $f(c_{ac}, self)$.
13. $f(c_{ac}, undecided)$ forbids to cause $f(c_{ac}, other)$.
14. $f(c_{ac}, undecided)$ forbids to cause $f(c_{ac}, environment)$.
15. $f(c_{ac}, other)$ forbids to cause $f(c_{ac}, undecided)$.
16. $f(c_{ac}, other)$ forbids to cause $f(c_{ac}, self)$.
17. $f(c_{ac}, other)$ forbids to cause $f(c_{ac}, environment)$.
18. $f(c_{ac}, environment)$ forbids to cause $f(c_{co}, low)$.
19. $f(c_{ac}, environment)$ forbids to cause $f(c_{co}, undecided)$.
20. $f(c_{co}, low)$ forbids to cause $f(c_{co}, undecided)$.
21. $f(c_{co}, low)$ forbids to cause $f(c_{co}, high)$.
22. $f(c_{co}, undecided)$ forbids to cause $f(c_{co}, low)$.
23. $f(c_{co}, undecided)$ forbids to cause $f(c_{co}, high)$.

These specifications enforce a structured adjustment of appraisal values, ensuring that emotional transitions occur in a controlled manner to prevent unintended effects. Specifically, they regulate the dynamics of emotional change according to the underlying emotion regulation principles—hedonic emotion regulation (HER) and utilitarian emotion regulation (UER). In Section 4.5, we analyze the formal properties of the action language to prove controlled emotional change. Before we proceed with the formal analysis, we present a prototype of the EG to support empirical analysis.

Implementation example: Emotional Reasoning

We present an implementation designed to capture the constraints of a specific EDG, based on AE-theory. The encoding focuses on how sequences of actions can influence the transition of emotion states from an initial state to a goal state. The actions involved in this prototype are translated to a predicate `holds(occurs(influence(MS,V)),T)`, such that `MS` is one of the four psychological classes in AE-theory, `V` is a psychological value in accordance with AE-theory (following Definition 16). Note that a time point `T` is included to define the index of each action in a trajectory. Similarly, fluents are implemented as the predicate `holds(mental_fluent(MS,V),T)`, where `MS` is a psychological class, `V` is a psychological value, and `T` is a time point. Influence actions are represented as `occurs(influence(MS,V))`, and their execution is tracked using the predicate `holds(occurs(influence(MS,V)),T)`, which specifies that the action occurred at time `T`. When an influence action occurs at `T`, it modifies the corresponding fluent at `T+1`.

Thus, the implementation supports influence actions of the form:

- `holds(occurs(influence(need_consistency,V)),T)`
- `holds(occurs(influence(goal_consistency,V)),T)`
- `holds(occurs(influence(control_potential,V)),T)`
- `holds(occurs(influence(accountability,V)),T)`

Particular action names corresponding to real-world interventions, e.g., ‘providing guides to increase control potential’, are not part of this implementation example. Instead, only transition constraints for the four *types* of AE-based actions are encoded.

The psychological classes and their valid values are defined in Listing 4.1. In Listing 4.2, we introduce a logic program P_{EG} that characterizes the state space of the emotion graph. In Listing 4.3 and Listing 4.4, two additional sets of integrity constraints are implemented, specialized for HER-based reasoning and UER-based reasoning, respectively. These integrity constraints are independent of the main logic program.

Let us clarify that the code in Listing 4.1-4.2 is combined into one ASP program, while either Listing 4.3 (HER-based constraints) or Listing 4.4 (UER-based constraints) are included in the program.

For instance, the HER-based constraint “Need consistency can only be increased if the goal consistency is high” can be expressed by the following integrity constraints (excerpt from Listing 4.3):

```
:- holds(mental_fluent(need_consistency, high), T+1),
    holds(mental_fluent(goal_consistency, low), T), time(T).
:- holds(mental_fluent(need_consistency, high), T+1),
    holds(mental_fluent(goal_consistency, undecided), T), time(T).
:- holds(mental_fluent(need_consistency, undecided), T+1),
    holds(mental_fluent(goal_consistency, low), T), time(T).
:- holds(mental_fluent(need_consistency, undecided), T+1),
    holds(mental_fluent(goal_consistency, undecided), T), time(T).
```

Similarly, the UER-based constraint “Need consistency can only be influenced to high” can be expressed by the following integrity constraints (excerpt from Listing 4.4):

```
:- holds(mental_fluent(need_consistency, low), T+1),
    holds(mental_fluent(need_consistency, undecided), T), time(T).
:- holds(mental_fluent(need_consistency, undecided), T+1),
    holds(mental_fluent(need_consistency, high), T), time(T).
```

By extending the program with different sets of integrity constraints, based on different emotion-regulation principles (e.g., HER or UER), we can analyze and compare their respective trajectories. We return to this comparison in our empirical analysis.

Listing 4.1: Definition of psychological classes and values.

```
1 % Psychological values
2 psychological_value(low).
3 psychological_value(undecided).
4 psychological_value(high).
5 psychological_value(other).
6 psychological_value(self).
7 psychological_value(environment).
8
9 % Psychological classes representing mental states
10 psychological_class(need_consistency).
11 psychological_class(goal_consistency).
12 psychological_class(accountability).
13 psychological_class(control_potential).
```

Let P_{EG} be the logic program in Listing 4.2:

Listing 4.2: Logic Program P_{EG} . Implementation of the emotion graph.

```
1 % Define mental fluents
2 mental_fluent(MS, V) :- wellformed_emotion_fluent(MS, V).
3
4 % Initial fluents
```

```

5 holds(mental_fluent(MS, V), 0) :- init_on(MS, V).
6
7 % If influence action occurs at T, it changes the mental fluent at T+1
8 holds(mental_fluent(MS, V), T+1) :-
9     holds(occurs(influence(MS, V)), T), time(T).
10
11 % Persistence: If no influence occurs, the previous value persists
12 holds(mental_fluent(MS, V), T+1) :-
13     holds(mental_fluent(MS, V), T),
14     not holds(occurs(influence(MS, _)), T), time(T).
15
16 % Ensure that each mental state has exactly one value at any time step
17 :- #count { V : holds(mental_fluent(MS, V), T) } != 1,
18     psychological_class(MS), time(T).
19
20 % Ensure that all goal states are reached by the final time step
21 :- goal_on(MS, V), not holds(mental_fluent(MS, V), t_max).

```

The EG transitions can be further constrained by considering HER or UER, implemented in terms of integrity constraints.

Listing 4.3: Integrity Constraints for Hedonic Emotion Regulation.

```

1 % HEDONIC EMOTION REGULATION
2 % Need consistency can only increase if goal consistency is high
3 :- holds(mental_fluent(need_consistency, high), T+1),
4     holds(mental_fluent(goal_consistency, low), T), time(T).
5 :- holds(mental_fluent(need_consistency, high), T+1),
6     holds(mental_fluent(goal_consistency, undecided), T), time(T).
7 :- holds(mental_fluent(need_consistency, undecided), T+1),
8     holds(mental_fluent(goal_consistency, low), T), time(T).
9 :- holds(mental_fluent(need_consistency, undecided), T+1),
10     holds(mental_fluent(goal_consistency, undecided), T), time(T).
11
12 % Goal consistency cannot decrease
13 :- holds(mental_fluent(goal_consistency, low), T+1),
14     holds(mental_fluent(goal_consistency, V1), T), V1 != low, time(T).
15 :- holds(mental_fluent(goal_consistency, undecided), T+1),
16     holds(mental_fluent(goal_consistency, V1), T), V1 != undecided, time(T).
17
18 % Prevent high control potential if goal consistency is low
19 :- holds(mental_fluent(control_potential, high), T+1),
20     holds(mental_fluent(goal_consistency, low), T), time(T).
21
22 % Prevent control potential from increasing to high
23 % when accountability is not the environment
24 % and there is a negative balance between goal and need consistency
25 :- holds(mental_fluent(control_potential, high), T+1),
26     holds(mental_fluent(need_consistency, high), T),
27     holds(mental_fluent(goal_consistency, low), T),
28     holds(mental_fluent(accountability, other), T), time(T).
29
30 :- holds(mental_fluent(control_potential, high), T+1),
31     holds(mental_fluent(need_consistency, high), T),
32     holds(mental_fluent(goal_consistency, undecided), T),
33     holds(mental_fluent(accountability, other), T), time(T).

```

```

34
35 :- holds(mental_fluent(control_potential, high), T+1),
36     holds(mental_fluent(need_consistency, high), T),
37     holds(mental_fluent(goal_consistency, low), T),
38     holds(mental_fluent(accountability, self), T), time(T).
39
40 :- holds(mental_fluent(control_potential, high), T+1),
41     holds(mental_fluent(need_consistency, high), T),
42     holds(mental_fluent(goal_consistency, undecided), T),
43     holds(mental_fluent(accountability, self), T), time(T).
44
45 :- holds(mental_fluent(control_potential, high), T+1),
46     holds(mental_fluent(need_consistency, high), T),
47     holds(mental_fluent(goal_consistency, low), T),
48     holds(mental_fluent(accountability, undecided), T), time(T).
49
50 :- holds(mental_fluent(control_potential, high), T+1),
51     holds(mental_fluent(need_consistency, high), T),
52     holds(mental_fluent(goal_consistency, undecided), T),
53     holds(mental_fluent(accountability, undecided), T), time(T).

```

Listing 4.4: Integrity Constraints for Utilitarian Emotion Regulation.

```

1 % UTILITARIAN EMOTION REGULATION
2 % Need consistency can only be influenced to high.
3 :- holds(mental_fluent(need_consistency, low), T+1),
4     holds(mental_fluent(need_consistency, V1), T), V1 != low, time(T).
5 :- holds(mental_fluent(need_consistency, undecided), T+1),
6     holds(mental_fluent(need_consistency, V1), T), V1 != undecided, time(T).
7
8 % The goal consistency can only be influenced to low or undecided,
9 % and the need consistency must be high.
10 :- holds(mental_fluent(goal_consistency, low), T+1),
11     holds(mental_fluent(need_consistency, low), T), time(T).
12 :- holds(mental_fluent(goal_consistency, low), T+1),
13     holds(mental_fluent(need_consistency, undecided), T), time(T).
14 :- holds(mental_fluent(goal_consistency, undecided), T+1),
15     holds(mental_fluent(need_consistency, low), T), time(T).
16 :- holds(mental_fluent(goal_consistency, undecided), T+1),
17     holds(mental_fluent(need_consistency, undecided), T), time(T).
18 :- holds(mental_fluent(goal_consistency, high), T+1),
19     holds(mental_fluent(need_consistency, undecided), T), time(T).
20 :- holds(mental_fluent(goal_consistency, high), T+1),
21     holds(mental_fluent(need_consistency, high), T), time(T).
22
23 % Accountability can only be influenced to self or environment.
24 % If environment, then the control must be high.
25 :- holds(mental_fluent(accountability, undecided), T+1),
26     holds(mental_fluent(accountability, V1), T), V1 != undecided, time(T).
27 :- holds(mental_fluent(accountability, other), T+1),
28     holds(mental_fluent(accountability, V1), T), V1 != other, time(T).
29 :- holds(mental_fluent(accountability, environment), T+1),
30     holds(mental_fluent(control_potential, low), T), time(T).
31 :- holds(mental_fluent(accountability, environment), T+1),
32     holds(mental_fluent(control_potential, undecided), T), time(T).
33

```

```
34 % Control potential can only be influenced to High.
35 :- holds(mental_fluent(control_potential, low), T+1),
36     holds(mental_fluent(control_potential, V1), T), V1 != low, time(T).
37 :- holds(mental_fluent(control_potential, undecided), T+1),
38     holds(mental_fluent(control_potential, V1), T), V1 != undecided, time(T).
```

The logic program P_{EG} , coupled with integrity constraints based on suitable emotion regulation principles, helps to filter the available actions of a system to deal with emotions in an interaction. Let us proceed with formal analysis before conducting experimental evaluation.

Table 4.2: Utilitarian Emotion Regulation as basis for Principles of Change

Utilitarian Emotion Regulation	Principle of Change
<p>Need Consistency/Importance: Valuing long-term goals over immediate pleasure makes individuals more likely to engage in UER. They are willing to endure temporary negative emotions to maximize utility and achieve important goals (supported by [96]).</p>	<p>High need consistency maximizes utility.</p>
<p>Goal Consistency/Attainability: Perceiving important goals as difficult to achieve in the short-term can motivate individuals to engage in UER strategies to increase their chances of long-term success (supported by [265]).</p>	<p>Low or Undecided goal consistency maximizes utility if the need consistency is high.</p>
<p>Accountability: Individuals perceiving themselves as primarily accountable for their long-term goals drive them to regulate emotions in a utilitarian manner. Taking responsibility for emotional experiences enhances their ability to deal with challenges effectively. Furthermore, if the control potential is high, the environment can be perceived as accountable, where an individual may be able to affect aspects of the environment for utilitarian gains (supported by [18]).</p>	<p>Accountability to self maximizes utility. Accountability to environment maximizes utility if the control potential is high.</p>
<p>Control Potential: Sense of control is an important utilitarian characteristic. Accepting and regulating temporary negative emotions to increase perceived control contribute to long-term goal attainment (supported by [286]).</p>	<p>High control potential maximizes utility.</p>

4.5 Formal Analysis

We start by establishing a link between the introduced formal definitions and answer set semantics. This lets us define the semantics of \mathcal{C}_{MT} in terms of answer sets, supporting implementations of the action language in Answer Set Programming (ASP). In the proceeding parts of the section, we analyze Safety properties (considering the invariance principle [130]) to be preserved in the generated trajectories.

Analysis of translation to answer set semantics

In order to define the semantics of \mathcal{C}_{MT} , we characterize trajectory models in terms of answer sets.

This is formalized by the following theorem:

Theorem 1 *Let $(D^{MT}, O_{initial})$ be an action theory such that $O_{initial}$ are the fluent observations of a dynamic environment in the initial state. Let Q be a query, according to Definition 27 and let*

$$A_Q = \{(a \text{ occurs_at } t_i) \mid a \in A_i, 1 \leq i \leq m\}.$$

Let $\mathcal{P}_{MT}(D^{MT}, O_{initial} \cup A_Q)$ be a \mathcal{C}_{MT} logic program according to the translation in Section 4.3.

1. If there is a trajectory model $\langle s_0, A_1, s_1, A_2, \dots, A_n, s_m \rangle$ where $A_i \subseteq \mathbf{A}$ ($0 \leq i \leq m$) of $\mathcal{C}_{MT}(D^{MT}, O_{initial} \cup A_Q)$,

then there is an answer set \mathcal{A} of logic program \mathcal{P}_{MT} such that for all fluents $f \in \mathbf{F}^E \cup \mathbf{F}^H$ and for all mental fluent $f(c, v) \in \mathbf{F}^H$ at the time points $0 \leq k \leq m$ the following statements hold true:

- (a) $holds(f, k) \in \mathcal{A}$, if $f \in s_k$,*
- (b) $holds(neg(f), k) \in \mathcal{A}$, if $f \notin s_k$*
- (c) $holds(occurs(a), k) \in \mathcal{A}$, if $a \in A_{k+1}$*
- (d) $holds(neg(occurs(a)), k) \in \mathcal{A}$, if $a \notin A_{k+1}$,*

2. If there is an answer set \mathcal{A} of logic program \mathcal{P}_{MT} and at time point $0 \leq k \leq m$

- (a) $s_k = \{f \mid holds(f, k) \in \mathcal{A}\} \cup \{\neg f \mid holds(neg(f), k) \in \mathcal{A}\}$*
- (b) $A_{k+1} = \{a \mid holds(occurs(a), k) \in \mathcal{A}\}$*

then there is a trajectory model $\langle s_0, A_1, s_1, A_2, \dots, A_m, s_m \rangle$ of $\mathcal{C}_{MT}(D^{MT}, O_{initial} \cup A_Q)$.

Proof: We will prove the theorem by constructing the necessary answer set and showing that it corresponds to a trajectory model, and vice versa. The proof will be divided into two parts, corresponding to each direction of the theorem

statement.

Part 1; Trajectory Model to Answer Set:

Assume there is a trajectory model $\langle s_0, A_1, s_1, A_2, \dots, A_n, s_m \rangle$ where $A_i \subseteq \mathbf{A}$ ($0 \leq i \leq m$) of \mathcal{C}_{MT} ($D^{MT}, O_{initial} \cup A_Q$).

Constructing the Answer Set: Define a set of atoms X such that:

$$X = \{holds(f, k) \mid f \in s_k, 0 \leq k \leq m\} \cup \{holds(neg(f), k) \mid f \notin s_k, 0 \leq k \leq m\} \cup \{holds(occurs(a), k) \mid a \in A_{k+1}, 0 \leq k < m\} \cup \{holds(neg(occurs(a)), k) \mid a \notin A_{k+1}, 0 \leq k < m\}$$

The remainder of X includes the heads of rules that originate from allowance, inhibition, facilitate, contravene, and forbid rules, which correspond to the following atoms:

- $holds(allow(occurs(a)), k)$: allows
- $holds(ab(occurs(a)), k)$: inhibits
- $holds(occurs(a), k)$: facilitates
- $holds(ab(occurs(a)), k)$: contravenes

The fact that X is an answer set can then be demonstrated by induction on the time parameter, leveraging the Splitting Theorem.

Inductive Argument: We will show by induction on time k that X is an answer set of the program \mathcal{P}_{MT} .

The translation in Section 4.3 includes, for every fluent $f \in s_k$, the atom $holds(f, k)$ and for every $\neg f \in s_k$, the atom $holds(neg(f), k)$. As in \mathcal{C}_{TAID} [97], there are five types of rules in the logic program \mathcal{P}_{MT} that generate these atoms in their heads. To include an atom $holds(f, k)$ in an answer set \mathcal{A} , one of these rules must be satisfied:

1. Fluent observations about the initial state:
 $holds(f, 0).$
 $holds(f, 0) : - neg(holds(neg(f), 0)).$
2. Effects of dynamic causal laws:
 $holds(f, k+1) : - holds(occurs(a), k), holds(g-1, k), \dots, holds(g-n, k),$
 $fluent(g-1), \dots, fluent(g-n), fluent(f),$
 $action(a), time(k), time(k+1).$
3. Default values for non-inertial fluents:
 $holds(f, k) : - neg(holds(neg(f), k), fluent(f), time(k)).$
4. Preservation of values for inertial fluents:
 $holds(f, k+1) : - holds(f, k), neg(holds(neg(f), k+1)), neg(default(f)),$
 $fluent(f), time(k), time(k+1).$

5. Effects of static causal laws:

$$\text{holds}(f, k) : - \text{holds}(g_{-1}, k), \dots, \text{holds}(g_{-n}, k), \text{fluent}(g_{-1}), \dots, \\ \text{fluent}(g_{-n}), \text{fluent}(f), \text{time}(k).$$

Moreover, for every mental fluent $f^h \in s_k$, the atom $\text{holds}(f^h, k)$ and for every mental fluent $\neg f^h \in s_k$, the atom $\text{holds}(\text{neg}(f^h), k)$. There are two additional types of rules in the logic program \mathcal{P}_{MT} that generate these atoms in their heads. To include an atom $\text{holds}(f^h, k)$ in an answer set \mathcal{A} , one of these rules must be satisfied:

6. Effects of dynamic causal laws for mental fluents:

$$\text{holds}(f^h, k+1) : - \text{holds}(\text{occurs}(a), k), \text{holds}(g_1, k), \dots, \text{holds}(g_m, k), \\ \text{fluent}(g_1), \dots, \text{fluent}(g_m), \text{mental_fluent}(f^h), \\ \text{action}(a), \text{time}(k).$$

7. Effects of static causal laws for mental fluents:

$$\text{holds}(f^h, k) : - \text{holds}(g_1, k), \dots, \text{holds}(g_m, k), \\ \text{fluent}(g_1), \dots, \text{fluent}(g_m), \text{mental_fluent}(f^h), \text{time}(k).$$

Additionally, the dynamic and static causal laws for mental fluents must adhere to forbidding constraints, ensuring that prohibited mental state transitions do not occur. These are represented as integrity constraints in the logic program \mathcal{P}_{MT} :

8. Forbidding constraints:

$$: - \text{holds}(f_1^h, k+1), \dots, \text{holds}(f_n^h, k+1), \text{holds}(g_1^h, k), \dots, \text{holds}(g_m^h, k), \\ \text{mental_fluent}(f_1^h), \dots, \text{mental_fluent}(f_n^h), \\ \text{mental_fluent}(g_1^h), \dots, \text{mental_fluent}(g_m^h), \text{time}(k).$$

Let us proceed with the inductive argument, starting with a base case, following up with an induction step:

Part1; Step 1; Base Case (k = 0):

For the initial state s_0 , the observations $O_{initial}$ specify the fluents that are true in s_0 . By definition of $O_{initial}$, we have that $f \in s_0$ if and only if $(f \text{ at } 0) \in O_{initial}$, and $f \notin s_0$ if and only if $(\text{neg}(f) \text{ at } 0) \in O_{initial}$. Moreover, by definition of $O_{initial}$, we have that $f(c, v) \in s_0$ if and only if $(f(c, v) \text{ at } 0) \in O_{initial}$, and $f(c, v) \notin s_0$ if and only if $(\text{neg}(f(c, v)) \text{ at } 0) \in O_{initial}$.

In the construction of the set X , this correspondence is captured as follows:

- For every fluent $f \in s_0$, $\text{holds}(f, 0) \in X$.
- For every fluent $f \notin s_0$, $\text{holds}(\text{neg}(f), 0) \in X$.
- For every mental fluent $f^h \in s_0$, $\text{holds}(f^h, 0) \in X$.
- For every mental fluent $f^h \notin s_0$, $\text{holds}(\text{neg}(f^h), 0) \in X$.

This implies that there must be a rule in \mathcal{P}_{MT} whose head corresponds to each such atom. Thus, the base case is satisfied, and X correctly represents the initial state observations in the context of \mathcal{P}_{MT} .

Part 1; Step 2; Induction Step: Assume the induction hypothesis holds true for time k . We need to show it holds for time $k + 1$. Let us proceed by considering Case 1-7 previously enumerated:

- **Case 1:** If $f \in s_k$ due to a fluent observation about the initial state, then $(f \text{ at } 0) \in O_{initial}$. The translation $(f \text{ at } 0)$ is the fact $holds(f, 0)$, which by definition of an answer set has to be included in the answer set X .
- **Case 2:** If $f \in s_k$ is the effect of a dynamic causal law for action a , we know that $a \in A_k$ and $holds(occurs(a), k - 1) \in X$. If a is an exogenous action, $holds(occurs(a), k - 1)$ is a fact. For triggered or allowed actions, $holds(occurs(a), k - 1)$ must be derived from the respective triggering or allowance rules, which are satisfied by induction hypothesis. This implies $holds(f, k) \in X$.
- **Case 3:** For a non-inertial fluent with a default value defined as true, if $f \in s_k$, there is no applicable dynamic or static causal law for $\neg f$. Thus, $holds(neg(f), k) \notin X$, satisfying the body of the default rule and ensuring $holds(f, k) \in X$.
- **Case 4:** If $f \in s_k$ is an inertial fluent such that $f \in s_{k-1}$ and $\neg f \notin s_k$, then $holds(f, k) \in X$ due to the preservation of values for inertial fluents. The body of the inertial rule is satisfied, ensuring $holds(f, k) \in X$.
- **Case 5:** If $f \in s_k$ is the effect of a static causal law, the preconditions g_1, \dots, g_n must hold in s_k . By induction hypothesis, $holds(g_1, k), \dots, holds(g_n, k) \in X$. Thus, the body of the translation of the static causal law is satisfied, ensuring $holds(f, k) \in X$.
- **Case 6:** If $f^h \in s_{k+1}$ due to a dynamic causal law, the preconditions g_1, \dots, g_m must hold at time k , and the action a must occur at time k . This implies $holds(f^h, k + 1) \in X$.

Assume by contradiction that $holds(f^h, k + 1) \in X$ but a forbidding constraint is violated. This means there exists a mental fluent g^h at time k that forbids f^h at time $k + 1$. By the logic program \mathcal{P}_{MT} , the integrity constraint would be violated:

$$: \neg holds(f^h, k + 1), holds(g^h, k).$$

Since this constraint is enforced, $holds(f^h, k + 1)$ cannot be in X when $holds(g^h, k)$ is present. This ensures that mental fluents respect the forbidding constraints and only permitted transitions occur.

- **Case 7:** If $f^h \in s_{k+1}$ due to a static causal law, the preconditions g_1, \dots, g_m must hold at time $k + 1$, implying $holds(f^h, k + 1) \in X$.

Assume by contradiction that $holds(f^h, k + 1) \in X$ but a forbidding constraint is violated. This means there exists a mental fluent g^h at time k that forbids f^h at time $k + 1$. The corresponding integrity constraint in \mathcal{P}_{MT} would be:

$$: \neg holds(f^h, k + 1), holds(g^h, k).$$

Since this constraint is enforced, $holds(f^h, k + 1)$ cannot be in X when $holds(g^h, k)$ is present. This ensures that mental fluents respect the forbidding constraints and that only allowed transitions occur.

The inductive argument ensures that X is an answer set of \mathcal{P}_{MT} at each time step k . Therefore, X correctly represents the trajectory model as specified.

Part 2; Answer Set to Trajectory Model

Assume there is an answer set \mathcal{A} of the program \mathcal{P} . We need to show there is a trajectory model $\langle s_0, A_1, s_1, A_2, \dots, A_m, s_m \rangle$ of $\mathcal{C}_{MT}(D^{MT}, O_{initial} \cup A_Q)$.

Constructing the Trajectory Model: Define the states and action sets as follows:

$$\begin{aligned} s_k &= \{f \mid holds(f, k) \in \mathcal{A}\} \cup \{\neg f \mid holds(neg(f), k) \in \mathcal{A}\} \\ &\quad \cup \{f^h \mid holds(f^h, k) \in \mathcal{A}\} \cup \{\neg f^h \mid holds(neg(f^h), k) \in \mathcal{A}\}, \\ A_{k+1} &= \{a \mid holds(occurs(a), k) \in \mathcal{A}\}. \end{aligned}$$

Part 2: Step 1 - Base Case ($k = 0$):

For the initial state s_0 , we need to show that it correctly represents the initial state observations as specified in $O_{initial}$.

By the definition of \mathcal{A} , which is an answer set of the logic program \mathcal{P}_{MT} , the set s_0 is constructed to include all fluents f for which $holds(f, 0) \in \mathcal{A}$ and to exclude all fluents f for which $holds(\neg f, 0) \in \mathcal{A}$.

This construction ensures that s_0 matches the observations in $O_{initial}$:

- For every fluent f observed to be true in $O_{initial}$, $holds(f, 0) \in \mathcal{A}$, thus $f \in s_0$.
- For every fluent f observed to be false in $O_{initial}$, $holds(\neg f, 0) \in \mathcal{A}$, thus $f \notin s_0$.

Similarly, for mental fluents f^h :

- For every mental fluent f^h observed to be true in $O_{initial}$, $holds(f^h, 0) \in \mathcal{A}$, thus $f^h \in s_0$.
- For every mental fluent f^h observed to be false in $O_{initial}$, $holds(\neg f^h, 0) \in \mathcal{A}$, thus $f^h \notin s_0$.

Therefore, s_0 satisfies the initial state conditions for both regular fluents and mental fluents, ensuring that the initial observations O_{initial} are represented in s_0 .

Furthermore, the trajectory model must adhere to forbidding constraints to ensure valid mental state transitions. If there exists a mental fluent g^h at time k that forbids f^h at time $k + 1$, then:

$$: -\text{holds}(f^h, k + 1), \text{holds}(g^h, k).$$

By the definition of the answer set semantics, this ensures that $f^h \in s_{k+1}$ only if no forbidding mental fluent is present at the preceding step.

Part 2: Step 2 - Induction Step:

Assume the induction hypothesis holds for time k . We need to show it holds for time $k + 1$.

- **Case 1:** Since $k > 0$, $\text{holds}(f, k)$ cannot be caused by an initial fluent observation. Therefore, $f \in s_k$ must hold due to other rules or laws.
- **Case 2:** If $\text{holds}(f, k + 1) \in X$ is the effect of a dynamic causal law for action a , then $a \in A_{k+1}$ and $\text{holds}(\text{occurs}(a), k) \in X$. By the definition of \mathcal{A} , $\text{holds}(\text{occurs}(a), k)$ must be true in s_k , ensuring a occurs at time k . The dynamic causal law a causes f if g_1, \dots, g_n implies that $\text{holds}(g_1, k), \dots, \text{holds}(g_n, k)$ are in X . By induction hypothesis, g_1, \dots, g_n must be in s_k , and thus $f \in s_{k+1}$.
- **Case 3:** If f is a non-inertial fluent with a default value of true, and $\text{holds}(f, k + 1) \in X$ holds due to the application of the default rule, then $\text{holds}(\text{neg}(f), k + 1) \notin X$. This implies there are no applicable static or dynamic causal laws for $\neg f$ or the actions preventing f . Therefore, $f \in s_{k+1}$.
- **Case 4:** If $\text{holds}(f, k) \in X$ and f is an inertial fluent where $\text{holds}(f, k + 1) \in X$ holds due to the application of inertial rules, then $\text{holds}(\text{neg}(f), k + 1) \notin X$. This ensures there are no applicable static or dynamic causal laws for $\neg f$ or the actions preventing f . Thus, $f \in s_{k+1}$.
- **Case 5:** If $\text{holds}(f, k + 1) \in X$ holds due to the application of a static causal law ($f \leftarrow g_1, \dots, g_n$), then $\text{holds}(g_1, k + 1), \dots, \text{holds}(g_n, k + 1) \in X$. By the definition of \mathcal{A} , this implies g_1, \dots, g_n are in s_{k+1} , ensuring $f \in s_{k+1}$.
- **Case 6:** If $\text{holds}(f^h, k + 1) \in \mathcal{A}$, this corresponds to $f^h \in s_{k+1}$ in the trajectory model $\langle s_0, A_1, s_1, \dots, A_m, s_m \rangle$. According to the logic program \mathcal{P}_{MT} , this implies that $\text{holds}(f^h, k + 1) \in X$, where X is the set of atoms in the answer set \mathcal{A} .

Assume by contradiction that $\text{holds}(f^h, k + 1) \in \mathcal{A}$ but a forbidding constraint is violated. This means there exists a mental fluent g^h at time k that forbids f^h at time $k + 1$. By the logic program \mathcal{P}_{MT} , the integrity

constraint would be violated:

$$: \text{-holds}(f^h, k + 1), \text{holds}(g^h, k).$$

Since this constraint holds in \mathcal{P}_{MT} , $\text{holds}(f^h, k + 1)$ cannot be in X when $\text{holds}(g^h, k)$ is present. This ensures that mental fluents respect the forbidding constraints and that only allowed state transitions occur.

- **Case 7:** If $\text{holds}(f^h, k + 1) \in \mathcal{A}$, this corresponds to $f^h \in s_{k+1}$ in the trajectory model $\langle s_0, A_1, s_1, \dots, A_m, s_m \rangle$. According to the logic program \mathcal{P}_{MT} , this implies that $\text{holds}(f^h, k + 1) \in X$, where X is the set of atoms in the answer set \mathcal{A} .

Assume by contradiction that $\text{holds}(f^h, k + 1) \in \mathcal{A}$ but a forbidding constraint is violated. This means there exists a mental fluent g^h at time k that forbids f^h at time $k + 1$. The corresponding integrity constraint in \mathcal{P}_{MT} would be:

$$: \text{-holds}(f^h, k + 1), \text{holds}(g^h, k).$$

Since this constraint is enforced, $\text{holds}(f^h, k + 1)$ cannot be in X when $\text{holds}(g^h, k)$ is present. This ensures that mental fluents respect the forbidding constraints and only permitted transitions occur.

By considering these cases and applying the induction hypothesis, we can conclude that $f \in s_{k+1}$ whenever $\text{holds}(f, k + 1) \in X$. This completes the induction step, establishing that $\langle s_0, A_1, s_1, \dots, A_m, s_m \rangle$ is a valid trajectory model of \mathcal{C}_{MT} .

Therefore, by induction, the trajectory model $\langle s_0, A_1, s_1, \dots, A_m, s_m \rangle$ correctly reflects the answer set \mathcal{A} of the logic program \mathcal{P}_{MT} , ensuring that the constructed trajectory satisfies all conditions specified by the initial observations and the dynamic and static causal laws governing the system. By demonstrating the construction of the answer set from the trajectory model and vice versa, we have proven the theorem.

Let us note that the splitting theorem [173] can be used to divide the logic program \mathcal{P}_{MT} into parts that can be solved separately, corresponding to different time steps in the trajectory. By ensuring that the base case and induction step hold for each split part, the entire program's answer set can be constructed step by step, and proven in a similar manner as in the provided proof.

It has been demonstrated that within a trajectory model $\langle s_0, A_1, s_1, A_2, \dots, A_m, s_m \rangle$, a fluent $f \in s_i$ implies the existence of an atom $\text{holds}(f, i) \in \mathcal{A}$ in the answer set \mathcal{A} of the logic program \mathcal{P}_{MT} . Conversely, if $\text{holds}(f, i) \in \mathcal{A}$, it follows that $f \in s_i$. This correspondence similarly applies to actions: if an action $a \in A_i$, then $\text{holds}(\text{occurs}(a), i - 1) \in \mathcal{A}$; conversely, if $\text{holds}(\text{occurs}(a), i - 1) \in \mathcal{A}$, then $a \in A_i$. The negation holds true for actions not occurring.

For mental fluents represented as f^h , their presence in s_i aligns with $\text{holds}(f^h, i) \in \mathcal{A}$, while their absence aligns with $\text{holds}(\text{neg}(f^h), i) \in \mathcal{A}$. This

adherence reflects the application of dynamic and static causal laws governing mental state transitions. Importantly, the enforcement of forbidding constraints ensures the safety of mental states and transitions. Any violation of these constraints is inherently prevented within \mathcal{A} , maintaining the integrity of the trajectory model.

■

Safety analysis: emotional change

In this subsection, we aim to prove that trajectories generated by \mathcal{C}_{MT} preserve certain safety properties by avoiding unintended emotional side-effects. This is proven by considering principles for emotional change based on HER and UER, respectively, using the *invariance principle* [130]. For each emotional change theory, HER and UER, an *invariance property* is defined, called an Emotional Invariant (*EI*), that holds for all states along a trajectory. An *EI* is preserved by the conditions of a specialized constraint satisfaction problem according to HER-based or UER-based principles.

To enhance the readability of the proofs, we define an emotion labeling scheme.

Definition 32 (Emotion labeling) *For any $\langle s_0, A_1, s_1, A_2, \dots, A_n, s_n \rangle$ of $D_{AE}^{MT}(\mathbf{A}, \mathbf{F})$, there is a transition emotion labeling $\langle E_0, \dots, E_n \rangle$ such that for any state s_i and a set of emotion fluents $\{f_0, \dots, f_m\} \subseteq s_i$, a projection of s_i w.r.t. $\{f_0, \dots, f_m\}$ is denoted by $s_i[f_0, \dots, f_m]$.*

For example, an emotion labeling projection of an initial state $s_0 = \{f(ne, undecided), f(go, high), f(ac, environment), f(co, low)\}$ is denoted $s_0[f(ne, undecided), f(go, high), f(ac, environment), f(co, low)]$. When an emotion labeling projection is associated with an emotion name (such as “hope” according to AE-theory), it can, for enhanced intuition, be denoted $hope : s_0[f(ne, undecided), f(go, high), f(ac, environment), f(co, low)]$.

Using the emotion labeling, we can define the emotional invariants. Let EI_{HER} denote the emotional invariant for HER-based constraints, and EI_{UER} denote the emotional invariant for UER-based constraints, defined as follows:

Definition 33 (Emotional Invariant: Hedonic) *Let $M = \langle s_0, A_1, s_1, A_2, \dots, A_n, s_n \rangle$ be a trajectory of $\mathcal{C}_{MT}(D_{AE}^{MT}, O_{initial} \cup A_Q)$, where $A_i \subseteq \mathbf{A}$ for $0 \leq i < n$.*

The trajectory M satisfies the hedonic emotional invariant, denoted EI_{HER} , if and only if for all transitions s_i to s_{i+1} :

$$s_i[f(ne, v_{ne}), f(go, v_{go}), f(ac, v_{ac}), f(co, v_{co})] \\ \implies s_{i+1}[f(ne, v'_{ne}), f(go, v'_{go}), f(ac, v'_{ac}), f(co, v'_{co})] \wedge v'_{ne} \leq v'_{go}$$

where $v_{ne}, v_{go}, v_{ac}, v_{co}, v'_{ne}, v'_{go}, v'_{ac}, v'_{co}$ are values in their respective domains, and $v_{ne}, v_{go} \in \langle low, undecided, high \rangle$, ranked as $low < undecided < high$.

Definition 34 (Emotional Invariant: Utilitarian) Let $M = \langle s_0, A_1, s_1, A_2, \dots, A_n, s_n \rangle$ be a trajectory of $\mathcal{C}_{MT}(D_{AE}^{MT}, O_{initial} \cup A_Q)$, where $A_i \subseteq \mathbf{A}$ for $0 \leq i < n$.

The trajectory M satisfies the utilitarian emotional invariant, denoted EI_{UER} , if and only if for all transitions s_i to s_{i+1} :

$$\begin{aligned}
& s_i[f(ne, v_{ne}), f(go, v_{go}), f(ac, v_{ac}), f(co, v_{co})] \\
& \implies s_{i+1}[f(ne, v'_{ne}), f(go, v'_{go}), f(ac, v'_{ac}), f(co, v'_{co})] \\
& \quad \wedge v'_{ne} = \text{high} \\
& \quad \wedge v'_{go} \leq v'_{ne} \\
& \quad \wedge (v'_{ac} = \text{self} \vee v'_{ac} = \text{environment}) \\
& \quad \wedge v'_{co} = \text{high}
\end{aligned}$$

where $v_{ne}, v_{go} \in \langle \text{low}, \text{undecided}, \text{high} \rangle$ and $v_{ac} \in \{\text{self}, \text{other}, \text{environment}\}$ and $v_{co} \in \langle \text{low}, \text{undecided}, \text{high} \rangle$.

The intuition behind EI_{HER} is that a positive balance between need consistency and goal consistency should be kept in any transition. Refer to Table 4.1 for HER-based principles of change.

Definition 35 (Utilitarian Emotional Invariant) For any trajectory model $M = \langle s_0, A_1, s_1, A_2, \dots, A_n, s_n \rangle$ where $A_i \subseteq \mathbf{A}$ ($0 \leq i \leq n$) of $\mathcal{C}_{MT}(D_{AE}^{MT}, O_{initial} \cup A_Q)$, all transitions s_i to s_{i+1} preserve a utilitarian emotional invariant, denoted EI_{UER} , such that:

If $s_i \models f(ne, \text{high}) \wedge f(go, V_{go}) \wedge f(ac, V_{ac}) \wedge f(co, V_{co})$, **and either** $V_{go} > \text{high}$ **or** $V_{ac} \notin \{\text{self}, \text{environment}\}$ **or** $V_{co} \neq \text{high}$, **then** $f(ne, \text{high})$ **forbids to cause** $f(go, V_{go})$, $f(ac, V_{ac})$, **or** $f(co, V_{co})$ **respectively in** s_{i+1} .

where $V_{ne}, V_{go} \in \langle \text{low}, \text{undecided}, \text{high} \rangle$ and are ranked as $\text{low} < \text{undecided} < \text{high}$.

The intuition behind EI_{UER} is to maintain a utilitarian state in every transition. According to UER-based principles of change, summarized in Table 4.2, this is achieved by ensuring 1) that goal consistency does not exceed need consistency; 2) the situation must be accountable either to the self or the environment; and 3) the individual must perceive high control potential.

Safety analysis: Hedonic emotional change

In the following theorem, we show that trajectories generated by HER-based constraints preserves the emotional invariant, EI_{HER} , such that if the initial state is a hedonic emotion state, then the following states will be hedonic emotion states.

Theorem 2 (Hedonic emotional change) Let $(D_{AE}^{MT}, O_{initial})$ be an action theory where D_{AE}^{MT} includes the hedonic emotion theory specification (Definition 31), and let $O_{initial}$ be the fluent observations of the initial state. Let Q be a query according to Definition 27, and let $A_Q = \{(a \text{ occurs_at } t_i) \mid a \in A_i, 1 \leq i \leq m\}$.

If there exists a trajectory model $M = \langle s_0, A_1, s_1, A_2, \dots, A_n, s_m \rangle$ of $\mathcal{C}_{MT}(D_{AE}^{MT}, O_{initial} \cup A_Q)$, where $A_i \subseteq \mathbf{A}$ for $0 \leq i \leq m$, then all transitions s_t to s_{t+1} in M preserve the emotional invariant EI_{HER} such that:

if $s_t[f(ne, v_{ne}), f(go, v_{go}), f(ac, v_{ac}), f(co, v_{co})]$,
then $s_{t+1}[f(ne, v'_{ne}), f(go, v'_{go}), f(ac, v'_{ac}), f(co, v'_{co})]$ with $v'_{ne} \leq v'_{go}$,
where $v_{ne}, v_{go}, v'_{ne}, v'_{go} \in \{low, undecided, high\}$ and are ordered as $low < undecided < high$.

Proof: We prove the theorem in two steps: (1) the base case shows that the initial state satisfies EI_{HER} ; (2) the inductive step shows that any transition from s_t to s_{t+1} preserves EI_{HER} .

Step 1: Base Case. Let s_0 be the initial state. From the assumptions of the theorem, $O_{initial} \subseteq s_0$ and $O_{initial}$ contains fluents corresponding to a hedonic emotion state. By Definition 33, EI_{HER} holds in s_0 if and only if for all fluents $f(ne, v_{ne}), f(go, v_{go}) \in s_0$ and corresponding $f(ne, v'_{ne}), f(go, v'_{go}) \in s_1$, we have:

$$s_0[f(ne, v_{ne}), f(go, v_{go}), f(ac, v_{ac}), f(co, v_{co})] \text{ implies} \\
s_1[f(ne, v'_{ne}), f(go, v'_{go}), f(ac, v'_{ac}), f(co, v'_{co})] \wedge v'_{ne} \leq v'_{go}.$$

By assumption, $v_{ne} \leq v_{go}$ holds in s_0 , and no transition has yet occurred. Hence, EI_{HER} holds trivially for s_0 .

Step 2: Inductive Step. Let s_t to s_{t+1} be any transition in the trajectory M . Suppose s_t satisfies the precondition in EI_{HER} :

$$s_t[f(ne, v_{ne}), f(go, v_{go}), f(ac, v_{ac}), f(co, v_{co})]$$

with $v_{ne} \leq v_{go}$. We must show that s_{t+1} satisfies the post-condition in EI_{HER} :

$$s_{t+1}[f(ne, v'_{ne}), f(go, v'_{go}), f(ac, v'_{ac}), f(co, v'_{co})] \wedge v'_{ne} \leq v'_{go}$$

Assume, for contradiction, that there exists a transition s_t to s_{t+1} such that $v'_{ne} > v'_{go}$ in s_{t+1} . Then s_{t+1} includes fluents $f(ne, v'_{ne})$ and $f(go, v'_{go})$ with $v'_{ne} > v'_{go}$.

By the semantics of \mathcal{C}_{MT} , a transition is valid only if $F(s_t) \cap s_{t+1} = \emptyset$, where $F(s_t)$ is the set of fluents forbidden to occur in s_{t+1} by active ‘forbids to cause’ rules in s_t .

Let us now enumerate the relevant rules from Definition 31 that enforce $v'_{ne} \leq v'_{go}$:

- If $f(ne, high) \in s_t$, then the following rules are active:

1. $f(ne, high)$ **forbids to cause** $f(go, low)$
 2. $f(ne, high)$ **forbids to cause** $f(go, undecided)$
 3. $f(ne, high)$ **forbids to cause** $f(go, high)$
- If $f(ne, undecided) \in s_t$, then the following rules are active:
 4. $f(ne, undecided)$ **forbids to cause** $f(go, low)$
 5. $f(ne, undecided)$ **forbids to cause** $f(go, undecided)$
 6. $f(ne, undecided)$ **forbids to cause** $f(go, high)$

Each of these rules defines fluents that are forbidden to hold in s_{t+1} when $f(ne, v_{ne})$ holds in s_t . So if $v'_{ne} = v_{ne}$ and v'_{go} is such that $v'_{ne} > v'_{go}$, then $f(go, v'_{go}) \in F(s_t)$, and the condition $F(s_t) \cap s_{t+1} = \emptyset$ is violated if $f(go, v'_{go}) \in s_{t+1}$. This contradicts the assumption that s_t to s_{t+1} is a valid transition.

Thus, $v'_{ne} > v'_{go}$ cannot hold in any s_{t+1} reachable from s_t that satisfies EI_{HER} . Hence, s_{t+1} must satisfy $v'_{ne} \leq v'_{go}$, and therefore satisfies EI_{HER} . ■

Safety analysis: Utilitarian emotional change

In order to prove invariance for utilitarian emotion regulation, we can follow a similar approach as the proof for hedonic emotion regulation, using the given constraints for UER. Let us define the theorem and present the proof.

Theorem 3 (Utilitarian emotional change) *Let $(D_{AE}^{MT}, O_{initial})$ be an action theory where D_{AE}^{MT} includes the utilitarian emotion theory specification (Definition 31), and let $O_{initial}$ be the fluent observations of the initial state. Let Q be a query according to Definition 27, and let $A_Q = \{(a \text{ occurs_at } t_i) \mid a \in A_i, 1 \leq i \leq m\}$.*

If there exists a trajectory model $M = \langle s_0, A_1, s_1, A_2, \dots, A_n, s_m \rangle$ of $\mathcal{C}_{MT}(D_{AE}^{MT}, O_{initial} \cup A_Q)$, where $A_i \subseteq \mathbf{A}$ for $0 \leq i \leq m$, then all transitions s_t to s_{t+1} in M preserve the emotional invariant EI_{UER} such that:

*if $s_t[f(ne, v_{ne}), f(go, v_{go}), f(ac, v_{ac}), f(co, v_{co})]$,
then $s_{t+1}[f(ne, v'_{ne}), f(go, v'_{go}), f(ac, v'_{ac}), f(co, v'_{co})]$ with:*

- $v'_{ne} = high$,
- $v'_{go} \leq v'_{ne}$,
- $v'_{ac} \in \{self, environment\}$,
- $v'_{co} = high$,

where values range over $v_{ne}, v_{go}, v_{co}, v'_{ne}, v'_{go}, v'_{co} \in \{low, undecided, high\}$ and are ordered as $low < undecided < high$.

Proof: We prove the theorem in two steps: (1) the base case shows that the initial state satisfies EI_{UER} ; (2) the inductive step shows that any transition from s_t to s_{t+1} preserves EI_{UER} .

Step 1: Base Case. Let s_0 be the initial state. From the assumptions of the theorem, $O_{initial} \subseteq s_0$ and $O_{initial}$ contains fluents corresponding to a utilitarian emotion state. By Definition 35, EI_{UER} holds true in s_0 if and only if:

$s_0[f(ne, high), f(go, v_{go}), f(ac, v_{ac}), f(co, high)]$ with $v_{go} \leq high$ and $v_{ac} \in \{self, environment\}$. This is satisfied by initial configurations such as:

Frustration : $s_0[f(ne, high), f(go, low), f(ac, environment), f(co, high)]$

Guilt : $s_0[f(ne, high), f(go, high), f(ac, self), f(co, high)]$

Thus, EI_{UER} holds true in s_0 .

Step 2: Inductive Step. Let s_t to s_{t+1} be any transition in the trajectory M . Suppose s_t satisfies the precondition in EI_{UER} :

$s_t[f(ne, high), f(go, v_{go}), f(ac, v_{ac}), f(co, high)]$

with $v_{go} \leq high$ and $v_{ac} \in \{self, environment\}$. We must show that s_{t+1} satisfies the post-condition in EI_{UER} :

$s_{t+1}[f(ne, high), f(go, v'_{go}), f(ac, v'_{ac}), f(co, high)]$ with $v'_{go} \leq high$, $v'_{ac} \in \{self, environment\}$

Assume, for contradiction, that s_{t+1} violates EI_{UER} . Then at least one of the following holds true:

- (i) $f(ne, v'_{ne}) \in s_{t+1}$ with $v'_{ne} \neq high$
- (ii) $f(go, v'_{go}) \in s_{t+1}$ with $v'_{go} > v'_{ne}$
- (iii) $f(ac, v'_{ac}) \in s_{t+1}$ with $v'_{ac} \notin \{self, environment\}$
- (iv) $f(co, v'_{co}) \in s_{t+1}$ with $v'_{co} \neq high$

By the definition of trajectory in \mathcal{C}_{MT} , a transition is valid only if $F(s_t) \cap s_{t+1} = \emptyset$, where $F(s_t)$ is the set of fluents forbidden in s_{t+1} by active ‘forbids to cause’ rules in s_t .

From Definition 31, the following ‘forbids to cause’ rules are active in s_t and ensure that transitions preserve EI_{UER} :

- Rules (1)–(4): If $f(ne, low)$ or $f(ne, undecided)$ holds true in s_t , then it is forbidden to cause $f(ne, high)$ in s_{t+1} , and vice versa. These rules prevent any value of $v'_{ne} \neq high$ in s_{t+1} , enforcing that $v'_{ne} = high$ must hold true.
- Rules (5)–(10): If $f(go, v_{go})$ holds true in s_t (for any value), then it is forbidden to cause $f(ne, v'_{ne})$ in s_{t+1} where $v'_{ne} \neq high$. This ensures that transitions cannot lead to $v'_{ne} < high$, and thus preserves the condition $v'_{go} \leq v'_{ne}$ required by EI_{UER} .

- Rules (11)–(16): If $f(ac, undecided)$ or $f(ac, other)$ holds true in s_t , then it is forbidden to cause $f(ac, v'_{ac})$ in s_{t+1} for any $v'_{ac} \in \{\text{self, environment}\}$, and vice versa. These rules prevent transitions that would result in $v'_{ac} \notin \{\text{self, environment}\}$.
- Rules (17)–(22): If $f(ac, environment)$, $f(co, low)$, or $f(co, undecided)$ holds true in s_t , then it is forbidden to cause $f(co, high)$ in s_{t+1} (Rules 17, 18, 20, 22), and also forbidden to cause transitions between $f(co, low)$ and $f(co, undecided)$ (Rules 19, 21). These rules prevent transitions that would result in $v'_{co} \neq \text{high}$ in s_{t+1} .

If any of the violating fluents in (i)–(iv) were to appear in s_{t+1} , they would be included in $F(s_t)$, and we would have $F(s_t) \cap s_{t+1} \neq \emptyset$, violating the transition condition. This contradicts the assumption that s_t to s_{t+1} is a valid transition.

Therefore, no such violating transition can occur, and all s_{t+1} reachable from s_t must satisfy EI_{UER} . Hence, the emotional invariant EI_{UER} is preserved across all transitions in the trajectory. ■

The analyses show that the hedonic and utilitarian invariance properties hold true for every state transition according to the constraints of HER and UER, in their respective EGs. Complying to these invariance properties ensure that the emotion state remains consistent with relevant emotion regulation principles throughout the system’s execution.

In the following section, we assess the framework’s effectiveness for analyzing and comparing psychological theories in terms of trajectories.

4.6 Experimental Evaluation

For gaining insights into the system’s behavior, we conducted a series of tests using a synthetic data set ran through the logic program P_{EG} . While the EG allows 108 emotional configurations to be modeled, considering 3 values for need_consistency \times 3 values for goal_consistency \times 3 values for control_potential \times 4 values for accountability, the following analysis is limited to the 16 emotion states defined by AE-theory. This allows us to evaluate a manageable state space. Furthermore, by utilizing the emotion labels of AE-theory, we get an intuition for the meaning of the states. Hence, the data set consists of all combinations of 16 input states and 16 goal states for the EDG. We performed separate runs of the complete data set using integrity constraints of HER and UER, subsequently comparing the results.

Each run is based on plan length 6. Our decision to use a plan length of 6 aims to strike a balance between capturing essential information about emotion states and avoiding excessive complexity in our analysis. Considering the four psychological classes from AE-theory—goal inconsistency, need consistency, accountability, and control potential—we found that a plan length of 6 give room for actions to influence each psychological class to reach the goal state, while

accommodating the constraints of emotional change. Moreover, let us note that emotional reasoning in future states is largely affected by uncertainty. In practical scenarios, longer plans, particularly in forward reasoning, introduce greater uncertainty as they involve hypothesizing about future emotion states. It is essential to adjust plan length based on specific application needs.

The test results comprise a total of 512 runs. This consists of 256 (16×16) runs with HER-based constraints and another 256 runs with UER-based constraints. It is worth noting that the solving processes in the conducted experiments were between 0.009 sec to 0.046 sec, with an average on 0.012 sec, considering both HER and UER based solving processes. Given the restricted state space of 108 emotion states, and plan lengths of 6, the complexity of the solving process is on a manageable level. The full set of test cases, the logic program P_{EG} and the data set can be seen in an online repository¹. To provide an overview, we have chosen specific samples to present. In Table 4.3, we display a sample of 16 runs, each corresponding to a different goal state, all with HER-based constraints. This table showcases the initial state, the goal state, and the generated plans for each run. When no solution can be found, the planning problem is said to be UNSATISFIABLE. Next, in Table 4.4, we compare the results between HER and UER. Similarly, we present a sample of 16 runs but with UER-based constraints. We have intentionally selected the same initial and goal states as in Table 4.3 to emphasize the differences. Lastly, Table 4.5 presents a different sample of 16 runs with UER-based constraints, further illustrating the behavior associated with UER. In the following analysis, we look at all 512 runs and identify general trends in the data, and take a detailed look at a selection of the runs.

Table 4.3: Sample Trajectories: HER.

Label (Init-Goal)	Init	Goal	Plan (Length:6)
Joy-Anger	(ne,h); (go,h); (ac,e); (co,u)	(ne,h); (go,l); (ac,o); (co,h)	UNSATISFIABLE
Fear-Hope	(ne,u); (go,l); (ac,e); (co,l)	(ne,u); (go,h); (ac,e); (co,l)	(co,u,1), (co,u,2), (co,u,3), (co,u,4), (co,l,5), ((go,h),6)
Frustration-Joy	(ne,h); (go,l); (ac,e); (co,h)	(ne,h); (go,h); (ac,e); (co,u)	(co,l,1), (co,l,2), (co,l,3), (co,l,4), (co,u,5), ((go,h),6)
Distress-Relief	(ne,l); (go,l); (ac,e); (co,l)	(ne,u); (go,h); (ac,e); (co,u)	(co,u,1), (ne,l,2), (co,u,3), (ne,l,4), ((go,h),5), (ne,u,6)
Joy-Dislike	(ne,h); (go,h); (ac,e); (co,u)	(ne,u); (go,l); (ac,o); (co,l)	UNSATISFIABLE
Anger-Frustration	(ne,h); (go,l); (ac,o); (co,h)	(ne,h); (go,l); (ac,e); (co,h)	UNSATISFIABLE
Anger-Liking	(ne,h); (go,l); (ac,o); (co,h)	(ne,u); (go,h); (ac,o); (co,u)	(co,l,1), (ne,l,2), (co,l,3), (co,u,4), ((go,h),5), (ne,u,6)
Fear-Regret	(ne,u); (go,l); (ac,e); (co,l)	(ne,u); (go,l); (ac,s); (co,l)	UNSATISFIABLE
Joy-Disgust	(ne,h); (go,h); (ac,e); (co,u)	(ne,l); (go,l); (ac,e); (co,h)	UNSATISFIABLE
Hope-Fear	(ne,u); (go,h); (ac,e); (co,l)	(ne,u); (go,l); (ac,e); (co,l)	UNSATISFIABLE
Hope-Pride	(ne,u); (go,h); (ac,e); (co,l)	(ne,u); (go,h); (ac,s); (co,u)	(ac,s,1), ((go,h),2), ((go,h),3), ((go,h),4), (co,u,5), ((go,h),6)
Sadness-Shame	(ne,h); (go,l); (ac,e); (co,l)	(ne,l); (go,l); (ac,s); (co,h)	UNSATISFIABLE
Regret-Sadness	(ne,u); (go,l); (ac,s); (co,l)	(ne,h); (go,l); (ac,e); (co,l)	UNSATISFIABLE
Hope-Distress	(ne,u); (go,h); (ac,e); (co,l)	(ne,l); (go,l); (ac,e); (co,l)	UNSATISFIABLE
Fear-Surprise	(ne,u); (go,l); (ac,e); (co,l)	(ne,u); (go,u); (ac,e); (co,u)	UNSATISFIABLE
Anger-Guilt	(ne,h); (go,l); (ac,o); (co,h)	(ne,h); (go,h); (ac,s); (co,h)	(co,l,1), (co,l,2), (co,u,3), ((go,h),4), (co,h,5), (ac,s,6)

¹Repository: <https://github.com/AndreasbCS/c-mt>

Table 4.4: Sample Trajectories: UER (Comparison with Table 4.3: HER).

Label (Init-Goal)	Init	Goal	Plan (Length:6)
Joy-Anger	(ne,h); (go,h); (ac,e); (co,u)	(ne,h); (go,l); (ac,o); (co,h)	UNSATISFIABLE
Fear-Hope	(ne,u); (go,l); (ac,e); (co,l)	(ne,u); (go,h); (ac,e); (co,l)	UNSATISFIABLE
Frustration-Joy	(ne,h); (go,l); (ac,e); (co,h)	(ne,h); (go,h); (ac,e); (co,u)	UNSATISFIABLE
Distress-Relief	(ne,l); (go,l); (ac,e); (co,l)	(ne,u); (go,h); (ac,e); (co,u)	UNSATISFIABLE
Joy-Dislike	(ne,h); (go,h); (ac,e); (co,u)	(ne,u); (go,l); (ac,o); (co,l)	UNSATISFIABLE
Anger-Frustration	(ne,h); (go,l); (ac,o); (co,h)	(ne,h); (go,l); (ac,e); (co,h)	(ac,e,1), (co,h,2), (ne,h,3), (co,h,4), (ne,h,5), (co,h,6)
Anger-Liking	(ne,h); (go,l); (ac,o); (co,h)	(ne,u); (go,h); (ac,o); (co,u)	UNSATISFIABLE
Fear-Regret	(ne,u); (go,l); (ac,e); (co,l)	(ne,u); (go,l); (ac,s); (co,l)	(ac,s,1), (ac,s,2), (ac,s,3), (ac,s,4), (ac,s,5), (ac,s,6)
Joy-Disgust	(ne,h); (go,h); (ac,e); (co,u)	(ne,l); (go,l); (ac,e); (co,h)	UNSATISFIABLE
Hope-Fear	(ne,u); (go,h); (ac,e); (co,l)	(ne,u); (go,l); (ac,e); (co,l)	UNSATISFIABLE
Hope-Pride	(ne,u); (go,h); (ac,e); (co,l)	(ne,u); (go,h); (ac,s); (co,u)	UNSATISFIABLE
Sadness-Shame	(ne,h); (go,l); (ac,e); (co,l)	(ne,l); (go,l); (ac,s); (co,h)	UNSATISFIABLE
Regret-Sadness	(ne,u); (go,l); (ac,s); (co,l)	(ne,h); (go,l); (ac,e); (co,l)	UNSATISFIABLE
Hope-Distress	(ne,u); (go,h); (ac,e); (co,l)	(ne,l); (go,l); (ac,e); (co,l)	UNSATISFIABLE
Fear-Surprise	(ne,u); (go,l); (ac,e); (co,l)	(ne,u); (go,u); (ac,e); (co,u)	UNSATISFIABLE
Anger-Guilt	(ne,h); (go,l); (ac,o); (co,h)	(ne,h); (go,h); (ac,s); (co,h)	UNSATISFIABLE

In order to provide a qualitative analysis of the generated trajectories, we established metrics which we refer to as *Emotional Reachability* and *Emotional Priority*, which we further define next.

Definition 36 (Emotional Reachability) Let $EG = \langle S, E \rangle$ be an emotion graph where S is a set of emotion states and $E \subseteq S \times S$ is a set of transition relations between emotion states. Let $D_{AE}^{MT}(\mathbf{A}, \mathbf{F})$ be a domain description. Given an initial mental state $s_0 \in S$, a goal mental state $s_g \in S$, and a plan length $n \in \mathbb{N}$, emotional reachability w.r.t. s_0 and s_g is satisfied if and only if there exists a trajectory $\langle s_0, A_1, s_1, A_2, \dots, A_n, s_n \rangle$, $A_i \subseteq \mathbf{A}$, $s_i \in S$, ($0 \leq i \leq n$), such that $s_n = s_g$.

Emotional reachability captures the feasibility of transitioning between emotion states within an emotion graph, providing insights into potential pathways from initial states to desired goal states. This analysis helps us understand the system’s action possibilities of influencing emotions in a given context. Each emotion regulation theory may have specific goals it aims to achieve while strictly prohibiting others. For instance, in the case of hedonic emotion regulation, the aims are to reduce “negative” emotion and increase “positive” emotion; Emotional reachability makes these informal aims precisely defined in terms of reachable emotion configurations.

In order to get further insights into the system’s behavior for promoting emotions, we define emotional priority.

Definition 37 (Emotional Priority) Let $D_{AE}^{MT}(\mathbf{A}, \mathbf{F})$ be a domain description, C be a set of psychological classes that define the emotion fluents in \mathbf{F} , and $Tr := PEG(D_{AE}^{MT}(\mathbf{A}, \mathbf{F}), n)$ be a set of emotional trajectories. For a psychological class $c \in C$ and an emotional fluent $f_c \in \mathbf{F}$ in the list of trajectories Tr , the emotional priority is determined by

$$P(c, Tr, i) = \frac{|\{f_c \mid T \in Tr, s_i, s_{i-1} \in T, f'_c \in s_i, f''_c \in s_{i-1}, f'_c \notin s_{i-1}\}|}{|Tr|}$$

Table 4.5: Sample Trajectories: UER.

Label (Init-Goal)	Init	Goal	Plan (Length:6)
Dislike-Anger	(ne,u); (go,l); (ac,o); (co,l)	(ne,h); (go,l); (ac,o); (co,h)	(co,h,1), (co,h,2), (co,h,3), (co,h,4), (co,h,5), (ne,h,6)
Dislike-Anger	(ne,u); (go,l); (ac,o); (co,l)	(ne,h); (go,l); (ac,o); (co,h)	(co,h,1), (co,h,2), (co,h,3), (co,h,4), (co,h,5), (ne,h,6)
Shame-Hope Relief-Joy	(ne,l); (go,l); (ac,s); (co,h) (ne,h); (go,h); (ac,e); (co,u)	(ne,u); (go,h); (ac,e); (co,l) (ne,h); (go,h); (ac,e); (co,u)	UNSATISFIABLE (ne,h,1), (ne,h,2), (ne,h,3), (ne,h,4), (ne,h,5), (ne,h,6)
Distress-Relief Joy-Dislike Distress-Frustration	(ne,l); (go,l); (ac,e); (co,l) (ne,h); (go,h); (ac,e); (co,u) (ne,l); (go,l); (ac,e); (co,l)	(ne,u); (go,h); (ac,e); (co,u) (ne,u); (go,l); (ac,o); (co,l) (ne,h); (go,l); (ac,e); (co,h)	UNSATISFIABLE UNSATISFIABLE (ac,s,1), (ac,s,2), (ac,s,3), (co,h,4), (ac,e,5), (ne,h,6)
Regret-Liking Dislike-Regret	(ne,u); (go,l); (ac,s); (co,l) (ne,u); (go,l); (ac,o); (co,l)	(ne,u); (go,h); (ac,o); (co,u) (ne,u); (go,l); (ac,s); (co,l)	UNSATISFIABLE (ac,s,1), (ac,s,2), (ac,s,3), (ac,s,4), (ac,s,5), (ac,s,6)
Distress-Disgust	(ne,l); (go,l); (ac,e); (co,l)	(ne,l); (go,l); (ac,e); (co,h)	(ac,s,1), (ac,s,2), (ac,s,3), (ac,s,4), (co,h,5), (ac,e,6)
Surprise-Fear Liking-Pride	(ne,u); (go,u); (ac,e); (co,u) (ne,u); (go,h); (ac,o); (co,u)	(ne,u); (go,l); (ac,e); (co,l) (ne,u); (go,h); (ac,s); (co,u)	UNSATISFIABLE (ac,s,1), (ac,s,2), (ac,s,3), (ac,s,4), (ac,s,5), (ac,s,6)
Disgust-Shame	(ne,l); (go,l); (ac,e); (co,h)	(ne,l); (go,l); (ac,s); (co,h)	(co,h,1), (ac,s,2), (ac,s,3), (ac,s,4), (ac,s,5), (ac,s,6)
Frustration-Sadness Hope-Distress Fear-Surprise Joy-Guilt	(ne,h); (go,l); (ac,e); (co,h) (ne,u); (go,h); (ac,e); (co,l) (ne,u); (go,l); (ac,e); (co,l) (ne,h); (go,h); (ac,e); (co,u)	(ne,h); (go,l); (ac,e); (co,l) (ne,l); (go,l); (ac,e); (co,l) (ne,u); (go,l); (ac,e); (co,u) (ne,h); (go,h); (ac,s); (co,h)	UNSATISFIABLE UNSATISFIABLE UNSATISFIABLE (ne,h,1), (ne,h,2), (ne,h,3), (ne,h,4), (ac,s,5), (co,h,6)

where $f'_c = f(c, v_1)$ and $f''_c = f(c, v_2)$, with $v_1 \neq v_2$ and $f'_c \neq f''_c$.

Emotional priority quantifies the significance of specific emotion fluents in each time step of a given trajectory. A set of emotional trajectories \mathcal{Tr} , obtained from the logic program P_{EG} applied to $D_{AE}^{MT}(\mathbf{A}, \mathbf{F})$, which represents the sequences of emotion states over time. For each psychological class c in \mathcal{C} , the Emotional Priority $P(c, \mathcal{Tr}, i)$ is computed by comparing the number of instances where specific emotion fluent f_c appears in state s_i but not in s_{i-1} . This count is divided by the total number of emotional trajectories $|\mathcal{Tr}|$, yielding a relative measure of the emotional priority of each fluent for each point in time.

Emotional reachability analysis

Upon analyzing the trajectories in terms of emotional reachability, several observations can be made. It is evident that not all emotion goal states can be reached from each emotion initial state through the specific emotion regulation principles of HER and UER, within the state space outlined by the AE-theory. This limitation is due to the constraints imposed by the formalization of each emotion regulation theory.

Firstly, analyzing the trajectories based on the constraints of HER (see Figure 4.4, marked in blue), we can observe that configurations labeled as Hope, Joy, Relief, Liking, Pride, and Guilt are reachable from all initial configurations. The goal states labeled Dislike, Regret, Fear, Sadness, and Surprise are only

reachable from the same state, allowing these states to remain stable. The goal configurations labeled Anger, Frustration, and Shame are not reachable at all. These observations align with previous empirical findings regarding HER [315], aiming to reduce negative emotions and enhance positive emotions. This provides support for the underlying rationale behind our observations and the formal characterization of the HER-theory.

Secondly, analyzing the trajectories based on the constraints of UER (see Figure 4.4, marked in red), we can observe that the configuration labeled Frustration is reachable from all initial configurations, encompassing each of the 16 AE-based emotions. Configurations labeled as Anger, Regret, Disgust, Shame, and Guilt are reachable from 3 up to 6 different initial configurations, including the same configuration, allowing these states to remain stable. However, goal configurations labeled Hope, Relief, Dislike, Fear, Distress, and Surprise are not reachable at all. The reasoning behind these observations, such as the inclusion of the goal of Frustration and the exclusion of the goal of Joy, bears similarity to prior empirical research on UER [284, 285, 224, 96, 265], indicating that individuals may opt for activities that elicit negative emotions when anticipating a challenging or threatening task, providing rationale for the formal characterization of the UER-theory.

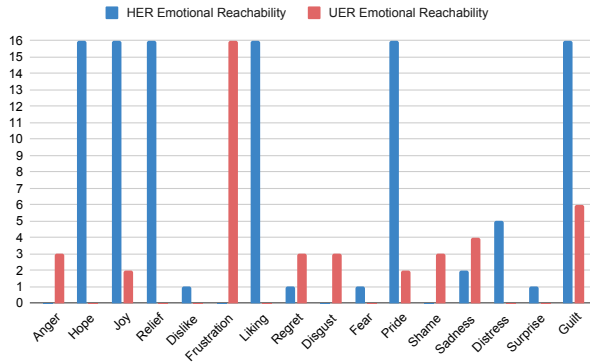


Figure 4.4: Emotional Reachability: Counting generated trajectories (HER and UER based) between each initial configuration and each goal configuration, considering the 16 emotions of AE-theory. The vertical axis represents the number of initial configurations with reachability to a specific goal configuration. The horizontal axis represents goal configurations, which are labeled by emotion to provide intuition.

By examining trajectories that adhere to HER-based constraints (see Figure 4.5) and trajectories that adhere to UER-based constraints (see Figure 4.6), we can gain a more comprehensive understanding of the system’s behavior. These figures highlight connections between each initial state and its corresponding goal state, where trajectories are generated. Any two states that are not con-

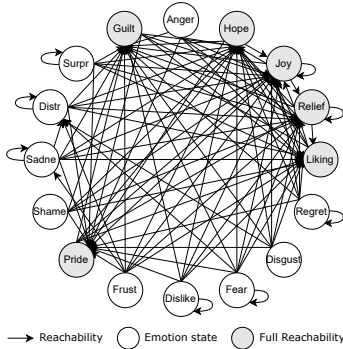


Figure 4.5: Reachability: HER

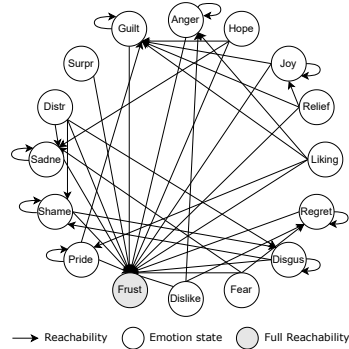


Figure 4.6: Reachability: UER

nected indicates that no valid trajectories were found. Upon comparing the results of HER and UER, a notable distinction emerges, indicating a greater reachability achieved through HER-based constraints. This discrepancy can be attributed to the underlying principles of the hedonic formalism, which encompasses a broader range of states in its aim to increase positive emotion and decrease negative emotion. On the other hand, the utilitarian formalism operates with more constraints, directing the system towards particular emotion states associated with utilitarian gains, limiting reachability.

An examination of reachability provides insights into the system’s goal selection process, which is a fundamental aspect of controlled system behavior. Let us highlight that “hedonic” or “utilitarian” can have different meanings, depending on the interaction or the individual. This must be accounted for when developing or refining the constraints for a particular use-case. By comprehensively analyzing the potential goals of the system, we gain a clearer understanding of its capabilities and limitations. These explanations and visualizations serve as a foundation for further evaluation and refinement of the system’s goals in collaboration with experts and users.

Emotional priority analysis

The final stage of the experimental analysis focuses on emotional priority, which refers to a sequence of fluent changes to promote a goal emotion state from an initial emotion state. Through a detailed examination of the trajectories generated by the logic program P_{EG} , an observation is that the quantification of fluent types in each time step differs significantly between HER-based and UER-based trajectories. Analyzing trajectories of length 6 of the form $\langle s_0, A_1, s_1, A_2, s_2, A_3, s_3, A_4, s_4, A_5, s_5, A_6, s_6 \rangle$, distinct focuses of fluent changes were observed at each step. This analysis was conducted for all 512 runs, calculating the degree of occurrence (in $[0,1]$) of each psychological

class; need_consistency, goal_consistency, control_potential and accountability, at each time step.

In the context of hedonic emotion regulation (HER), an analysis was conducted on all HER-based trajectories, following the 256 test cases to determine the priorities of different influences at each step (see Figure 4.7). Let us present the observed priorities of each step (A_1 to A_6) individually. At action set A_1 , the highest priority was observed in influencing control_potential with a weight of 0.8. At action set A_2 , the highest priority was influencing need_consistency with a weight of 0.5. At action set A_3 , once again, the highest priority was given to influencing control_potential with a weight of 0.6. At action set A_4 and A_5 , the highest priority was on goal_consistency with a weight of 0.6 in A_4 and 0.5 in A_5 . A priority on need_consistency with a weight of 0.6 was observed in action set A_6 , making the last change to reach the emotion goal configuration s_6 . This observed trend in the trajectories can be intuitively explained by the objective of hedonic emotion regulation, which aims to increase positive emotion and decrease negative emotion [286]. Recall that according to our interpretation of the AE-theory, the balance between need_consistency and goal_consistency determines the experience of positive and negative emotions (such that $\text{need_consistency} \leq \text{goal_consistency}$ means positive emotion), while control_potential and accountability regulate the intensity of the emotion by managing the feeling of control [226], and redirecting the focus on who/what is accountable [248] for a situation. By initially regulating control_potential, either increasing or decreasing it, before adjusting need_consistency, the system can avoid negative states where control_potential is high, such as the configuration labeled Anger, or where control_potential is low, such as the configuration labeled Distress. Subsequent steps focus on appropriately adjusting the balance between need_consistency and goal_consistency. Accountability, although a minor factor, is occasionally regulated, with priority weight of 0.2 and 0.3, in steps A_1 and A_5 , respectively. This can be explained by accountability not playing a significant role in the balance between positive and negative emotions [248].

Table 4.6: Emotional Priority: HER

Action Type	A_1	A_2	A_3	A_4	A_5	A_6
influence need_consistency	0.0	0.5	0.1	0.1	0.0	0.6
influence goal_consistency	0.0	0.3	0.3	0.6	0.5	0.3
influence control_potential	0.8	0.2	0.6	0.3	0.2	0.1
influence accountability	0.2	0.0	0.0	0.0	0.3	0.1

In the context of utilitarian emotion regulation, an analysis was conducted on all UER-based trajectories, following the 256 test cases to determine the priorities of different influences at each step (see Figure 4.8). Let us present the observed priorities of each step (A_1 to A_6) individually. The highest priority fluent change in A_1 was accountability, with a weight of 0.6. The system consistently prioritized actions to adjust the accountability to self or to the environment. This influence, in turn, promoted emotional configurations such as Frustration,

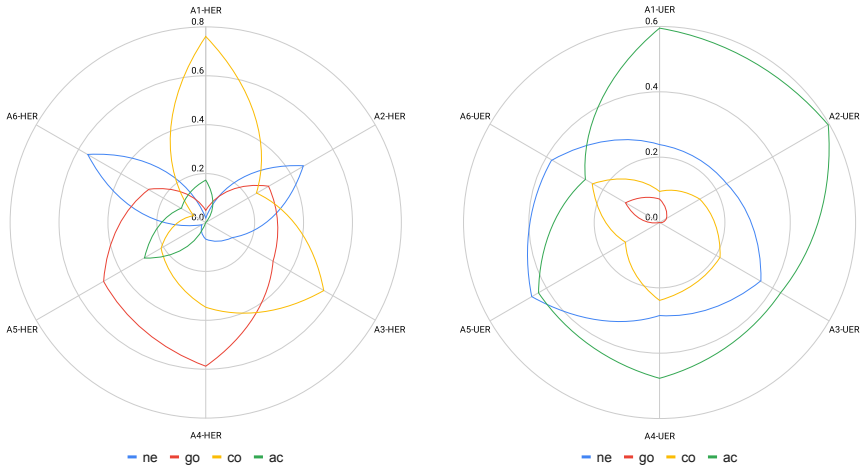


Figure 4.7: Emotional Priority: HER Figure 4.8: Emotional Priority: UER

Guilt or Regret. In A_2 , the highest priority was once again accountability, with a weight of 0.6. At A_3 , the highest priority was need_consistency, most often by increasing it, and accountability, both with weights of 0.4. At A_4 , one again accountability had highest priority, with a weight of 0.5. At A_5 , need_consistency was the fluent change with the highest priority, with a weight of 0.5. Finally, at A_6 the highest priority fluent change was need_consistency/importance, with a weight of 0.4. Goal_consistency/attainability was mostly unaffected. This intuitively reflects that utilitarian gains, such as self responsibility [18], and high need_consistency [96] and high control_potential [286], prevailed over a priority to reach positive emotions (need_consistency \leq goal_consistency), a significant difference from the HER-based trajectories.

Table 4.7: Emotional Priority: UER

Action Type	A_1	A_2	A_3	A_4	A_5	A_6
influence need_consistency	0.2	0.2	0.4	0.3	0.5	0.4
influence goal_consistency	0.1	0.0	0.0	0.0	0.0	0.1
influence control_potential	0.1	0.1	0.2	0.2	0.1	0.2
influence accountability	0.6	0.6	0.4	0.5	0.4	0.3

This analysis has provided insights into the different behaviors emerging from the formalizations of hedonic versus utilitarian emotion regulation approaches (both within the state-space defined by AE-theory). The emotional reachability analysis revealed that not all emotion goal-states can be reached from every initial state. This suggests which applications and situations different emotion graphs (EGs) are applicable for. While HER focuses on balancing the need_consistency/importance and goal_consistency/attainability to achieve positive emotions, UER prioritizes self/environment-accountability and high need_consistency/importance, which intuitively aims to increase utilitarian at-

tributes, such as responsibility and motivation. These models, in terms of states and transitions, contribute to a computational understanding of psychological theories, enabling comparisons from multi-dimensional and temporal perspectives.

4.7 Chapter Remarks

In this chapter, we introduce a computational method to reason about dynamics of mental states through formalizations of psychological theories. We introduce the action language \mathcal{C}_{MT} , and the so-called Belief Graph (BG), which through different specializations, is able to capture multi-dimensional representations of mental states, and principles of mental change, from different psychological theories. We have presented an application in the setting of emotions, using Appraisal theory of Emotion [248], Hedonic Emotion Regulation [315], and Utilitarian Emotion Regulation [284]. Through the introduced methodology, other psychological theories, such as emotion theories based on Arousal and Valence dimensions [157], or theories about motivation, such as the Theory of Planned Behavior [6, 46], can be captured and compared in terms of trajectories.

A BG is a representation, and we expect human mental states and dynamics to be captured there. In that respect, the BG creates a Theory of Mind of the human as a multi-dimensional abstraction based on psychological theories. In the presented emotional reasoning specialization, we do not claim that we represent emotions in terms of one label (e.g., “Sadness” or “Joy”), but make an abstraction in terms of multiple variables (“Need consistency”, “Goal consistency”, “Accountability”, and “Control potential”). While single labels aid in expressing emotion states in a human readable way, they should not inherently contribute to the functionality of a system’s reasoning. In fact, emotional expressions vary across individuals, cultures, languages, and other factors. Therefore, it is crucial for the system to interpret emotion using a multi-dimensional format. Let us further note that the action languages supports knowledge elicitation and engineering. Domain experts can through the language specify high level descriptions of actions, fluents and constraints. Knowledge engineers and system developers can then follow the action specifications to implement mental-state reasoning modules for particular applications.

The proposed framework has notable limitations. In real-world interactions, the impact of a fluent, action, or transition between mental states is rarely fixed. The weight or relevance of these elements can shift depending on the context—even for the same individual. For instance, the emotional or motivational significance of an event may vary based on prior experiences, social dynamics, or momentary goals. Capturing such context-sensitive variation is crucial for developing realistic models of mental-state change. To address this, the framework can be extended with weighted representations of fluents and actions. These weights reflect the relative strength or salience of differ-

ent mental state components in a given situation. This is approached in the following two chapters, where we explore how quantitative bipolar argumentation frameworks (QBAFs) can model weighted beliefs of agents. Moreover, while the current chapter focused on the internal modeling of mental states and their transitions, an important next step is to link this reasoning to interaction protocols—that is, how agents communicate their belief in structured exchanges. To address this, we explore agent interactions through dialogue games, and incorporate QBAFs act as mental-state models, enabling agents to reason with weighted mental constructs in dynamic, rule-governed dialogues, in the following chapter.

Chapter 5

Goal Hiding Dialogues

In this chapter, we examine information-seeking dialogues where a seeker agent strategically delays revealing its true goal—seeking sensitive information—until the respondent is perceived as willing to discuss it. To achieve this, the seeker tactfully introduces intermediary topics that steer the conversation toward its ultimate goal. We propose a formal dialogue game where each dialogue state is associated with a Quantitative Bipolar Argumentation Framework (QBAF)-based willingness model. The seeker maintains a QBAF representation of the respondent’s willingness for different topics, where initial willingness scores are updated dynamically based on how topics promote (support) or demote (attack) each other. This model allows the seeker to reason about when and how to introduce topics to gradually increase the respondent’s willingness to discuss the goal. We define a method for inferring topic relationships from shared information and introduce a gradual semantics to estimate willingness shifts over the course of the dialogue. Our formal analysis and empirical evaluation demonstrate that the proposed dialogue system adheres to privacy-preserving safety properties, ensuring controlled information disclosure. Understanding goal-hiding strategies enables practical applications, such as designing agents that strategically plan dialogues to enhance privacy in human-agent interactions or developing observer agents that detect manipulative goal-hiding tactics in online interactions.

5.1 Introduction

In interpersonal human conversations, participants commonly employ planned behaviors to reach specific goals or outcomes [135]. This includes strategic maneuvers for “setting the stage” [277] to make it easier to approach certain

topics or statements, particularly prominent in interactions concerning sensitive topics that may not be suitable to introduce directly. Speakers continuously observe their listeners for signs of sensitivity, misunderstanding or disagreement and tailor their expressions [28]. The concept of *reference episodes*, wherein speakers utter aspects related to a subsequently introduced entity, such as a topic, statement, or story element, has been posited as a multi-stage process [277]:

- Pre-introduction: Utterances to set the stage for introducing a new entity, without mentioning it directly.
- Formal introduction: The first expression directly referring to the entity.
- Self-repair: Adjusting the characterization of the entity as needed.
- Grounding: Acknowledging or negotiating the entity’s representation with the audience.

This multi-stage strategic approach can more generally be understood as a series of tactful transitions between related topics, gradually adjusting the characterization (e.g., suitability or willingness) of a goal topic that initially is assumed to be difficult, inappropriate or unwanted by the listener. We will refer to these types of dialogues, where a goal is postponed or initially hidden, as Goal-Hiding Dialogues (GHDs). While planned behavior of this kind is present in everyday conversation [135], there are contexts where it is particularly prevalent, such as criminal interrogations [147], medical assessments [114], or health promotion dialogues [140] involving sensitive information that can be difficult for the respondent to talk about or to admit directly. Conversely, in malicious online interactions, such as social media extortion and fraud [83], similar strategies may aim to manipulate individuals into revealing information or agreeing to unfavorable terms.

In this chapter, we seek a formal and computational understanding of GHDs, enabling us to represent and compute the involved strategies for selecting sequences of topics. This includes the development of methods for representing and reasoning about the dynamic characterization of topics, such as the estimated change in a respondent’s willingness for topics, throughout an interaction. A formal understanding of GHDs opens avenues for various applications that deal with strategic maneuvering in (human) interactions. For instance, in human-AI interactions, it can enable conversational systems to proactively avoid discussing matters that may breach user privacy, deferring sensitive topics until the user accepts them. Moreover, in the setting of cybersecurity, a computational understanding of GHDs could enable an observer agent (e.g., third-party software consulted by the user) to overwatch the user’s interactions and detect malicious goal-hiding strategies employed by the user’s interlocutors. The following research questions arise:

RQ5.1: How to recognize arrangements of topics that increases an agent’s willingness to engage in a goal topic?

RQ5.2: What are the formal properties to maintain in such a topic arrangement?

With the aim of approaching these research questions, this chapter is concerned with defining a framework for formal information-seeking dialogues, referred to as *Goal-Hiding Information-Seeking Dialogues*, between a seeker agent and a respondent agent, where it is assumed that the respondent initially is unwilling to disclose the information that the seeker wants. A central concern regards continuously constructing a hypothesis about the interlocutor’s beliefs, such as their current willingness to discuss potential future topics within the ongoing dialogue [37]. This is an ability referred to as Theory of Mind (ToM), which in the setting of formal dialogues often is referred to as Opponent Modeling [129]. In an information-seeking dialogue, the opponent model is the basis for any adaptive changes to the seeker agent’s behavior, e.g., which topics that are opened in a dialogue. Such opponent models can be represented as *gradual* changes between different configurations of mental state properties, and as effects of different events, or an agent’s actions, on the agent’s mind. Therefore, a seeker agent must compute strategies to acquire requisite knowledge for approaching topics without prematurely introducing them. Let us call such a decision process *Epistemic Topic Planning (ETP)*.

ETP requires a formalism to enable representing and reasoning about an agent’s degree of willingness for topics. Furthermore, due to that an agent’s willingness to topics may change depending on how the dialogue develops, we require the formalism to be non-monotonic w.r.t. the state of the dialogue. Moreover, for the purpose of selecting a sequence of topics to reach a goal topic, the formalism must capture relations among topics. 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) other topics.

A non-monotonic formalism which meets our requirements is Quantitative Bipolar Argumentation Frameworks (QBAFs) [29] (refer to Chapter 2.3 for the formal introduction), a non-monotonic formalism that extends classical argumentation theory [94] by incorporating support and attack relations between arguments, deriving weights, so-called strengths, of arguments based on their interactions. In this work, we propose QBAFs for modeling a respondent’s dynamic willingness regarding topics. In this model, arguments represent topics, and the supports and attacks among these topics determine the respondent’s willingness in a given dialogue state. We introduce a game-theoretic gradual semantics to handle changes in willingness as new topics emerge in the dialogue. Additionally, we present a method for analyzing a respondent’s asserted beliefs to construct and incrementally adapt a QBAF-based opponent model throughout the dialogue. By estimating a willingness model for each dialogue state, a seeker agent can strategically promote willingness for a goal topic before introducing it into the conversation. We define the essential formal structures and semantics, which is formally analyzed and implemented as a proof of concept prototype. We further proposes several novel contributions to the state of the

art in the areas of formal argumentation and formal dialogues:

- On the formal theoretical side, we introduce a new formal approach for understanding how a QBAF changes over time, which is an area in need of further exploration in the state of the art.
- A QBAF-based dynamic model is introduced, as a sequence of QBAFs, to track a respondent’s changing willingness for dialogue topics, guiding topic selection.
- A new gradual semantics is introduced that estimates and adjusts the strength of a respondent’s willingness for potential future topics.
- On the application side, contributions regard providing control in the interaction, addressing a key challenge in chatbot development [210], where a system may need to plan its interactions to anticipate and reduce unwanted side effects of its actions, such as avoiding to open topics that may violate privacy or intimacy.
- The formal framework is implemented as a proof of concept prototype in the form of a Web application, including the proposed QBAF-based reasoning engine and an interactive tool, shared as open-source material.

The rest of this chapter is organized as follows. In Section 5.2, we present a brief motivational background. In Section 5.3, a formal framework for goal-hiding dialogues is presented. In Section 5.4, a formal analysis of the proposed framework is conducted. In Section 5.5 and 5.6, a series of running examples are presented. In Section 5.7, an experimental evaluation is presented. In Section 5.8, related work is discussed. In Section 5.9, the chapter is concluded.

5.2 Motivational Background

Verifying safe interactions and ensuring responsible information management have become increasingly critical—especially in cases where agents, whether deliberately or unintentionally, may deviate from these principles. For instance, in social media [144], a malicious human seeker might tactically steer a conversation to gather sensitive information. In Artificial Intelligence (AI)-driven applications [91, 25], controlling the flow and content of information exchanged with, e.g., a chatbot is crucial to ensure that only appropriate topics and related information are discussed. These settings underscore the need for rigorous methods to formally verify and manage how information is presented, requested, or withheld, in digital communication.

Computationally dealing with strategic topic selection in multi-agent environments is a challenging strategic reasoning task [291, 214, 98] involving the systematic process through which an agent utilizes computational models to navigate decision spaces and select (sub)optimal strategies (e.g., sequences of topics to influence a respondent’s willingness for a goal topic). The complexity of these decision spaces has been noted to expand significantly. It has

been observed that real dialogues are games of incomplete knowledge where, e.g., goals, beliefs, and payoffs may be private [124]. Moreover, it has been shown that the problem of identifying the most suitable utterances in a dialogue is NP-complete [123, 184]. By considering these challenges, most of the current dialogue-based frameworks typically consider complete knowledge, and the problem of identifying the most suitable utterances in an ongoing dialogue is not taken into consideration [12, 38, 107, 230]. To efficiently navigate decision spaces in long-term interactions involving multiple exchanges among participants, a system must account for a dynamic user model. This necessitates the ability to represent and reason about the beliefs of the involved agents and understand how actions may influence and alter these beliefs, such as willingness to engage with certain topics. Such capabilities can inform the system’s strategies for selecting appropriate moves in an interaction.

Contemporary interactive software agent solutions, such as chatbots [43] based on different ML [169, 269, 270], NLP [34, 92], and LLM-based solutions [287, 211, 289, 242, 15, 1], avoid tackling the challenge of deliberating over decision spaces, let alone the dynamic changes in mental states. Interactions are often long-term, consisting of sequences of moves which change the overall narrative, for which dialogue systems based on ML/LLM technologies have inherent difficulties in keeping track of [241]. These limitations regard a key challenge in the area of dialogue systems to provide controlled system behaviors [210], where a system may need to constrain and plan its interactions to anticipate and reduce unwanted side effects of its actions. For instance, a system designed to provide information to a human user needs to constrain its responses to mitigate undesirable outcomes, such as the dissemination of harmful information. Similarly, a system designed to seek information from a human user must anticipate and actively minimize the potential exposure of a user’s private information due to improper topics or questions. Dialogue systems based on ML/LLM technologies to solve such tasks would require large amounts of conversation data [310], typically not available in settings where sensitive topics are discussed. Given that most dialogue systems primarily, or in sub-processes, focus on information-seeking procedures, wherein an agent either seeks information or is requested to provide information to another agent, we specifically emphasize this type of dialogue. In order to reason about the dynamics of mental states in such interactions, we require formal dialogue methods able to model interactions between rational agents.

In the following section, we draw inspiration from these prior works and introduce a dialogue framework for managing what we refer to as goal-hiding dialogues. We define the involved vocabulary, agents, moves, and a particular protocol for this type of information-seeking dialogue. Moreover, we define a strategic reasoning process where QBAFs are applied to represent and reason about a respondent’s willingness for topics.

5.3 Formal Framework

The contributions of this work begin in this section. In addressing research question RQ5.1, we establish a formal dialogue framework that connects the theoretical foundations of QBAFs and Formal Dialogue Systems with their application in modeling and tracking a respondent agent’s changing willingness for topics in goal-hiding information-seeking dialogues. A key component is the so-called belief-topic dependency graph, a seeker’s theory about a respondent’s beliefs, enabling the system to infer relations among topics from asserted beliefs. In addressing RQ5.2, we define key properties to maintain in such dialogues, specifically the so-called sensitivity interval and strength monotonicity. Finally, we develop an algorithm for recognizing topic arrangements, referred to as strategies, that aim to gradually increase the willingness for a given goal topic, ultimately enabling its introduction into the dialogue.

By considering QBAFs of the form $Q = \langle X, R^-, R^+, \tau \rangle$, and δ as a strength function, these constructs are applied as follows:

- X represent topics available for introduction in a dialogue. $\mathcal{U}^T \subseteq X$ is called the vocabulary of topics.
- R^- correspond to negative relations between topics, where the intended meaning of $(\beta, \alpha) \in R^-$ is that talking about topic β decreases willingness to talk about topic α .
- R^+ correspond to positive relations between topics, where the intended meaning of $(\beta, \alpha) \in R^+$ is that talking about topic β increases willingness to talk about topic α .
- τ represents the initial willingness to talk about topics, denoted as $\tau(t)$ for each topic $t \in \mathcal{U}^T$.
- δ evaluates the gradual willingness strength of topics, denoted as $\delta(t)$ for each topic $t \in \mathcal{U}^T$, by dynamically adjusting their willingness scores based on the support and attack relations between topics.

Goal-hiding information-seeking dialogues regard interactions as dialogue games between a *seeker agent* (also called seeker) and a *respondent agent* (also called respondent). This is a process of collecting information/beliefs, where the seeker agent opens topics and the respondent replies by asserting beliefs related to the current topic. The seeker aims to find a strategic sequence of topics to, incrementally, strengthen the respondent’s willingness to talk about the goal topic(s) to share the sought-after information (see Figure 5.1).

Two vocabularies are considered to define utterances of an agent a , the vocabulary of topics, \mathcal{U}_a^T , and the vocabulary of beliefs, \mathcal{U}_a^B . For a given agent a , $\mathcal{U}_a := \mathcal{U}_a^T \cup \mathcal{U}_a^B$ denotes all the possible utterances from an agent a , $\mathcal{U}_a^T := \{t_1, \dots, t_p\}$ denotes the known topics by the agent a , and $\mathcal{U}_a^B := \{b_1, \dots, b_q\}$ denotes the known beliefs by the agent a . Sometimes, we just use $\mathcal{U} := \mathcal{U}^T \cup \mathcal{U}^B$ to denote topics and beliefs in general terms. Utterances are delivered to a dialogue through the so-called dialogue moves of an agent:

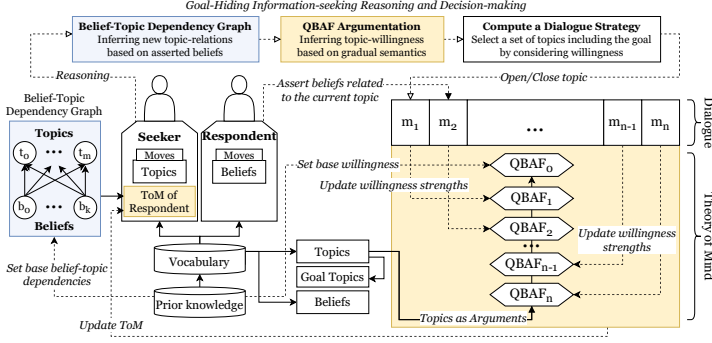


Figure 5.1: Goal-Hiding Dialogue Framework.

Definition 38 (Dialogue moves) *The possible moves of an agent a are:*

- $\langle a, \text{open}, t \rangle$ such that $t \in \mathcal{U}_a^T$, called an open move.
- $\langle a, \text{close}, t \rangle$ such that $t \in \mathcal{U}_a^T$, called a close move.
- $\langle a, \text{assert}, b \rangle$ such that $b \in \mathcal{U}_a^B$, called an assert move.

The set of all possible moves of an agent a is denoted by \mathcal{M}_a .

In this definition, we outline the potential moves available to an agent within a dialogue setting. These moves include “open” moves, marked as $\langle a, \text{open}, t \rangle$, where the agent initiates discussion on a topic t ; “close” moves, denoted as $\langle a, \text{close}, t \rangle$, signaling the conclusion of discussion on topic t ; and “assert” moves, represented by $\langle a, \text{assert}, b \rangle$, where the agent asserts a statement b . In the following, we assume that the seeker only plays open topic and close topic moves, while the respondent only plays assert belief moves, such that the seeker asks for information and the respondent provides information.

A dialogue is a sequence of moves and is defined as follows:

Definition 39 (Dialogue) *Given a finite set of agents \mathcal{I} , a dialogue D_r^n is an ordered sequence of moves $[m_r, \dots, m_n]$, where each move $m_i \in \mathcal{M}_a$ such that $a \in \mathcal{I}$, ($r \leq i \leq n$) and $r, n \in \mathbb{N}$. \mathcal{D} is the set of all possible dialogues, such that:*

Sender(m_i) $\in \mathcal{I}$, in which $\text{Sender} : \mathcal{M} \rightarrow \mathcal{I}$ is a function where \mathcal{M} is the set of all the possible moves of the agents of \mathcal{I} .

A dialogue state denotes the sequence of moves from an initial point r up to point i in a dialogue:

Definition 40 (Dialogue state) *Given a dialogue $D_r^n = [m_r, \dots, m_n]$, $m_i \in \mathcal{M}$, ($r \leq i \leq n$), and $r, n \in \mathbb{N}$, a dialogue state at step i is denoted as: $D_r^i = [m_r, \dots, m_i]$.*

We assume that a seeker agent *se* aims to introduce its goal topics $G \subseteq \mathcal{U}_{se}^T$. To infer strategic sequences that increase willingness for a topic in G , the seeker agent uses a method to relate topics based on how each one promotes or demotes willingness to discuss another. These topic relations are determined by analyzing how beliefs, asserted by the respondent, are connected to topics. For each topic $t \in \mathcal{U}_a^T$, there is a set of beliefs $\mathcal{U}_a^{B^t} \subseteq \mathcal{U}_a^B$ that are related (e.g., contextually) to the topic. Each belief-topic relation is assigned a value in $[-1, 1]$, which intuitively represents the positive, neutral, or negative “attitude,” “tone,” or “sentiment” expressed in the belief towards the topic, providing insights into the respondent’s willingness to discuss the current topic or another topic it influences. A belief can be associated with at most two topics, establishing a one-directional relation from one topic to the other. If a belief b is related to only one topic t , then assigned value of the belief-topic relation provides insight solely into the respondent’s attitude towards that topic. If b is related to two topics, $t1$ and $t2$, the pair of belief-topic relations suggest a directional relationship from $t1$ to $t2$, suggesting how opening topic $t1$ increases or decreases the willingness to discuss $t2$.

These belief-topic relations are captured in a *belief-topic dependency graph* (see Figure 5.2), which encompasses all potential beliefs and topics under consideration that can be asserted in a dialogue. When a belief is asserted, it establishes support or attack relations between the associated topics within a corresponding QBAF. The gradual semantics then uses these relations to infer the strengths of the topics based on their interdependencies.

Let us define the belief-topic dependency graph and a set of notations for working with this structure.

Definition 41 (Belief-Topic Dependency Graph) *Let \mathcal{U}^T be a set of topics and \mathcal{U}^B a set of beliefs. A belief-topic dependency graph is $BT = \langle N, E, v \rangle$, where:*

- $N = \mathcal{U}^T \cup \mathcal{U}^B$ is the set of nodes.
- $E \subseteq \mathcal{U}^B \times \mathcal{U}^T$ is the set of directed edges, with $|\{(b, t) \mid (b, t) \in E\}| \leq 2$ for each $b \in \mathcal{U}^B$.
- $v : E \rightarrow [-1, 1]$ assigns each edge $(b, t) \in E$ a value in $[-1, 1]$.

The role of BT in the proposed framework is to infer relations in QBAF, as follows:

Definition 42 (Belief-Topic Dependency) *Given a belief-topic dependency graph $BT = \langle N, E, v \rangle$, we define the following:*

- $dependent_topics^-(b, \mathcal{U}^T) := \{t \mid t \in \mathcal{U}^T \text{ and } (b, t) \in E \text{ and } v(b, t) < 0\}$.
- $dependent_topics^+(b, \mathcal{U}^T) := \{t \mid t \in \mathcal{U}^T \text{ and } (b, t) \in E \text{ and } v(b, t) > 0\}$.

Definition 43 (Inferring Support and Attack Relations) *Let \mathcal{U}^T be a set of topics, \mathcal{U}^B be a set of beliefs, a be an agent, and D_r^n be a dialogue.*

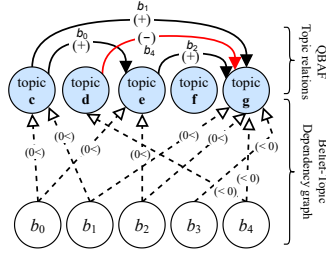


Figure 5.2: In a Belief-Topic Dependency Graph, a belief is either positively dependent (>0) or negatively dependent (<0) to topics, pairs of dependencies imply support or attack relations between topics in an associated QBAF.

- $sup(D_r^n, \mathcal{U}^T) := \{(t, t') \mid t, t' \in \mathcal{U}^T, \{t, t'\} \subseteq dependent_topics^+(b, \mathcal{U}^T) \text{ and } m \text{ appears in } D_r^n \text{ such that } m = \langle a, assert, b \rangle\}$
- $att(D_r^n, \mathcal{U}^T) := \{(t, t') \mid t, t' \in \mathcal{U}^T, \{t, t'\} \subseteq dependent_topics^-(b, \mathcal{U}^T) \text{ and } m \text{ appears in } D_r^n \text{ such that } m = \langle a, assert, b \rangle\}$

By considering Definition 41, if a belief b of an agent a is dependent on more than one topic, then those topics are related through b . We assume that a belief is either positively or negatively dependent on a topic based on whether the quantitative dependency is above or below 0, respectively. If the quantitative dependency equals 0, then there is no dependency. We assume that a knowledge-base exists with, for example, the respondent's prior preferences, to create dependencies between potential beliefs and topics in the belief-topic dependency graph.

For example, consider a QBAF $Q = \langle X, R^-, R^+, \tau \rangle$, where $X = \{x_0, x_1, x_2, x_3, x_4, x_5\}$ is the set of topics. Let the belief-topic dependency graph $BT = \langle N, E, v \rangle$ be specified by:

$$\begin{aligned}
 N &= X \cup \mathcal{U}^B, \text{ where } \mathcal{U}^B = \{b_0, b_1, b_2, b_3, b_4\}, \\
 E &= \{(b_0, x_0), (b_0, x_5), (b_1, x_1), (b_1, x_5), (b_2, x_2), (b_2, x_5), (b_3, x_3), \\
 &\quad (b_3, x_5), (b_4, x_4), (b_4, x_5)\}, \text{ and } v(b_0, x_0) = -1, v(b_0, x_5) = -1, \\
 &\quad v(b_1, x_1) = -1, v(b_1, x_5) = -1, v(b_2, x_2) = -1, v(b_2, x_5) = -1, \\
 &\quad v(b_3, x_3) = 1, v(b_3, x_5) = 1, v(b_4, x_4) = 1, v(b_4, x_5) = 1.
 \end{aligned}$$

By considering the belief-topic dependencies in BT , the following relations can (potentially, if each belief is asserted) be inferred in Q :

$$\begin{aligned}
 R^+ &= \{(x_3, x_5), (x_4, x_5)\} \\
 R^- &= \{(x_0, x_5), (x_1, x_5), (x_2, x_5)\}
 \end{aligned}$$

We situate this example in the context of strength evaluation later on in this section.

Let us proceed by defining the participating agents of an information-seeking dialogue; seeker and respondent.

Definition 44 (A seeker agent) A seeker agent se is a tuple of the form $se := \langle G, \mathcal{U}_{se}, \delta, TS \rangle$ such that:

- $G \subseteq \mathcal{U}_{se}^T$ is a non-empty set of topics denoting goal topics of the agent.
- \mathcal{U}_{se} is a finite set of utterances of the agent.
- $\delta : \mathcal{U}_{se}^T \rightarrow [0, 1]$; we call $\delta(t)$ the theory of a respondent's willingness of a topic $t \in \mathcal{U}_{se}^T$.
- $TS : \mathcal{D} \times G \rightarrow \mathcal{U}_{se}^T$. For any dialogue $D_r^n \in \mathcal{D}$, and a goal topic $g \in G$, $TS(D_r^n)$ returns a topic for the move m_{n+1} .

Definition 45 (A respondent agent) A respondent agent re is a tuple of the form $re := \langle \mathcal{U}_{re}, BS \rangle$ such that:

- \mathcal{U}_{re} is a finite set of utterances of the agent.
- $BS : \mathcal{D} \rightarrow \mathcal{U}_{re}^B$. For any dialogue $D_r^n \in \mathcal{D}$, $BS(D_r^n)$ returns a belief for the move m_{n+1} .

An information-seeking dialogue is a process of collecting beliefs in which the seeker opens topics (e.g., a topic can be a set of questions) and the respondent “replies” by asserting beliefs. A well-formed information-seeking dialogue meets certain protocol conditions, including starting with a topic t opened by the seeker followed by beliefs asserted by the respondent, such that a belief b is connected to an opened topic t , i.e., $v(b, t) \neq 0$. The protocol further states that each topic that is opened must later be closed. When a topic is opened without closing the previous topic, we say that the dialogue is nested.

Definition 46 (An information-seeking dialogue) Let $\mathcal{I} = \{se, re\}$ such that $se = \langle G, \mathcal{U}_{se}, \delta, TS \rangle$ denotes a seeker agent and $re = \langle \mathcal{U}_{re}, BS \rangle$ denotes a respondent agent. An information-seeking dialogue is a tuple of the form $\langle \mathcal{I}, D_r^n \rangle$ in which D_r^n is a dialogue. An information-seeking dialogue is well-formed when the following conditions hold true:

1. D_r^n is a dialogue of the form $[m_r, \dots, m_n]$ such that $Sender(m_i) \in \mathcal{I}$ ($r \leq i \leq n$);
2. m_r is an open move $\langle se, open, t \rangle$ such that $t \in \mathcal{U}_{se}^T$;
3. If m_j ($r < j < n$) is an assert move action of the form $\langle re, assert, b \rangle$, $b \in \mathcal{U}_{re}^B$, then it exists an open move m_i ($r \leq i < j$) of the form $\langle se, open, t \rangle$, $t \in \mathcal{U}_{se}^T$, such that $v(b, t) \neq 0$;
4. If $Sender(m_i) = se$, ($r \leq i < n - 1$), then there exists a possibly empty sequence of moves $\mathcal{R}e = [m_{i+1}, \dots, m_{i+k}]$ ($i + 1 \leq p \leq i + k < n - i$), such that $Sender(m_p) = re$;
5. for all open move $m_i = \langle se, open, t \rangle$ ($r \leq i < n$) there is a close move $m_j = \langle se, close, t \rangle$ ($r < j \leq n$) such that $i < j$ and m_n is a close move;
6. $Sender(m_r) = Sender(m_n)$;
7. $m_i \neq m_j$ for all $r \leq i < j \leq n$;

8. If $\exists t < n$ and $\exists g \in G$ such that g appears in D_r^t , then $n = t + 1$;
9. If $\exists t < n$ and $\forall g \in G$, g does not appear in D_r^t and $\forall m_i \in \mathcal{U}_{se} \cup \mathcal{U}_{re}$ ($r \leq i \leq n$), $m_i \in D_r^n$, then $n = t + 1$;

In the above definition, we can intuitively describe the **protocol** for the dialogue as follows:

1. Each move in the dialogue $[m_r, \dots, m_n]$ is sent by an agent in $\mathcal{I} = \{se, re\}$;
2. The dialogue is started with an open move by se ;
3. All beliefs asserted by re are related to a preceding topic opened by se ;
4. All topics opened by se are followed by 0 or more asserted beliefs by re ;
5. All topics opened are eventually closed by se ;
6. The dialogue is started and ended by se ;
7. All moves in the dialogue are unique;
8. The dialogue terminates successfully if a goal topic by se appears in the dialogue.
9. The dialogue terminates unsuccessfully if no goal topics appear in the dialogue and there are no moves by se nor re that are not already in the dialogue.

Recall that the seeker agent's aim, apart from asking about the sought-after information, is to find a strategic sequence of topics, w.r.t. the respondent's willingness, to promote willingness for the goal topic(s). When the respondent asserts beliefs, inferences are made to find dependencies among asserted beliefs and topics, resulting in relations among topics. Hence, throughout the dialogue, the seeker identifies topic relations to construct links of supporting relations to the goal topic(s).

As seen in Definition 46, an information-seeking dialogue is a composition of topics and asserted beliefs. In turn, the (overarching) information-seeking dialogue is composed of sub-sequences of utterances. Such a sub-sequence is called a sub-dialogue.

Definition 47 (Sub-dialogue) *Let $\gamma = \langle \mathcal{I}, D_r^n \rangle$ be a well-formed information-seeking dialogue. $\gamma' = \langle \mathcal{I}, D_g^h \rangle$ is a sub-dialogue if $\mathcal{I} = \{se, re\}$, D_g^h ($r \leq g \leq h \leq n$) is a sub-sequence of utterances in D_r^n , and γ' is a well-formed information-seeking dialogue.*

An information-seeking dialogue γ , in its minimal form, has at least one sub-dialogue.

Proposition 1 *Let $\gamma = \langle \mathcal{I}, D_r^n \rangle$ be an information-seeking dialogue. If γ is a well-formed information-seeking dialogue, there is at least one sub-dialogue γ' of γ that is well-formed.*

Proof: The proof is direct by the construction of well-formed information-seeking dialogues. ■

In a goal-hiding dialogue, we further assume that each topic is weighted in terms of how willingly beliefs under that topic are disclosed. Let us introduce an information-seeking dialogue system, which is an information-seeking dialogue $\langle \mathcal{I}, D_r^n \rangle$ extended with a QBAF-based *opponent model* (a sequence of QBAFs, Q_r^n) to reason about a respondent's willingness for topics in each dialogue state.

Definition 48 (An information-seeking dialogue system) *Given an information seeking dialogue $\langle \mathcal{I}, D_r^n \rangle$ such that $\mathcal{I} = \{se, re\}$, $se = \langle G, \mathcal{U}_{se}, \delta, TS \rangle$ and $re = \langle \mathcal{U}_{re}, BS \rangle$, an information-seeking dialogue system γ is a tuple of the form $\langle \mathcal{I}, D_r^n, Q_r^n \rangle$ in which*

- Q_r^n is a sequence of QBAFs of the form $[q_r, \dots, q_n]$ such that
 - $q_i = \langle X_i, R_i^-, R_i^+, \tau_i \rangle$, ($r \leq i \leq n$) where
 - $X_i := \mathcal{U}_{se}^T$,
 - $R_i^- := att(D_r^i, \mathcal{U}_{se}^T)$,
 - $R_i^+ := sup(D_r^i, \mathcal{U}_{se}^T)$,
 - $\tau_i(t) := \tau_r(t)$, $t \in \mathcal{U}_{se}^T$.

The dialogue starts with an initial assumption of willingness for each topic $t \in \mathcal{U}^T$, denoted by $\tau(t) \in [0, 1]$. Additionally, it begins with an initial belief-topic dependency graph to reason about relations among topics, and an initial QBAF. The initial QBAF starts with no predefined relations between topics, given that no beliefs yet have been asserted. As the interaction unfolds, and beliefs are asserted, relations among topics emerge. The initial willingness assumptions serve as the base for calculating topic strengths, allowing topic strengths to be traced back to the base scores of the initial state.

Let us define some important concepts for understanding the process of epistemic topic planning. In order for a seeker agent to keep focus on the aim of increasing the willingness for the goal topic(s), the strength of a goal topic should steadily increase. We call this property *Strength monotonicity*, which should be characterized throughout the dialogue.

Definition 49 (Strength Monotonicity) *Let $\gamma = \langle \mathcal{I}, D_r^n, Q_r^n \rangle$ be an information-seeking dialogue system such that $Q_r^n = [q_r, \dots, q_n]$ and $q_i = \langle X_i, R_i^-, R_i^+, \tau_i \rangle$, ($r \leq i \leq n$) and δ be a strength function. δ is monotonic consistent w.r.t. γ and $x \in X_i$ if $\forall i \in \{r, \dots, n-1\}, \exists j \in \{i+1, \dots, n\}$, $t_i \in X_i, t_j \in X_j$, $v_i := \delta(t_i)$ w.r.t. q_i , $v_j := \delta(t_j)$ w.r.t. q_j and $t_i = x$, $t_j = x$ then $v_i \leq v_j$.*

The concept of strength monotonicity relates to the dynamics of information-seeking dialogues. It describes a property essential for maintaining the coherence and progression of such dialogues. The definition specifies that the strength function δ should be monotonic consistent with respect to the dialogue system γ . Intuitively, strength monotonicity means that the strength of

a topic t should either increase or remain constant in between a dialogue state in time i and some future dialogue state in time j . It is important to note that this condition permits local decreases in the topic's strength in the dialogue states between i and j . This maintains a goal-oriented dialogue structure.

In order to respect a respondent's willingness for topics, the seeker requires a measure for when a topic is allowed to be opened. We define a so-called *sensitivity interval* with an upper and lower bound, $\rho^u \in [0, 1]$, $\rho^l \in [0, 1]$, respectively. Any topic t with a willingness strength within a sensitivity interval can be opened. The intuition behind the sensitivity interval is to moderate the extent to which willingness for topics is either promoted or demoted. When a topic's willingness surpasses the upper boundary ρ^u or falls below the lower boundary ρ^l , it means that the respondent either has too low willingness for the topic, or that a too extreme promotion of willingness has been conducted. Surpassing these bounds may explain certain seeker behaviors, and may serve as constraints for effective dialogue management. Specification of upper and lower bounds depends on particular context requirements.

Definition 50 (Sensitivity interval) *Let $\gamma = \langle \mathcal{I}, D_r^n, Q_r^n \rangle$ be an information-seeking dialogue system such that $Q_r^n = [q_r, \dots, q_n]$ and $q_i = \langle X_i, R_i^-, R_i^+, \tau_i \rangle$, ($r \leq i \leq n$), δ is a strength function, $t_j \in X_i$ and the willingness of t_j is $w := \delta(t_j)$ w.r.t. q_i .*

- A sensitivity upper-bound of D_r^i w.r.t. t_j is a constant $\rho^u \in [0, 1]$ such that t_j does not appear in D_r^i and $w < \rho^u$.
- A sensitivity lower-bound of D_r^i w.r.t. t_j is a constant $\rho^l \in [0, 1]$ such that t_j does not appear in D_r^i and $w > \rho^l$.

$[\rho^l, \rho^u]$, such that $\rho^l < \rho^u$, is called the *sensitivity interval* of D_r^i w.r.t. t_j .

Goal-hiding dialogues comply with strength monotonicity (Definition 49) and sensitivity interval (Definition 50) w.r.t. the goal topic(s) to ensure that information-seeking is effective in terms of information extraction while being personalized w.r.t. a respondent agent's willingness for topics.

Definition 51 (A goal-hiding dialogue) *Let $\gamma = \langle \mathcal{I}, D_r^n, Q_r^n \rangle$ be an information-seeking dialogue system such that $\mathcal{I} = \{se, re\}$, $se = \langle G, \mathcal{U}_{se}, \delta, TS \rangle$, $re = \langle \mathcal{U}_{re}, BS \rangle$, $Q_r^n = [q_r, \dots, q_n]$ and $q_i = \langle X_i, R_i^-, R_i^+, \tau_i \rangle$, ($r \leq i \leq n$), $g \in G$ is a goal topic, δ be a strength function, and $[\rho^l, \rho^u]$ is a sensitivity interval w.r.t. g . D_r^n is a goal-hiding dialogue w.r.t. g if the following conditions hold true:*

- $\forall i, \exists h$ with $r \leq i < h \leq n$, such that $v_i := \delta(g)$ w.r.t. q_i , $v_h := \delta(g)$ w.r.t. q_h , and $v_i \leq v_h$.
- $v_i := \delta(g)$ w.r.t. q_i and g does not appear in D_r^i if $v_i > \rho^u$ or $v_i < \rho^l$.
- $v_i := \delta(g)$ w.r.t. q_i and g appears in D_r^i if $\rho^u \geq v_i \geq \rho^l$.

The three conditions of a goal-hiding dialogue define a so-called Goal-Hiding Invariant, which must be preserved in any dialogue state of a goal-hiding dialogue.

Definition 52 (A goal-hiding invariant) Let $\gamma = \langle \mathcal{I}, D_r^n, Q_r^n \rangle$ be an information-seeking dialogue system such that $\mathcal{I} = \{se, re\}$, $se = \langle G, \mathcal{U}_{se}, \delta, TS \rangle$, $re = \langle \mathcal{U}_{re}, BS \rangle$, $Q_r^n = [q_r, \dots, q_n]$ and $q_i = \langle X_i, R_i^-, R_i^+, \tau_i \rangle$, ($r \leq i \leq n$), where $g \in G$ is a goal topic, δ is a strength function, and $[\rho^l, \rho^u]$ is a sensitivity interval w.r.t. g . The Goal-Hiding Invariant (GI) is maintained if the following conditions hold true:

- C1:** $\forall i, \exists h$ with $r \leq i < h \leq n$, such that $v_i := \delta(g)$ w.r.t. q_i , $v_h := \delta(g)$ w.r.t. q_h , and $v_i \leq v_h$.
- C2:** $v_i := \delta(g)$ w.r.t. q_i and g does not appear in D_r^i if $v_i > \rho^u$ or $v_i < \rho^l$.
- C3:** $v_i := \delta(g)$ w.r.t. q_i and g appears in D_r^i if $\rho^u \geq v_i \geq \rho^l$.

For evaluating the strengths of topics in X , we need a gradual semantics to determine gradual strengths, i.e., how the willingness of a topic is weakened or strengthened depending on its relations with other topics.

In these gradual semantics, we consider a notion we call strategy, defined as an ordered sequence of topics. Such strategies belong to a strategy space with the possible combination of moves to reach a goal topic. We define a dialogue strategy space as a set of topics. In each such set, a goal topic of a seeker must be included. Let us assume that the seeker has a set of strategy spaces \mathcal{P} w.r.t g , i.e., $\mathcal{P} \subseteq 2^X$ and $\forall P \in \mathcal{P}, g \in P$. The strategy space P , in turn, suggests a set of (ordered) permutations of the topics, where each such permutation $\langle t_1, \dots, t_k \rangle$, $k = |P|$, so-called strategy, is a possible dialogue sequence leading up to the goal.

Definition 53 (Strategy space) Let $\gamma = \langle \mathcal{I}, D_r^n, Q_r^n \rangle$ be an information-seeking dialogue system, $Q_r^n = [q_r, \dots, q_n]$ and $q_i = \langle X_i, R_i^-, R_i^+, \tau_i \rangle$, ($r \leq i \leq n$), where $x \in X_i$ is a topic, $g \in G$ is a goal topic. $\mathcal{P} \subseteq 2^X$ is the set of all possible strategies such that $\forall P \in \mathcal{P}, g \in P$. Let $\Pi(P)$ denote the set of permutations of a strategy space $P \in \mathcal{P}$. Each permutation $\pi \in \Pi(P)$ is represented as $\pi = \langle t_1, \dots, t_k \rangle$, where $k = |P|$. The notation $\Pi(P)$ represents all such permutations for the strategy space P .

Intuitively, the strategy space refers to all possible sets of topics that the system can choose from during the dialogue. The set of strategy spaces \mathcal{P} is a set containing subsets of topics. Each subset P represents a specific strategy space that the dialogue system can employ. For example, if the set of topics is $X = \{x_1, x_2, x_3\}$, \mathcal{P} might contain subsets like $\{x_1, x_2\}$, $\{x_1, x_3\}$, $\{x_2, x_3\}$, $\{x_1\}$, $\{x_2\}$, $\{x_3\}$, etc. Given a strategy space $P \in \mathcal{P}$, the set of permutations $\Pi(P)$ consists of all possible arrangements of the topics within that strategy space. Each permutation $\pi \in \Pi(P)$ represents a specific ordering or sequence of topics. For instance, if $P = \{x_1, x_2, x_3\}$, then $\Pi(P)$ would contain all permutations of P : $\langle x_1, x_2, x_3 \rangle$, $\langle x_1, x_3, x_2 \rangle$, $\langle x_2, x_1, x_3 \rangle$, $\langle x_2, x_3, x_1 \rangle$, $\langle x_3, x_1, x_2 \rangle$, and $\langle x_3, x_2, x_1 \rangle$.

If a permutation $\pi \in \Pi(P)$ contains the goal topic, then it is considered a candidate strategy. Each such candidate strategy must be further evaluated by considering the gradual semantics. We look into the details of this strategy validation process, and an algorithm to compute it, later on in this section.

The respondent, who only can assert beliefs, is assumed to not be goal-oriented w.r.t. topics. Hence, any topic is included in the respondent's strategy space $O \in \mathcal{O}$, such that $\mathcal{O} \subseteq 2^{\mathcal{U}^{re}}$, where attack and support relations from topics in $O \in \mathcal{O}$ towards topics in $P \in \mathcal{P}$ are inferred from a given belief-topic dependency graph.

Gradual strength evaluation is defined in the following definition.

Definition 54 (Gradual strength evaluation) *Let $\gamma = \langle \mathcal{I}, D_r^n, Q_r^n \rangle$ be an information-seeking dialogue system, such that $\mathcal{I} = \{se, re\}$, $se = \langle G, \mathcal{U}_{se}, \delta, TS \rangle$, $re = \langle \mathcal{U}_{re}, BS \rangle$, $Q_r^n = [q_r, \dots, q_n]$ and $q_i = \langle X_i, R_i^-, R_i^+, \tau_i \rangle$, ($r \leq i \leq n$), where $x \in X_i$ is a topic, $g \in G$ is a goal topic, \mathcal{P} is the set of strategy spaces w.r.t. g of the seeker, and $\mathcal{O} \subseteq 2^{\mathcal{U}^{re}}$ are the strategy spaces of the respondent. $\beta \in [0, 1]$ is a discount constant. The gradual strength evaluation of topic x w.r.t. the strategy spaces $P \in \mathcal{P}$ and $O \in \mathcal{O}$, denoted by $\phi^i(x, P, O)$, is determined by:*

$$\phi^i(x, P, O) = \begin{cases} (\text{Case E1}) : \tau_i(x) - \beta * |P^{\leftarrow O}(R_i^-)| * w(P^{\leftarrow O}(R_i^-)) - \\ |P^{\leftarrow P \cup O}(R_i^+)| * w(P^{\leftarrow P \cup O}(R_i^+)) \\ \text{if } |P^{\leftarrow O}(R_i^-)| * w(P^{\leftarrow O}(R_i^-)) > \\ |P^{\leftarrow P \cup O}(R_i^+)| * w(P^{\leftarrow P \cup O}(R_i^+)); \\ (\text{Case E2}) : \tau_i(x) + \beta * |P^{\leftarrow O}(R_i^-)| * w(P^{\leftarrow O}(R_i^-)) - \\ |P^{\leftarrow P \cup O}(R_i^+)| * w(P^{\leftarrow P \cup O}(R_i^+)) \\ \text{if } |P^{\leftarrow O}(R_i^-)| * w(P^{\leftarrow O}(R_i^-)) < \\ |P^{\leftarrow P \cup O}(R_i^+)| * w(P^{\leftarrow P \cup O}(R_i^+)); \end{cases}$$

where $Y^{\leftarrow X}(R^-) = \{(x, y) | (x, y) \in X \times Y, (x, y) \in R^-\}$ and $Y^{\leftarrow X}(R^+) = \{(x, y) | (x, y) \in X \times Y, (x, y) \in R^+\}$, and $w(X) = \text{agg}(\{\delta(z) \mid z \in X\})$ is a strength aggregation function, where $\text{agg} \in \{\min, \max, \text{average}\}$.

The determination of willingness strength for a specific topic x involves assessing the discrepancy between the aggregated strengths associated with supporting and attacking topics related to x , while also accounting for the cardinality of supporting and attacking topics. This computation incorporates a discount constant denoted as $\beta \in [0, 1]$, called the willingness influence factor. It allows for different aggregation methods for strengths of supporting or attacking topics, such as min, max, or average. Specifically:

- If the aggregated strength of the attacking topics of x , taking into account their cardinality, surpasses the aggregated strength of the supporting topics of x , and their respective cardinality, the willingness strength of x is weakened.

- Conversely, if the aggregated strength of the supporting topics of x , and their cardinality, exceeds the aggregated strength of the attacking topics of x , and their respective cardinality, the willingness strength of x is strengthened.

For example, consider a QBAF $Q_0 = \langle X_0, R_0^-, R_0^+, \tau \rangle$, where $X_0 = \{x_0, x_1, x_2, x_3, x_4, x_5\}$ is the set of topics, $\tau(x_0) = 0.6, \tau(x_1) = 0.6, \tau(x_2) = 0.6, \tau(x_3) = 0.7, \tau(x_4) = 0.7, \tau_i(x_5) = 0.3$ are base scores, and the relations $R_0^- = \{\}$ and $R_0^+ = \{\}$ are initially empty. t_5 is a goal topic.

Let the belief-topic dependency graph $BT = \langle N, E, v \rangle$ be specified by: $N = X_0 \cup \mathcal{U}^B$, where $\mathcal{U}^B = \{b_0, b_1, b_2, b_3, b_4\}$, $E = \{(b_0, x_0), (b_0, x_5), (b_1, x_1), (b_1, x_5), (b_2, x_2), (b_2, x_5), (b_3, x_3), (b_3, x_5), (b_4, x_4), (b_4, x_5)\}$, and $v(b_0, x_0) = -1, v(b_0, x_5) = -1, v(b_1, x_1) = -1, v(b_1, x_5) = -1, v(b_2, x_2) = -1, v(b_2, x_5) = -1, v(b_3, x_3) = 1, v(b_3, x_5) = 1, v(b_4, x_4) = 1, v(b_4, x_5) = 1$.

Considering a strategy space $P = \{x_0, x_1, x_2, x_3, x_4, x_5\}$, then $\Pi(P)$ contains permutations such as $\langle x_0, x_1, x_2, x_3, x_4, x_5 \rangle, \langle x_0, x_2, x_1, x_3, x_4, x_5 \rangle, \langle x_1, x_0, x_3, x_2, x_4, x_5 \rangle$.

In this example, we assume a case where all relations in R_i^+ and R_i^- in time $i > 0$ are inferred in response to beliefs in BT , and we particularly consider the strategy candidate $\pi = \langle x_0, x_1, x_2, x_3, x_4, x_5 \rangle$. Hence, by considering the belief-topic dependencies in BT , the following relations are inferred in time $i > 0$:

$$R_i^+ = \{(x_3, x_5), (x_4, x_5)\}$$

$$R_i^- = \{(x_0, x_5), (x_1, x_5), (x_2, x_5)\}$$

The respondent's strategy space O consists of topics that attack or support the topics in P according to BT . Again, by considering $P = \{x_0, x_1, x_2, x_3, x_4, x_5\}$, then O consists of $\{x_0, x_1, x_2, x_3, x_4\}$ as by R_i^+ and R_i^- . Let us proceed with the gradual strength evaluation of x_5 considering candidate strategy π by using the formula $\phi_i(x_5, P, O)$:

$$\beta * |P^{\leftarrow O}(R_i^-)| * w(P^{\leftarrow O}(R_i^-)) - |P^{\leftarrow P \cup O}(R_i^+)| * w(P^{\leftarrow P \cup O}(R_i^+))$$

Here, $P^{\leftarrow O}(R_i^-) = \{x_0, x_1, x_2\}$, with $|P^{\leftarrow O}(R_i^-)| = 3$ counting the topics attacking a topic in P , and $P^{\leftarrow P \cup O}(R_i^+) = \{x_3, x_4\}$, with $|P^{\leftarrow P \cup O}(R_i^+)| = 2$ counting the topics supporting a topic in P . In this example, we use an aggregation function w based on average, then, by considering the specified base scores, the average strengths are $w(P^{\leftarrow O}(R_i^-)) = 0.6$ and $w(P^{\leftarrow P \cup O}(R_i^+)) = 0.7$. By considering a discount constant $\beta = 0.5$, we calculate the strength change, *sc*:

$$sc := 0.5 * (3 * 0.6 - 2 * 0.7)$$

$$sc := 0.5 * (1.8 - 1.4)$$

$$sc := 0.5 * 0.4$$

$$sc := 0.2$$

We must then check the conditions of **(Case E1)** and **(Case E2)**:

- **(Case E1)**: If attack relations are stronger than support relations, then subtract sc from $\tau_i(x_5)$.
- **(Case E2)**: If support relations are stronger than attack relations, then add sc to $\tau_i(x_5)$.

Since the attacks are stronger than the supports according to:

$$|P^{\leftarrow O}(R_i^-)| * w(P^{\leftarrow O}(R_i^-)) > |P^{\leftarrow P \cup O}(R_i^+)| * w(P^{\leftarrow P \cup O}(R_i^+));$$

$$3 * 0.6 > 2 * 0.7,$$

then **(Case E1)** holds true, and the final strength change score $sc = 0.2$ is subtracted from $\tau_i(x_5) = 0.3$, reflecting a weakened strength of x_5 from 0.3 to 0.1, considering both supporting and attacking topics. We can conclude that this strategy would not increase willingness for the goal topic x_5 . By evaluating all candidate strategies in this way, a seeker can estimate the potential of each available strategy for increasing goal topic strength.

This approach of gradual evaluation draws inspiration from prior game-theoretic gradual semantics [30]. It is worth noting that within specific dialogue contexts, various alternative semantics may be applicable [29].

The strength evaluation of a topic x in the next state, w.r.t. the strategy spaces P and O , can be assumed to be unchanged from the current state if no new relations (support or attack) have emerged from the previous state. This assumption aligns with the principle of inertia in planning, suggesting that if the relationships among topics persist unchanged, the ongoing evaluation of the topic persists. Hence, gradual strength evaluation of a topic x becomes relevant when the situation has changed due to new knowledge acquired through the dialogue.

Definition 55 (Strength evaluation) *Let $\gamma = \langle \mathcal{I}, D_r^n, Q_r^n \rangle$ be an information-seeking dialogue system, such that $\mathcal{I} = \{se, re\}$, $se = \langle G, \mathcal{U}_{se}, \delta, TS \rangle$, $re = \langle \mathcal{U}_{re}, BS \rangle$, $Q_r^n = [q_r, \dots, q_n]$ and $q_i = \langle X_i, R_i^-, R_i^+, \tau_i \rangle$, ($r \leq i \leq n$), where $x \in X_i$ is a topic, $g \in G$ is a goal topic, \mathcal{P} is the set of strategy spaces w.r.t. g of the seeker, and $\mathcal{O} \subseteq 2^{\mathcal{U}_{re}}$ are the strategy spaces of the respondent, $P \in \mathcal{P}$, $O \in \mathcal{O}$. The strength evaluation of topic x in the next state, w.r.t. the strategy spaces P and O , denoted by $\delta^{i+1}(x)$, is determined by:*

$$\delta^{i+1}(x) = \begin{cases} \tau_i(x), & \text{if } i > r, R_i^- = R_{i-1}^- \text{ and } R_i^+ = R_{i-1}^+; (\text{inertia}); \\ \phi^i(x, P, O), & \text{otherwise;} \end{cases}$$

The formal framework for goal-hiding information-seeking dialogues regards an information-seeking strategy, where a seeker opens a series of topics in a dialogue to extract information on particular goal topics while respecting a respondent's willingness for topics.

By considering a strategy space $P \subseteq X_i$ where $g \in P$, where each permutation is evaluated through the gradual strength evaluation, a strategy π is a permutation of topics $\langle t_1, \dots, t_k \rangle$ satisfying a Goal-Hiding Invariant (GI_S or GI_W) at each step of the permutation.

Definition 56 (Goal-hiding strategy) *Let $\gamma = \langle \mathcal{I}, D_r^n, Q_r^n \rangle$ be an information-seeking dialogue system, where $\mathcal{I} = \{se, re\}$, $se = \langle G, \mathcal{U}_{se}, \delta, TS \rangle$, $re = \langle \mathcal{U}_{re}, BS \rangle$, $Q_r^n = [q_r, \dots, q_n]$ and $q_i = \langle X_i, R_i^-, R_i^+, \tau_i \rangle$, ($r \leq i \leq n$), with $g \in G$, δ being a strength function, $[\rho^l, \rho^u]$ being a sensitivity interval w.r.t. g , and GI is a goal-hiding invariant.*

A goal-hiding strategy is a permutation of topics $\pi = \langle t_{n+1}, \dots, t_{n+k}, g \rangle$, where $n+1 \leq j \leq n+k$ and $t_j, g \in \mathcal{U}_{se}^T$ such that GI is satisfied in π w.r.t. g .

We now proceed by establishing a process for recognizing goal-hiding strategies.

Epistemic Topic Planning

We have established the necessary components for what we term as “epistemic topic planning”. The following algorithm presents the control loop of a seeker agent within an information-seeking dialogue system, leveraging epistemic topic planning as a process for goal-hiding strategy recognition (see Algorithm 1).

Initialization: The algorithm begins by initializing an information-seeking dialogue system $\gamma = \langle \mathcal{I}, D_r^n, Q_r^n \rangle$, a belief-topic dependency graph BT , and the bounds of a sensitivity interval $[\rho^l, \rho^u]$. A goal topic g is predefined, directing the epistemic topic planning process.

Main Loop: The algorithm operates within a loop until the seeker agent either accomplishes its initial goal or the dialogue is otherwise terminated. In each iteration, a series of procedures are conducted:

1. **Belief revision and dialogue state initialization:** A respondent belief $b \in \mathcal{U}_{re}^B$ is collected from the dialogue, representing an assertion of the belief b in the move $m_{n+1} = \langle re, assert, b \rangle$. The move m_{n+1} is appended to the current dialogue, initiating a new dialogue state D_r^{n+1} . Concurrently, it constructs a QBAF Q_r^{n+1} based on the current dialogue state, incorporating supports and attacks according to topic relations of b in BT .
2. **Strategy candidate generation:** Once the dialogue state is updated, the algorithm generates candidate strategies Π by permuting topics from the historical dialogue D_r^n up to the current goal g . It then filters these candidate strategies, checking the existence of support relations to g , suggesting a possible strategy to reach g .
3. **Goal-hiding strategy selection:** For each candidate strategy $\pi \in \Pi$, the algorithm performs forward-checking to evaluate the strength of each topic transition with respect to the goal g . This is accomplished by constructing a counterfactual QBAF instance (hypothetical willingness

state) for each move in the strategy up to g . It checks whether each topic transition maintains the topic's strength within the sensitivity interval. Each strategy that complies with the sensitivity interval for each move up to g is further verified by considering the strength monotonicity of g , comparing the strength of g in the initial dialogue state D_7^n and in the final state. Strategies that meet these criteria are retained.

4. **Execution:** Finally, the algorithm executes the selected strategy by instructing the seeker agent to open the next corresponding topic in the selected strategy, thereby advancing the dialogue towards achieving its objectives while strategically postponing the goal topic.

The algorithm provides a systematic approach for goal-hiding strategy identification to adaptively select the most promising course of action based on the dialogue context and the underlying belief-topic dependencies. If a goal-hiding strategy exists, the algorithm will find it. Let us provide soundness and completeness proofs in the following section.

Algorithm 1: Epistemic Topic Planning (ETP)

Input: $\gamma = \langle \mathcal{I}, D_r^n, Q_r^n \rangle$, $\mathcal{I} = \{se, re\}$, $se = \langle G, \mathcal{U}_{se}, \delta, TS \rangle$, $re = \langle \mathcal{U}_{re}, BS \rangle$ #
 Information-seeking dialogue system;
Input: BT # Belief-topic dependency graph;
Input: $[\rho^l, \rho^u]$ # Sensitivity interval where $\rho^l, \rho^u \in [0, 1]$;
 1 $g \leftarrow g_0 \in G$ # Initial goal topic;
 2 $t \leftarrow null$;
 3 **while not** $(\langle se, open, t \rangle \in D_r^n$ or $\langle se, open, g \rangle \in D_r^n)$ **do**
 4 $b \leftarrow getBeliefObservation(\mathcal{U}_{re}^B)$ # such that $b \in \mathcal{U}_{re}^B$;
 5 $m_{n+1} \leftarrow \langle re, assert, b \rangle$ # Get move with asserted belief b ;
 6 $D_r^{n+1} \leftarrow D_r^n \cup \{m_{n+1}\}$ # Append new move to the dialogue;
 7 $Q_r^{n+1} \leftarrow createQBAF(Q_r^n, b, BT)$ # Append supports and attacks to Q_r^{n+1}
 according to dependencies in BT w.r.t. belief $b \in m_{n+1}$;
 8 $\Pi \leftarrow findCandidateStrategies(D_r^{n+1}, g, BT)$ # Generate permutations of topics
 and match with support relations to g ;
 9 $\pi \leftarrow selectGoalHidingStrategy(\Pi, g, Q_r^{n+1}, BT)$ # Reduce strategies by verifying
 Strength monotonicity and Sensitivity interval;
 10 **if** π **is not null** **then**
 11 $t \leftarrow \pi[n+1]$;
 12 **execute** $(\langle se, move, t \rangle)$: $move \in \{open, close\}$, $t \in \mathcal{U}_{se}^T$;
 13 **Function** $findCandidateStrategies(D_r^n, g, BT)$
 14 $\Pi \leftarrow generatePermutations(D_r^n, g)$ # Permutations of the form
 $\langle t_r, \dots, t_n, t_{n+1}, \dots, t_k, g \rangle$, $(r \leq j \leq k)$, $t_j, g \in \mathcal{U}_{se}^T$, where $\{t_r, \dots, t_n\}$ and g are
 fixed topics, and $\{t_{n+1}, \dots, t_k\}$ represents new topics arranged in different orders
 (According to Definition 53);
 15 $candidateStrategies \leftarrow \emptyset$;
 16 **foreach** π **in** Π **do**
 17 $hasSupportToGoalTopic \leftarrow \text{False}$;
 18 **foreach** $t \in \pi$ **do**
 19 **if** $\exists b \in \mathcal{U}^B$: $t \in dependent_topics^+(b, \mathcal{U}^T)$ **and** $g \in$
 20 $dependent_topics^+(b, \mathcal{U}^T)$ **then**
 21 $hasSupportToGoalTopic \leftarrow \text{True}$;
 break;
 22 **if** $hasSupportToGoalTopic$ **then**
 23 $candidateStrategies \leftarrow candidateStrategies \cup \{\pi\}$;
 24 **return** $candidateStrategies$;
 25 **Function** $selectGoalHidingStrategy(candidateStrategies, g, Q_r^n, BT)$
 26 $selectedStrategy \leftarrow null$;
 27 $goalHidingStrategies \leftarrow null$;
 28 **foreach** π **in** $candidateStrategies$ **do**
 29 $cQ_r^n \leftarrow Q_r^n$ # Create a counterfactual QBAF instance;
 30 **for** $i \leftarrow n+1$ **to** $length(\pi) - 1$ **do**
 31 $t_i \leftarrow \pi[i]$;
 32 $cQ_r^i \leftarrow createQBAF(cQ_r^{i-1}, t_i, BT)$ # Append supports and attacks to
 Q_r^i according to dependencies in BT w.r.t. topic t_i ;
 33 # Check sensitivity interval w.r.t. topic t_i (Strength evaluation according
 to Definitions 54-55);
 34 **if not** $\rho^l \leq I_\delta(cQ_r^i, t_i) \leq \rho^u$ **then**
 35 **break**;
 36 # Check strength monotonicity w.r.t. the goal topic g (Strength evaluation
 according to Definitions 54-55);
 37 **if** $I_\delta(cQ_r^i, g) \geq I_\delta(Q_r^i, g)$ # Strength monotonicity;
 38 **then**
 39 $goalHidingStrategies \leftarrow goalHidingStrategies \cup \{\pi\}$;
 40 $selectedStrategy \leftarrow selectStrategy(goalHidingStrategies)$ # Based on an
 appropriate optimization criteria;
 41 **return** $selectedStrategy$;

5.4 Formal analysis

In this section, we conduct a formal analysis of the proposed dialogue framework. We analyze the properties of goal-hiding dialogues to understand their behavior. We then analyze the introduced gradual semantics, considering a set of 11 basic ideas of gradual semantics in QBAFs introduced by [29].

Analysis of Goal-hiding Dialogues

Let us begin by showing that *goal-hiding dialogues* generated by an information-seeking dialogue system preserve a safety property, such that any topic, when being opened, complies with a sensitivity interval. Meanwhile, we show that the dialogue steadily progresses towards a goal topic by complying with the property of strength monotonicity. First, we show the existence of goal-hiding strategies. Then, we show that the conditions of goal-hiding dialogues are preserved in any dialogue state.

Theorem 4 (Goal-hiding strategy) *Let $\gamma = \langle \mathcal{I}, D_r^n, Q_r^n \rangle$ be an information-seeking dialogue system, such that $\mathcal{I} = \{se, re\}$, $se = \langle G, \mathcal{U}_{se}, \delta, TS \rangle$, $re = \langle \mathcal{U}_{re}, BS \rangle$, $Q_r^n = [q_r, \dots, q_n]$ and $q_i = \langle X_i, R_i^-, R_i^+, \tau_i \rangle$, ($r \leq i \leq n$), where $x \in X_i$ is a topic, $g \in G$ is a goal topic, δ is a strength function, $[\rho^l, \rho^u]$ is a sensitivity interval w.r.t. g , \mathcal{P} is the set of strategies w.r.t. g of the seeker, and $\mathcal{O} \subseteq \mathcal{2}^{\mathcal{U}_{re}}$ are the strategies of the respondent, $O \in \mathcal{O}$, and $w(X) = \text{agg}(\{\delta(z) \mid z \in X\})$ is a strength aggregation function, where $\text{agg} \in \{\min, \max, \text{average}\}$. If D_r^n is a goal-hiding dialogue w.r.t. g , and g appears in D_r^n , then there exists $\pi \in \Pi(P)$ and $P \subseteq \mathcal{P}$ is a strategic space w.r.t. g such that $|P^{\leftarrow O}(R_i^-)| * w(P^{\leftarrow O}(R_i^-)) < |P^{\leftarrow P \cup O}(R_i^+)| * w(P^{\leftarrow P \cup O}(R_i^+))$ holds true.*

Proof: Let us start by assuming that there does not exist $\pi \in \Pi(P)$ and $P \subseteq \mathcal{P}$ w.r.t. g such that $|P^{\leftarrow O}(R_i^-)| * w(P^{\leftarrow O}(R_i^-)) < |P^{\leftarrow P \cup O}(R_i^+)| * w(P^{\leftarrow P \cup O}(R_i^+))$ ($r \leq i \leq n$).

Let us look for a contradiction by observing each condition of a goal-hiding dialogue. The conditions are:

- C1:** $v_i := \delta(g)$ w.r.t. q_i , $v_h := \delta(g)$ w.r.t. q_h ($r \leq i, h \leq n$), $i < h$ and $v_i < v_h$.
- C2:** $v_i := \delta(g)$ w.r.t. q_i ($r \leq i \leq n$) and g does not appear in D_r^i if $v_i > \rho^u$ or $v_i < \rho^l$.
- C3:** $v_i := \delta(g)$ w.r.t. q_i ($r \leq i \leq n$) and g appears in D_r^i if $\rho^u \geq v_i \geq \rho^l$.

If D_r^n is a goal-hiding dialogue w.r.t. g , then, there is a $\pi = \langle t_1, \dots, t_k \rangle \in \Pi(P)$ and $P \in \mathcal{P}$ such that g appears in P and all the topic D_r^n are the same to the topics of π . Since D_r^n is a goal-hiding dialogue w.r.t. g , then π holds true C1 and C2. Let us observe that if C1 holds true then $|P^{\leftarrow O}(R_i^-)| * w(P^{\leftarrow O}(R_i^-)) <$

$|P^{\leftarrow P \cup O}(R_i^+)| * w(P^{\leftarrow P \cup O}(R_i^+)) (r \leq i \leq n)$ holds true. Then we have a contradiction w.r.t π .

■

We have to show that dialogues generated by a goal-hiding information-seeking dialogue system preserve a *safety property* that complies to the conditions of a goal-hiding dialogue, in each state of the dialogue.

Theorem 5 (Goal-hiding invariance) *Let $\gamma = \langle \mathcal{I}, D_r^n, Q_r^n \rangle$ be an information-seeking dialogue system, such that $\mathcal{I} = \{se, re\}$, $se = \langle G, \mathcal{U}_{se}, \delta, TS \rangle$, $re = \langle \mathcal{U}_{re}, BS \rangle$, $Q_r^n = [q_r, \dots, q_n]$ and $q_i = \langle X_i, R_i^-, R_i^+, \tau_i \rangle$, ($r \leq i \leq n$), where $x \in X_i$ is a topic, $g \in G$ is a goal topic, δ is a strength function, $[\rho^l, \rho^u]$ is a sensitivity interval w.r.t. g , \mathcal{P} is the set of strategies w.r.t. g of the seeker, and $\mathcal{O} \subseteq 2^{\mathcal{U}_{re}}$ are the strategies of the respondent, $O \in \mathcal{O}$, and $w(X) = \text{agg}(\{\delta(z) \mid z \in X\})$ is a strength aggregation function, where $\text{agg} \in \{\min, \max, \text{average}\}$.*

*If D_r^n is a goal-hiding dialogue w.r.t. g and there exists $\pi \in \Pi(P)$ and $P \subseteq \mathcal{P}$ w.r.t. g such that $|P^{\leftarrow O}(R_i^-)| * w(P^{\leftarrow O}(R_i^-)) < |P^{\leftarrow P \cup O}(R_i^+)| * w(P^{\leftarrow P \cup O}(R_i^+))$, then any sub-dialogue $\gamma_i \subseteq D_r^n$ at time point i preserve the Goal-Hiding Invariant (GI) w.r.t. g defined by:*

C1: $v_i := \delta(g)$ w.r.t. q_i , $v_h := \delta(g)$ w.r.t. q_h ($r \leq i, h \leq n$), $i < h$ and $v_i < v_h$.

C2: $v_i := \delta(g)$ w.r.t. q_i ($r \leq i \leq n$) and g does not appear in D_r^i if $v_i > \rho^u$ or $v_i < \rho^l$.

C3: $v_i := \delta(g)$ w.r.t. q_i ($r \leq i \leq n$) and g appears in D_r^i if $\rho^u \geq v_i \geq \rho^l$.

Proof: Assuming that a given dialogue state D_r^i holds according to GI, and g does not appear in D_r^i . Then, let us consider two cases for dialogue state D_r^{i+1} :

- (A) If a strategy $\pi \in \Pi(P)$ is executed and ($\delta(g)$ w.r.t. $q_{i+1} < \rho^l$ or $\delta(g)$ w.r.t. $q_{i+1} > \rho^u$), then g does not appear in D_r^{i+1} .
- (B) If strategy $\pi \in \Pi(P)$ is executed and $\rho^u \geq \delta(g)$ w.r.t. $q_{i+1} \geq \rho^l$, then g appears in D_r^{i+1} .

It is clear that condition C1 holds, since an execution of a strategy $\pi \in \Pi(P)$ will result in $\delta(g)$ w.r.t. $q_i < \delta(g)$ w.r.t. q_{i+1} . Also, condition C2 and C3 hold, since in case A, g does not appear in D_r^{i+1} , and in case B, g appears in D_r^{i+1} . Hence, we can conclude that γ preserves the conditions of GI.

■

Corollary 1 (Willingness monotonic strategy) *Let $\gamma = \langle \mathcal{I}, D_r^n, Q_r^n \rangle$ be an information-seeking dialogue system, such that $\mathcal{I} = \{se, re\}$, $se = \langle G, \mathcal{U}_{se}, \delta, TS \rangle$, $g \in G$ is a goal topic, δ is a strength function, \mathcal{P} is the set of strategy spaces w.r.t. g of the seeker. If D_r^n is a goal-hiding dialogue and g appears in D_r^n , then δ is monotonic consistent w.r.t. γ and g .*

We have looked at the process of goal-hiding dialogues and concluded that an information-seeking dialogue system γ preserves the conditions of the goal-hiding invariant GI , ensuring that information-seeking is goal-oriented in terms of information extraction, w.r.t. a goal topic g , while respecting a respondent's willingness for g .

By considering these conclusions, let us proceed by analyzing the soundness and completeness of the ETP Algorithm (Algorithm 1).

Theorem 6 (ETP - Soundness) *Let $ETP(\gamma, BT, \mathcal{E}, [\rho^l, \rho^u])$ be an epistemic topic planner, where $\gamma = \langle \mathcal{I}, D_r^n, Q_r^n \rangle$ is an information-seeking dialogue system, $\mathcal{I} = \{se, re\}$, $se = \langle G, \mathcal{U}_{se}, \delta, TS \rangle$, $re = \langle \mathcal{U}_{re}, BS \rangle$, $Q_r^n = [q_r, \dots, q_n]$ and $q_i = \langle X_i, R_i^-, R_i^+, \tau_i \rangle$, ($r \leq i \leq n$), $g \in G$ is a goal topic, δ is a strength function, BT is a belief-topic dependency graph, \mathcal{E} is an environment sensor, $[\rho^l, \rho^u]$ is a sensitivity interval with respect to $\forall t \in X_i$. \mathcal{P} is the set of strategy spaces with respect to g of the seeker.*

If $\pi := ETP(\gamma, BT, \mathcal{E}, [\rho^l, \rho^u])$, where $\pi \in \Pi(P)$ and $P \in \mathcal{P}$, then π satisfies the Goal-Hiding Invariant (GI).

Proof: Let π be a strategy of the form $\langle t_r, \dots, t_n, t_{n+1}, \dots, t_k \rangle$, ($r \leq j \leq k$), $t_j \in \mathcal{U}_{se}^T$, $g = t_k$, inferred by ETP . This means that π belongs to the set of candidate strategies Π identified by Algorithm 1, where Π are all the permutations of the set of topics in the strategy space $P = \{t_r, \dots, t_k\} \in \mathcal{P}$ (according to Definition 53).

ETP , through forward-checking, iteratively constructs QBAFs for each topic in each strategy: $\forall \pi' \in \Pi$, $\forall t_p \in \pi'$, ($n+1 \leq p \leq k$), $Q_r^p = Q_r^{p-1} \cup \{q_p\}$, $q_p = \langle X_p, R_p^-, R_p^+, \tau_p \rangle$, for which the following conditions are checked:

1. evaluate the sensitivity interval: $\rho^l \leq I_\delta(Q_r^p, t_p) \leq \rho^u$ (according to GI condition C2 and C3 in Theorem 5).
2. evaluate strength monotonicity: $I_\delta(Q_r^n, g) \leq I_\delta(Q_r^p, g)$ (according to GI condition C1 in Theorem 5).

Any strategy $\pi' \in \Pi$ violating the conditions of GI is discarded. Therefore, if $\pi := ETP(\gamma, BT, \mathcal{E}, [\rho^l, \rho^u])$, then π satisfies the Goal-Hiding Invariant (GI).

■

Theorem 7 (ETP - Completeness) *Let $ETP(\gamma, BT, \mathcal{E}, [\rho^l, \rho^u])$ be an epistemic topic planner, where $\gamma = \langle \mathcal{I}, D_r^n, Q_r^n \rangle$ is an information-seeking dialogue system, $\mathcal{I} = \{se, re\}$, $se = \langle G, \mathcal{U}_{se}, \delta, TS \rangle$, $re = \langle \mathcal{U}_{re}, BS \rangle$, $Q_r^n = [q_r, \dots, q_n]$ and $q_i = \langle X_i, R_i^-, R_i^+, \tau_i \rangle$, ($r \leq i \leq n$), $g \in G$ is a goal topic, δ is a strength function, BT is a belief-topic dependency graph, \mathcal{E} is an environment sensor, $[\rho^l, \rho^u]$ is a sensitivity interval w.r.t. $\forall t \in X_i$. \mathcal{P} is the set of strategy spaces w.r.t. g of the seeker.*

If there exists a strategy $\pi \in \Pi(P)$ and $P \in \mathcal{P}$ such that satisfies the Goal-Hiding Invariant GI (as per Theorem 5), then $\pi' := ETP(\gamma, BT, \mathcal{E}, [\rho^l, \rho^u])$ and $\pi' = \pi$.

Proof: The strategy space $P = \{t_r, \dots, t_k\}$ defines a finite set of permutations of the form $\langle t_r, \dots, t_n, t_{n+1}, \dots, t_k \rangle$, where $(r \leq j \leq k)$, $t_j \in \mathcal{U}_{se}^T$, with $g = t_k$. In these permutations, $\{t_r, \dots, t_n\} \cup t_k$ are topics fixed in position, while $\{t_{n+1}, \dots, t_{k-1}\}$ represents a unique arrangement of topics. Let π be a strategy (permutation) derived from P that satisfies GI in each move (according to Definition 53). The function $generatePermutations(D_r^n, g)$ in ETP generates all permutations of topics in P . Any candidate strategy $\pi \in P$ must contain at least one topic that supports the goal topic g through a belief $b \in \mathcal{U}^B$, according to a belief-topic dependency graph BT (Definition 41). This support relation is satisfied if there exists a pair (t', g) such that $t' \in dependent_topics^+(b, \mathcal{U}^T)$ and $g \in dependent_topics^+(b, \mathcal{U}^T)$ for some b . Since all strategies are filtered in $findCandidateStrategies(D_r^n, g, BT)$ based on if they contain at least one topic with a support relation to g , any valid permutation π will be included in the set of candidate strategies generated by ETP : $\pi \in \Pi \leftarrow findCandidateStrategies(D_r^n, g, BT)$. Then, through forward-checking, the function $selectGoalHidingStrategy(\Pi, g, Q_r^n, BT)$ iteratively creates QBAFs for each topic in each strategy: $\forall \pi' \in \Pi, \forall t_p \in \pi', (n+1 \leq p \leq k)$, $Q_r^p = Q_r^{p-1} \cup \{q_p\}$, $q_p = \langle X_p, R_p^-, R_p^+, \tau_p \rangle$, for which the following conditions are checked:

1. evaluate the sensitivity interval: $\rho^l \leq I_\delta(Q_r^p, t_p) \leq \rho^u$ (according to GI condition C2 and C3 in Theorem 5),
2. evaluate strength monotonicity: $I_\delta(Q_r^n, g) \leq I_\delta(Q_r^p, g)$ (according to GI condition C1 in Theorem 5).

Any strategy $\pi'' \in \Pi$ violating the conditions of GI is discarded. Therefore, if there exists a strategy π in strategy space P that satisfies GI , then $\pi' := ETP(\gamma, BT, \mathcal{E}, [\rho^l, \rho^u])$ and $\pi' = \pi$. ■

Soundness ensures that Algorithm 1 correctly identifies strategies that adhere to the goal-hiding invariant, thereby upholding the properties of sensitivity interval and strength monotonicity throughout the dialogue system's operation. On the other hand, completeness guarantees that Algorithm 1 captures all relevant strategies that satisfy the goal-hiding invariant, affirming its efficacy in exploring the solution space.

Analysis of Gradual Semantics

In order to better understand the behavior of the introduced gradual semantics (Definitions 54-55), we will in this section analyze the semantics in relation to a set of 11 basic ideas of QBAFs introduced in [29]. These basic ideas outline

how the strengths of arguments in a QBAF are expected to be weakened or strengthened in relation to the balance of attacks and supports within a QBAF.

The basic ideas introduced in [29] are the following:

- **Basic idea 1:** The strength of an argument differs from its base score only if the argument is the dialectical target of other arguments.
Formally: For an argument α , if $R^-(\alpha) = \emptyset$ and $R^+(\alpha) = \emptyset$ then $\delta(\alpha) = \tau(\alpha)$.
- **Basic idea 2:** In the absence of supporters, if there is at least an attacker then the strength of an argument is lower than its base score.
Formally: For an argument α , if $R^-(\alpha) \neq \emptyset$ and $R^+(\alpha) = \emptyset$, then $\delta(\alpha) \leq \tau(\alpha)$.
- **Basic idea 3:** In the absence of attackers, if there is at least a supporter then the strength of an argument is greater than its base score.
Formally: For an argument α , if $R^-(\alpha) = \emptyset$ and $R^+(\alpha) \neq \emptyset$, then $\tau(\alpha) \leq \delta(\alpha)$.
- **Basic idea 4:** If the strength of an argument is lower than its base score then the argument has at least one attacker.
Formally: For an argument α , if $\delta(\alpha) < \tau(\alpha)$, then $R^-(\alpha) \neq \emptyset$.
- **Basic idea 5:** If the strength of an argument is higher than its base score then the argument has at least one supporter.
Formally: For an argument α , if $\delta(\alpha) > \tau(\alpha)$, then $R^+(\alpha) \neq \emptyset$.
- **Basic idea 6:** Arguments with equal conditions in terms of attackers, supporters and base score have the same strength.
Formally: For an argument α and an argument β , if $R^-(\alpha) = R^-(\beta)$, $R^+(\alpha) = R^+(\beta)$, and $\tau(\alpha) = \tau(\beta)$, then $\delta(\alpha) = \delta(\beta)$.
- **Basic idea 7:** A strictly larger set of attackers determines a lower strength.
Formally: For an argument α and an argument β , if $R^-(\alpha) \subseteq R^-(\beta)$, $R^+(\alpha) = R^+(\beta)$, and $\tau(\alpha) = \tau(\beta)$, then $\delta(\beta) \leq \delta(\alpha)$.
- **Basic idea 8:** A strictly larger set of supporters determines a higher strength.
Formally: For an argument α and an argument β , if $R^-(\alpha) = R^-(\beta)$, $R^+(\alpha) \subseteq R^+(\beta)$, and $\tau(\alpha) = \tau(\beta)$, then $\delta(\alpha) \leq \delta(\beta)$.
- **Basic idea 9:** A higher base score gives a higher strength.
Formally: For an argument α and an argument β , if $R^-(\alpha) = R^-(\beta)$, $R^+(\alpha) = R^+(\beta)$, and $\tau(\alpha) < \tau(\beta)$, then $\delta(\alpha) \leq \delta(\beta)$.
- **Basic idea 10:** A weaker set of attackers determines a higher strength.
Formally: For an argument α and an argument β , if $\delta(R^-(\alpha)) < \delta(R^-(\beta))$, $\delta(R^+(\alpha)) = \delta(R^+(\beta))$, and $\tau(\alpha) = \tau(\beta)$, then $\delta(\beta) \leq \delta(\alpha)$.
- **Basic idea 11:** A stronger set of supporters determines a higher strength.
Formally: For an argument α and an argument β , if $\delta(R^-(\alpha)) = \delta(R^-(\beta))$, $\delta(R^+(\alpha)) > \delta(R^+(\beta))$, and $\tau(\alpha) = \tau(\beta)$, then $\delta(\beta) \leq \delta(\alpha)$.

Our analysis of gradual semantics considers the following premises: Let $\gamma = \langle \mathcal{I}, D_r^n, Q_r^n \rangle$ be an information-seeking dialogue system, such that $\mathcal{I} = \{se, re\}$, $se = \langle G, \mathcal{U}_{se}, \delta, TS \rangle$, $re = \langle \mathcal{U}_{re}, BS \rangle$, $Q_r^n = [q_r, \dots, q_n]$ and $q_i = \langle X_i, R_i^-, R_i^+, \tau_i \rangle$, ($r \leq i \leq n$), where $x \in X_i$ is a topic, $g \in G$ is a goal topic, δ is a strength function, \mathcal{P} is the set of strategies w.r.t. g of the seeker, and $\mathcal{O} \subseteq 2^{\mathcal{U}_{re}}$ are the strategies of the respondent, $P \in \mathcal{P}$ and $O \in \mathcal{O}$, and $w(X) = agg(\{\delta(z) \mid z \in X\})$ is a strength aggregation function, where $agg \in \{min, max, average\}$. We then formalize a proposition and proof for each of the basic ideas in terms of the gradual semantics (Definitions 54-55). The resulting propositions are summarized in Table 5.1.

According to *Basic idea 1* [29]: The strength of an argument differs from its base score only if the argument is the dialectical target of other arguments. Formally stated as: For an argument α , if $R^-(\alpha) = \emptyset$ and $R^+(\alpha) = \emptyset$ then

Table 5.1: Gradual semantics analysis considering the 11 basic ideas of QBAFs [29].

#	Basic idea expression	Prop
1	For an argument $x \in X_i$, if $P^{\leftarrow O}(R_i^-(x)) = \emptyset$ and $P^{\leftarrow P \cup O}(R_i^+(x)) = \emptyset$, then $\delta_i(x) = \tau_i(x)$.	Proposition 2
2	For an argument $x \in X_i$, if $P^{\leftarrow O}(R_i^-(x)) \neq \emptyset$ and $P^{\leftarrow P \cup O}(R_i^+(x)) = \emptyset$, then $\delta_i(x) \leq \tau_i(x)$.	Proposition 3
3	For an argument $x \in X_i$, if $ P^{\leftarrow O}(R_i^-(x)) * w(P^{\leftarrow O}(R_i^-(x))) = 0$ and $ P^{\leftarrow P \cup O}(R_i^+(x)) * w(P^{\leftarrow P \cup O}(R_i^+(x))) \neq 0$, then $\tau_i(x) \leq \delta_i(x)$.	Proposition 4
4	For an argument $x \in X_i$, if $\delta_i(x) < \tau_i(x)$ then $P^{\leftarrow O}(R_i^-(x)) \neq \emptyset$.	Proposition 5
5	For an argument $x \in X_i$, if $\delta_i(x) > \tau_i(x)$ then $P^{\leftarrow P \cup O}(R_i^+(x)) \neq \emptyset$.	Proposition 6
6	For the arguments $x, y \in X_i$, if $ P^{\leftarrow O}(R_i^-(x)) * w(P^{\leftarrow O}(R_i^-(x))) = P^{\leftarrow O}(R_i^-(y)) * w(P^{\leftarrow O}(R_i^-(y)))$, $ P^{\leftarrow P \cup O}(R_i^+(x)) * w(P^{\leftarrow P \cup O}(R_i^+(x))) = P^{\leftarrow P \cup O}(R_i^+(y)) * w(P^{\leftarrow P \cup O}(R_i^+(y)))$, and $\tau_i(x) = \tau_i(y)$, then $\delta_i(x) = \delta_i(y)$.	Proposition 7
7	For the arguments $x, y \in X_i$, if $ P^{\leftarrow O}(R_i^-(x)) * w(P^{\leftarrow O}(R_i^-(x))) \geq P^{\leftarrow O}(R_i^-(y)) * w(P^{\leftarrow O}(R_i^-(y)))$, $ P^{\leftarrow P \cup O}(R_i^+(x)) * w(P^{\leftarrow P \cup O}(R_i^+(x))) = P^{\leftarrow P \cup O}(R_i^+(y)) * w(P^{\leftarrow P \cup O}(R_i^+(y)))$, and $\tau_i(x) = \tau_i(y)$, then $\delta_i(x) \leq \delta_i(y)$.	Proposition 8
8	For the arguments $x, y \in X_i$, if $ P^{\leftarrow P \cup O}(R_i^+(x)) * w(P^{\leftarrow P \cup O}(R_i^+(x))) \geq P^{\leftarrow P \cup O}(R_i^+(y)) * w(P^{\leftarrow P \cup O}(R_i^+(y)))$, $ P^{\leftarrow O}(R_i^-(x)) * w(P^{\leftarrow O}(R_i^-(x))) = P^{\leftarrow O}(R_i^-(y)) * w(P^{\leftarrow O}(R_i^-(y)))$, and $\tau_i(x) = \tau_i(y)$, then $\delta_i(x) \geq \delta_i(y)$.	Proposition 9
9	For the arguments $x, y \in X_i$, if $ P^{\leftarrow O}(R_i^-(x)) * w(P^{\leftarrow O}(R_i^-(x))) = P^{\leftarrow O}(R_i^-(y)) * w(P^{\leftarrow O}(R_i^-(y)))$, $ P^{\leftarrow P \cup O}(R_i^+(x)) * w(P^{\leftarrow P \cup O}(R_i^+(x))) = P^{\leftarrow P \cup O}(R_i^+(y)) * w(P^{\leftarrow P \cup O}(R_i^+(y)))$, and $\tau_i(x) < \tau_i(y)$, then $\delta_i(x) \leq \delta_i(y)$.	Proposition 10
10	For the arguments $x, y \in X_i$, if $ P^{\leftarrow O}(R_i^-(x)) * w(P^{\leftarrow O}(R_i^-(x))) > P^{\leftarrow O}(R_i^-(y)) * w(P^{\leftarrow O}(R_i^-(y)))$, $ P^{\leftarrow P \cup O}(R_i^+(x)) * w(P^{\leftarrow P \cup O}(R_i^+(x))) = P^{\leftarrow P \cup O}(R_i^+(y)) * w(P^{\leftarrow P \cup O}(R_i^+(y)))$, and $\tau_i(x) = \tau_i(y)$, then $\delta_i(x) \leq \delta_i(y)$.	Proposition 11
11	For the arguments $x, y \in X_i$, if $ P^{\leftarrow P \cup O}(R_i^+(x)) * w(P^{\leftarrow P \cup O}(R_i^+(x))) > P^{\leftarrow P \cup O}(R_i^+(y)) * w(P^{\leftarrow P \cup O}(R_i^+(y)))$, $ P^{\leftarrow O}(R_i^-(x)) * w(P^{\leftarrow O}(R_i^-(x))) = P^{\leftarrow O}(R_i^-(y)) * w(P^{\leftarrow O}(R_i^-(y)))$, and $\tau_i(x) = \tau_i(y)$, then $\delta_i(x) \geq \delta_i(y)$.	Proposition 12

$\delta(\alpha) = \tau(\alpha)$. In order to show whether this holds for the proposed gradual semantics (Definitions 54 and 55), we specify the following proposition and proof.

Proposition 2 *Let $\gamma = \langle \mathcal{I}, D_r^n, Q_r^n \rangle$ be an information-seeking dialogue system, such that $\mathcal{I} = \{se, re\}$, $se = \langle G, \mathcal{U}_{se}, \delta, TS \rangle$, $re = \langle \mathcal{U}_{re}, BS \rangle$, $Q_r^n = [q_r, \dots, q_n]$ and $q_i = \langle X_i, R_i^-, R_i^+, \tau_i \rangle$, ($r \leq i \leq n$), where $x \in X_i$ is a topic, $g \in G$ is a goal topic, δ is a strength function, \mathcal{P} is the set of strategies w.r.t. g of the seeker, and $\mathcal{O} \subseteq 2^{\mathcal{U}_{re}}$ are the strategies of the respondent, $P \in \mathcal{P}$ and $O \in \mathcal{O}$, and $w(X) = \text{agg}(\{\delta(z) \mid z \in X\})$ is a strength aggregation function, where $\text{agg} \in \{\min, \max, \text{average}\}$.*

For an argument $x \in X_i$, if $P^{\leftarrow O}(R_i^-(x)) = \emptyset$ and $P^{\leftarrow P \cup O}(R_i^+(x)) = \emptyset$, then $\delta_i(x) = \tau_i(x)$.

Proof: Let x be an argument, $\tau_i(x)$ represent its base score and $\delta_i(x)$ represent its strength. By considering the gradual strength evaluation (Definitions 54 and 55), where $P^{\leftarrow O}(R^-(x))$ is the attacks on x in strategy P , and $P^{\leftarrow P \cup O}(R^+(x))$ is the supports on x in strategy P , we can observe the following cases:

- (A) If $P^{\leftarrow O}(R_i^-(x)) = \emptyset$ and $P^{\leftarrow P \cup O}(R_i^+(x)) = \emptyset$, then $|P^{\leftarrow O}(R^-(x))| * w(P^{\leftarrow O}(R^-(x))) = |P^{\leftarrow P \cup O}(R^+(x))| * w(P^{\leftarrow P \cup O}(R^+(x)))$. Consequently, if $|P^{\leftarrow O}(R^-(x))| * w(P^{\leftarrow O}(R^-(x))) = |P^{\leftarrow P \cup O}(R^+(x))| * w(P^{\leftarrow P \cup O}(R^+(x)))$, then $\tau_{i+1}(x) = \delta_{i+1}(x)$.
- (B) If $|P^{\leftarrow O}(R^-(x))| * w(P^{\leftarrow O}(R^-(x))) < |P^{\leftarrow P \cup O}(R^+(x))| * w(P^{\leftarrow P \cup O}(R^+(x)))$, then $\tau_{i+1}(x) < \delta_{i+1}(x)$.
- (C) If $|P^{\leftarrow O}(R^-(x))| * w(P^{\leftarrow O}(R^-(x))) > |P^{\leftarrow P \cup O}(R^+(x))| * w(P^{\leftarrow P \cup O}(R^+(x)))$, then $\tau_{i+1}(x) > \delta_{i+1}(x)$.

Only cases B and C, where there is an imbalance between attacks and supports for x , result in a strength of x that differs from the base score of x . Since the cases exhaust the possibilities, we can conclude that Basic idea 1 [29] is followed by the gradual semantics. ■

According to *Basic idea 2* [29]: In the absence of supporters, if there is at least an attacker then the strength of an argument is lower than its base score. Formally stated as: For an argument α , if $R^-(\alpha) \neq \emptyset$ and $R^+(\alpha) = \emptyset$, then $\delta(\alpha) \leq \tau(\alpha)$. In order to show whether this holds for the proposed gradual semantics (Definitions 54 and 55), we specify the following proposition and proof.

Proposition 3 *Let $\gamma = \langle \mathcal{I}, D_r^n, Q_r^n \rangle$ be an information-seeking dialogue system, such that $\mathcal{I} = \{se, re\}$, $se = \langle G, \mathcal{U}_{se}, \delta, TS \rangle$, $re = \langle \mathcal{U}_{re}, BS \rangle$, $Q_r^n = [q_r, \dots, q_n]$ and $q_i = \langle X_i, R_i^-, R_i^+, \tau_i \rangle$, ($r \leq i \leq n$), where $x \in X_i$ is a topic, $g \in G$ is a goal topic, δ is a strength function, \mathcal{P} is the set of strategies w.r.t. g of the seeker, and $\mathcal{O} \subseteq 2^{\mathcal{U}_{re}}$ are the strategies of the respondent, $P \in \mathcal{P}$ and $O \in \mathcal{O}$, and $w(X) = \text{agg}(\{\delta(z) \mid z \in X\})$ is a strength aggregation function, where $\text{agg} \in \{\min, \max, \text{average}\}$.*

For an argument $x \in X_i$, if $P^{\leftarrow O}(R_i^-(x)) \neq \emptyset$ and $P^{\leftarrow P \cup O}(R_i^+(x)) = \emptyset$, then $\delta_i(x) \leq \tau_i(x)$.

Proof: Let x be an argument, $\tau_i(x)$ represent its base score and $\delta_i(x)$ represent its strength. By considering the gradual strength evaluation (Definitions 54 and 55), where $P^{\leftarrow O}(R^-(x))$ is the attacks on x in strategy P , and $P^{\leftarrow P \cup O}(R^+(x))$ is the supports on x in strategy P , we can observe the following cases:

- (A) If $|P^{\leftarrow O}(R^-(x))| * w(P^{\leftarrow O}(R^-(x))) = 0$ and $|P^{\leftarrow P \cup O}(R^+(x))| * w(P^{\leftarrow P \cup O}(R^+(x))) = 0$, then $\delta_{i+1}(x) = \tau_{i+1}(x)$.
- (B) If $|P^{\leftarrow O}(R^-(x))| * w(P^{\leftarrow O}(R^-(x))) > 0$ and $|P^{\leftarrow P \cup O}(R^+(x))| * w(P^{\leftarrow P \cup O}(R^+(x))) = 0$, then $\delta_{i+1}(x) \leq \tau_{i+1}(x)$.

Only case B, where there is an absence of supporters of x and at least an attacker of x , results in a strength of x lower than the base score of x . Since the cases exhaust the possibilities, we can conclude that Basic idea 2 [29] is followed by the gradual semantics. ■

According to *Basic idea 3* [29]: In the absence of attackers, if there is at least a supporter then the strength of an argument is greater than its base score. Formally stated as: For an argument α , if $R^-(\alpha) = \emptyset$ and $R^+(\alpha) \neq \emptyset$, then $\tau(\alpha) \leq \delta(\alpha)$. In order to show whether this holds for the proposed gradual semantics (Definitions 54 and 55), we specify the following proposition and proof.

Proposition 4 *Let $\gamma = \langle \mathcal{I}, D_r^n, Q_r^n \rangle$ be an information-seeking dialogue system, such that $\mathcal{I} = \{se, re\}$, $se = \langle G, \mathcal{U}_{se}, \delta, TS \rangle$, $re = \langle \mathcal{U}_{re}, BS \rangle$, $Q_r^n = [q_r, \dots, q_n]$ and $q_i = \langle X_i, R_i^-, R_i^+, \tau_i \rangle$, ($r \leq i \leq n$), where $x \in X_i$ is a topic, $g \in G$ is a goal topic, δ is a strength function, \mathcal{P} is the set of strategies w.r.t. g of the seeker, and $\mathcal{O} \subseteq 2^{\mathcal{U}_{re}}$ are the strategies of the respondent, $P \in \mathcal{P}$ and $O \in \mathcal{O}$, and $w(X) = \text{agg}(\{\delta(z) \mid z \in X\})$ is a strength aggregation function, where $\text{agg} \in \{\min, \max, \text{average}\}$.*

*For an argument $x \in X_i$, if $|P^{\leftarrow O}(R_i^-(x))| * w(P^{\leftarrow O}(R_i^-(x))) = 0$ and $|P^{\leftarrow P \cup O}(R_i^+(x))| * w(P^{\leftarrow P \cup O}(R_i^+(x))) \neq 0$, then $\tau_i(x) \leq \delta_i(x)$.*

Proof: Let x be an argument, $\tau_i(x)$ represent its base score and $\delta_i(x)$ represent its strength. By considering the gradual strength evaluation (Definitions 54 and 55), where $P^{\leftarrow O}(R^-(x))$ is the attacks on x in strategy P , and $P^{\leftarrow P \cup O}(R^+(x))$ is the supports on x in strategy P , we can observe the following cases:

- (A) If $|P^{\leftarrow P \cup O}(R^+(x))| * w(P^{\leftarrow P \cup O}(R^+(x))) = 0$ and $|P^{\leftarrow O}(R^-(x))| * w(P^{\leftarrow O}(R^-(x))) = 0$, then $\delta_{i+1}(x) = \tau_{i+1}(x)$.
- (B) If $|P^{\leftarrow P \cup O}(R^+(x))| * w(P^{\leftarrow P \cup O}(R^+(x))) > 0$ and $|P^{\leftarrow O}(R^-(x))| * w(P^{\leftarrow O}(R^-(x))) = 0$, then $\delta_{i+1}(x) \geq \tau_{i+1}(x)$.

Only case B, where there is an absence of attackers of x and at least a supporter of x , results in a strength of x higher than the base score of x . Since the cases exhaust the possibilities, we can conclude that Basic idea 2 [29] is followed by the gradual semantics. ■

According to *Basic idea 4* [29]: If the strength of an argument is lower than its base score then the argument has at least one attacker. Formally stated as: For an argument α , if $\delta(\alpha) < \tau(\alpha)$, then $R^-(\alpha) \neq \emptyset$. In order to show whether this holds for the proposed gradual semantics (Definitions 54 and 55), we specify the following proposition and proof.

Proposition 5 *Let $\gamma = \langle \mathcal{I}, D_r^n, Q_r^n \rangle$ be an information-seeking dialogue system, such that $\mathcal{I} = \{se, re\}$, $se = \langle G, \mathcal{U}_{se}, \delta, TS \rangle$, $re = \langle \mathcal{U}_{re}, BS \rangle$, $Q_r^n = [q_r, \dots, q_n]$ and $q_i = \langle X_i, R_i^-, R_i^+, \tau_i \rangle$, ($r \leq i \leq n$), where $x \in X_i$ is a topic, $g \in G$ is a goal topic, δ is a strength function, \mathcal{P} is the set of strategies w.r.t. g of the seeker, and $\mathcal{O} \subseteq \mathcal{2}^{\mathcal{U}_{re}}$ are the strategies of the respondent, $P \in \mathcal{P}$ and $O \in \mathcal{O}$, and $w(X) = \text{agg}(\{\delta(z) \mid z \in X\})$ is a strength aggregation function, where $\text{agg} \in \{\min, \max, \text{average}\}$.*

For an argument $x \in X_i$, if $\delta_i(x) < \tau_i(x)$ then $P^{\leftarrow O}(R_i^-(x)) \neq \emptyset$.

Proof: Let $x \in X_i$ be an argument with base score $\tau_i(x)$ and strength $\delta_i(x)$. Given $\delta_i(x) < \tau_i(x)$, we have $\beta * |P^{\leftarrow(R_i^+(x))}| * w(P^{\leftarrow(R_i^+(x))}) - |P^{\leftarrow O}(R_i^-(x))| * w(P^{\leftarrow O}(R_i^-(x))) < 0$. It follows that $|P^{\leftarrow(R_i^+(x))}| * w(P^{\leftarrow(R_i^+(x))}) < |P^{\leftarrow O}(R_i^-(x))| * w(P^{\leftarrow O}(R_i^-(x)))$, implying $P^{\leftarrow O}(R_i^-(x)) \neq \emptyset$. ■

According to *Basic idea 5* [29]: If the strength of an argument is higher than its base score then the argument has at least one supporter. Formally stated as: For an argument α , if $\delta(\alpha) > \tau(\alpha)$, then $R^+(\alpha) \neq \emptyset$. In order to show whether this holds for the proposed gradual semantics (Definitions 54 and 55), we specify the following proposition and proof.

Proposition 6 *Let $\gamma = \langle \mathcal{I}, D_r^n, Q_r^n \rangle$ be an information-seeking dialogue system, such that $\mathcal{I} = \{se, re\}$, $se = \langle G, \mathcal{U}_{se}, \delta, TS \rangle$, $re = \langle \mathcal{U}_{re}, BS \rangle$, $Q_r^n = [q_r, \dots, q_n]$ and $q_i = \langle X_i, R_i^-, R_i^+, \tau_i \rangle$, ($r \leq i \leq n$), where $x \in X_i$ is a topic, $g \in G$ is a goal topic, δ is a strength function, \mathcal{P} is the set of strategies w.r.t. g of the seeker, and $\mathcal{O} \subseteq \mathcal{2}^{\mathcal{U}_{re}}$ are the strategies of the respondent, $P \in \mathcal{P}$ and $O \in \mathcal{O}$, and $w(X) = \text{agg}(\{\delta(z) \mid z \in X\})$ is a strength aggregation function, where $\text{agg} \in \{\min, \max, \text{average}\}$.*

For an argument $x \in X_i$, if $\delta_i(x) > \tau_i(x)$ then $P^{\leftarrow P \cup O}(R_i^+(x)) \neq \emptyset$.

Proof: Let $x \in X_i$ be an argument with base score $\tau_i(x)$ and strength $\delta_i(x)$. Given $\delta_i(x) > \tau_i(x)$, we have $\beta * |P^{\leftarrow(R_i^+(x))}| * w(P^{\leftarrow(R_i^+(x))}) - |P^{\leftarrow O}(R_i^-(x))| * w(P^{\leftarrow O}(R_i^-(x))) > 0$. It follows that $|P^{\leftarrow(R_i^+(x))}| * w(P^{\leftarrow(R_i^+(x))}) > |P^{\leftarrow O}(R_i^-(x))| * w(P^{\leftarrow O}(R_i^-(x)))$, implying $P^{\leftarrow P \cup O}(R_i^+(x)) \neq \emptyset$. ■

According to *Basic idea 6* [29]: Arguments with equal conditions in terms of attackers, supporters and base score have the same strength. Formally stated as: For an argument α and an argument β , if $R^-(\alpha) = R^-(\beta)$, $R^+(\alpha) = R^+(\beta)$, and $\tau(\alpha) = \tau(\beta)$, then $\delta(\alpha) = \delta(\beta)$. In order to show whether this holds for the proposed gradual semantics (Definitions 54 and 55), we specify the following proposition and proof.

Proposition 7 Let $\gamma = \langle \mathcal{I}, D_r^n, Q_r^n \rangle$ be an information-seeking dialogue system, such that $\mathcal{I} = \{se, re\}$, $se = \langle G, \mathcal{U}_{se}, \delta, TS \rangle$, $re = \langle \mathcal{U}_{re}, BS \rangle$, $Q_r^n = [q_r, \dots, q_n]$ and $q_i = \langle X_i, R_i^-, R_i^+, \tau_i \rangle$, ($r \leq i \leq n$), where $x \in X_i$ is a topic, $y \in X_i$ is a topic, $g \in G$ is a goal topic, δ is a strength function, \mathcal{P} is the set of strategies w.r.t. g of the seeker, and $\mathcal{O} \subseteq 2^{\mathcal{U}_{re}}$ are the strategies of the respondent, $P \in \mathcal{P}$ and $O \in \mathcal{O}$, and $w(X) = \text{agg}(\{\delta(z) \mid z \in X\})$ is a strength aggregation function, where $\text{agg} \in \{\min, \max, \text{average}\}$.

For the arguments $x, y \in X_i$, if $|P^{\leftarrow O}(R_i^-(x))| * w(P^{\leftarrow O}(R_i^-(x))) = |P^{\leftarrow O}(R_i^-(y))| * w(P^{\leftarrow O}(R_i^-(y)))$, $|P^{\leftarrow P \cup O}(R_i^+(x))| * w(P^{\leftarrow P \cup O}(R_i^+(x))) = |P^{\leftarrow P \cup O}(R_i^+(y))| * w(P^{\leftarrow P \cup O}(R_i^+(y)))$, and $\tau_i(x) = \tau_i(y)$, then $\delta_i(x) = \delta_i(y)$.

Proof: Let x and y be an arguments, $\tau(x)$ and $\tau(y)$ represent their base scores and $\delta(x)$ and $\delta(y)$ represent their strengths. By considering the gradual strength evaluation (Definitions 54 and 55), where $P^{\leftarrow O}(R^-(x))$ is the attacks on x in strategy P , $P^{\leftarrow P \cup O}(R^+(x))$ is the supports on x in strategy P , $P^{\leftarrow O}(R^-(y))$ is the attacks on y in strategy P , and $P^{\leftarrow P \cup O}(R^+(y))$ is the supports on y in strategy P , we can observe the following cases:

- (A) If $|P^{\leftarrow O}(R_i^-(x))| * w(P^{\leftarrow O}(R_i^-(x))) \neq |P^{\leftarrow O}(R_i^-(y))| * w(P^{\leftarrow O}(R_i^-(y)))$, $|P^{\leftarrow P \cup O}(R_i^+(x))| * w(P^{\leftarrow P \cup O}(R_i^+(x))) = |P^{\leftarrow P \cup O}(R_i^+(y))| * w(P^{\leftarrow P \cup O}(R_i^+(y)))$, and $\tau_i(x) = \tau_i(y)$, then $\delta_i(x) \neq \delta_i(y)$.
- (B) If $|P^{\leftarrow O}(R_i^-(x))| * w(P^{\leftarrow O}(R_i^-(x))) = |P^{\leftarrow O}(R_i^-(y))| * w(P^{\leftarrow O}(R_i^-(y)))$, $|P^{\leftarrow P \cup O}(R_i^+(x))| * w(P^{\leftarrow P \cup O}(R_i^+(x))) \neq |P^{\leftarrow P \cup O}(R_i^+(y))| * w(P^{\leftarrow P \cup O}(R_i^+(y)))$, and $\tau_i(x) = \tau_i(y)$, then $\delta_i(x) \neq \delta_i(y)$.
- (C) If $|P^{\leftarrow O}(R_i^-(x))| * w(P^{\leftarrow O}(R_i^-(x))) = |P^{\leftarrow O}(R_i^-(y))| * w(P^{\leftarrow O}(R_i^-(y)))$, $|P^{\leftarrow P \cup O}(R_i^+(x))| * w(P^{\leftarrow P \cup O}(R_i^+(x))) = |P^{\leftarrow P \cup O}(R_i^+(y))| * w(P^{\leftarrow P \cup O}(R_i^+(y)))$, and $\tau_i(x) \neq \tau_i(y)$, then $\delta_i(x) \neq \delta_i(y)$.
- (D) If $|P^{\leftarrow O}(R_i^-(x))| * w(P^{\leftarrow O}(R_i^-(x))) \neq |P^{\leftarrow O}(R_i^-(y))| * w(P^{\leftarrow O}(R_i^-(y)))$, $|P^{\leftarrow P \cup O}(R_i^+(x))| * w(P^{\leftarrow P \cup O}(R_i^+(x))) \neq |P^{\leftarrow P \cup O}(R_i^+(y))| * w(P^{\leftarrow P \cup O}(R_i^+(y)))$, and $\tau_i(x) = \tau_i(y)$, then $\delta_i(x) \neq \delta_i(y)$.
- (E) If $|P^{\leftarrow O}(R_i^-(x))| * w(P^{\leftarrow O}(R_i^-(x))) = |P^{\leftarrow O}(R_i^-(y))| * w(P^{\leftarrow O}(R_i^-(y)))$, $|P^{\leftarrow P \cup O}(R_i^+(x))| * w(P^{\leftarrow P \cup O}(R_i^+(x))) = |P^{\leftarrow P \cup O}(R_i^+(y))| * w(P^{\leftarrow P \cup O}(R_i^+(y)))$, and $\tau_i(x) = \tau_i(y)$, then $\delta_i(x) = \delta_i(y)$.

Only case E, with equal conditions in terms of attackers, supporters and base score, results in equal strength between x and y . Since the cases exhaust the possibilities, we can conclude that Basic Idea 6 [29] is followed by the gradual semantics. ■

According to *Basic idea 7* [29]: A strictly larger set of attackers determines a lower strength. Formally stated as: For an argument α and an argument β , if $R^-(\alpha) \subseteq R^-(\beta)$, $R^+(\alpha) = R^+(\beta)$, and $\tau(\alpha) = \tau(\beta)$, then $\delta(\beta) \leq \delta(\alpha)$. In order to show whether this holds for the proposed gradual semantics (Definitions 54 and 55), we specify the following proposition and proof.

Proposition 8 Let $\gamma = \langle \mathcal{I}, D_r^n, Q_r^n \rangle$ be an information-seeking dialogue system, such that $\mathcal{I} = \{se, re\}$, $se = \langle G, \mathcal{U}_{se}, \delta, TS \rangle$, $re = \langle \mathcal{U}_{re}, BS \rangle$, $Q_r^n = [q_r, \dots, q_n]$ and $q_i = \langle X_i, R_i^-, R_i^+, \tau_i \rangle$, ($r \leq i \leq n$), where $x \in X_i$ is a topic,

$y \in X_i$ is a topic, $g \in G$ is a goal topic, δ is a strength function, \mathcal{P} is the set of strategies w.r.t. g of the seeker, and $\mathcal{O} \subseteq 2^{\mathcal{U}^{re}}$ are the strategies of the respondent, $P \in \mathcal{P}$ and $O \in \mathcal{O}$, and $w(X) = \text{agg}(\{\delta(z) \mid z \in X\})$ is a strength aggregation function, where $\text{agg} \in \{\min, \max, \text{average}\}$.

For the arguments $x, y \in X_i$, if $|P^{\leftarrow O}(R_i^-(x))| * w(P^{\leftarrow O}(R_i^-(x))) \geq |P^{\leftarrow O}(R_i^-(y))| * w(P^{\leftarrow O}(R_i^-(y)))$, $|P^{\leftarrow P \cup O}(R_i^+(x))| * w(P^{\leftarrow P \cup O}(R_i^+(x))) = |P^{\leftarrow P \cup O}(R_i^+(y))| * w(P^{\leftarrow P \cup O}(R_i^+(y)))$, and $\tau_i(x) = \tau_i(y)$, then $\delta_i(x) \leq \delta_i(y)$.

Proof: Let x and y be an arguments, $\tau(x)$ and $\tau(y)$ represent their base scores and $\delta(x)$ and $\delta(y)$ represent their strengths. By considering the gradual strength evaluation (Definitions 54 and 55), where $P^{\leftarrow O}(R^-(x))$ is the attacks on x in strategy P , $P^{\leftarrow P \cup O}(R^+(x))$ is the supports on x in strategy P , $P^{\leftarrow O}(R^-(y))$ is the attacks on y in strategy P , and $P^{\leftarrow P \cup O}(R^+(y))$ is the supports on y in strategy P , we can observe the following cases:

- (A) If $|P^{\leftarrow O}(R_i^-(x))| * w(P^{\leftarrow O}(R_i^-(x))) = |P^{\leftarrow O}(R_i^-(y))| * w(P^{\leftarrow O}(R_i^-(y)))$, $|P^{\leftarrow P \cup O}(R_i^+(x))| * w(P^{\leftarrow P \cup O}(R_i^+(x))) = |P^{\leftarrow P \cup O}(R_i^+(y))| * w(P^{\leftarrow P \cup O}(R_i^+(y)))$, and $\tau_i(x) = \tau_i(y)$, then $\delta_i(x) = \delta_i(y)$.
- (B) If $|P^{\leftarrow O}(R_i^-(x))| * w(P^{\leftarrow O}(R_i^-(x))) \leq |P^{\leftarrow O}(R_i^-(y))| * w(P^{\leftarrow O}(R_i^-(y)))$, $|P^{\leftarrow P \cup O}(R_i^+(x))| * w(P^{\leftarrow P \cup O}(R_i^+(x))) = |P^{\leftarrow P \cup O}(R_i^+(y))| * w(P^{\leftarrow P \cup O}(R_i^+(y)))$, and $\tau_i(x) = \tau_i(y)$, then $\delta_i(x) \geq \delta_i(y)$.
- (C) If $|P^{\leftarrow O}(R_i^-(x))| * w(P^{\leftarrow O}(R_i^-(x))) \geq |P^{\leftarrow O}(R_i^-(y))| * w(P^{\leftarrow O}(R_i^-(y)))$, $|P^{\leftarrow P \cup O}(R_i^+(x))| * w(P^{\leftarrow P \cup O}(R_i^+(x))) = |P^{\leftarrow P \cup O}(R_i^+(y))| * w(P^{\leftarrow P \cup O}(R_i^+(y)))$, and $\tau_i(x) = \tau_i(y)$, then $\delta_i(x) \leq \delta_i(y)$.

Only case C, with a larger set of attackers of x , results in a lower strength of x . Since the cases exhaust the possibilities, we can conclude that Basic Idea 7 [29] is followed by the gradual semantics. ■

According to *Basic idea 8* [29]: A strictly larger set of supporters determines a higher strength. Formally stated as: For an argument α and an argument β , if $R^-(\alpha) = R^-(\beta)$, $R^+(\alpha) \subseteq R^+(\beta)$, and $\tau(\alpha) = \tau(\beta)$, then $\delta(\alpha) \leq \delta(\beta)$. In order to show whether this holds for the proposed gradual semantics (Definitions 54 and 55), we specify the following proposition and proof.

Proposition 9 Let $\gamma = \langle \mathcal{I}, D_r^n, Q_r^n \rangle$ be an information-seeking dialogue system, such that $\mathcal{I} = \{se, re\}$, $se = \langle G, \mathcal{U}_{se}, \delta, TS \rangle$, $re = \langle \mathcal{U}_{re}, BS \rangle$, $Q_r^n = [q_r, \dots, q_n]$ and $q_i = \langle X_i, R_i^-, R_i^+, \tau_i \rangle$, ($r \leq i \leq n$), where $x \in X_i$ is a topic, $y \in X_i$ is a topic, $g \in G$ is a goal topic, δ is a strength function, \mathcal{P} is the set of strategies w.r.t. g of the seeker, and $\mathcal{O} \subseteq 2^{\mathcal{U}^{re}}$ are the strategies of the respondent, $P \in \mathcal{P}$ and $O \in \mathcal{O}$, and $w(X) = \text{agg}(\{\delta(z) \mid z \in X\})$ is a strength aggregation function, where $\text{agg} \in \{\min, \max, \text{average}\}$.

For the arguments $x, y \in X_i$, if $|P^{\leftarrow P \cup O}(R_i^+(x))| * w(P^{\leftarrow P \cup O}(R_i^+(x))) \geq |P^{\leftarrow P \cup O}(R_i^+(y))| * w(P^{\leftarrow P \cup O}(R_i^+(y)))$, $|P^{\leftarrow O}(R_i^-(x))| * w(P^{\leftarrow O}(R_i^-(x))) = |P^{\leftarrow O}(R_i^-(y))| * w(P^{\leftarrow O}(R_i^-(y)))$, and $\tau_i(x) = \tau_i(y)$, then $\delta_i(x) \geq \delta_i(y)$.

Proof: Let x and y be an arguments, $\tau(x)$ and $\tau(y)$ represent their base scores and $\delta(x)$ and $\delta(y)$ represent their strengths. By considering the gradual strength evaluation (Definitions 54 and 55), where $P^{\leftarrow O}(R^-(x))$ is the

attacks on x in strategy P , $P^{\Leftarrow P \cup O}(R^+(x))$ is the supports on x in strategy P , $P^{\Leftarrow O}(R^-(y))$ is the attacks on y in strategy P , and $P^{\Leftarrow P \cup O}(R^+(y))$ is the supports on y in strategy P , we can observe the following cases:

- (A) If $|P^{\Leftarrow P \cup O}(R_i^+(x))| * w(P^{\Leftarrow P \cup O}(R_i^+(x))) = |P^{\Leftarrow P \cup O}(R_i^+(y))| * w(P^{\Leftarrow P \cup O}(R_i^+(y)))$, $|P^{\Leftarrow O}(R_i^-(x))| * w(P^{\Leftarrow O}(R_i^-(x))) = |P^{\Leftarrow O}(R_i^-(y))| * w(P^{\Leftarrow O}(R_i^-(y)))$, and $\tau_i(x) = \tau_i(y)$, then $\delta_i(x) = \delta_i(y)$.
- (B) If $|P^{\Leftarrow P \cup O}(R_i^+(x))| * w(P^{\Leftarrow P \cup O}(R_i^+(x))) \leq |P^{\Leftarrow P \cup O}(R_i^+(y))| * w(P^{\Leftarrow P \cup O}(R_i^+(y)))$, $|P^{\Leftarrow O}(R_i^-(x))| * w(P^{\Leftarrow O}(R_i^-(x))) = |P^{\Leftarrow O}(R_i^-(y))| * w(P^{\Leftarrow O}(R_i^-(y)))$, and $\tau_i(x) = \tau_i(y)$, then $\delta_i(x) \leq \delta_i(y)$.
- (C) If $|P^{\Leftarrow P \cup O}(R_i^+(x))| * w(P^{\Leftarrow P \cup O}(R_i^+(x))) \geq |P^{\Leftarrow P \cup O}(R_i^+(y))| * w(P^{\Leftarrow P \cup O}(R_i^+(y)))$, $|P^{\Leftarrow O}(R_i^-(x))| * w(P^{\Leftarrow O}(R_i^-(x))) = |P^{\Leftarrow O}(R_i^-(y))| * w(P^{\Leftarrow O}(R_i^-(y)))$, and $\tau_i(x) = \tau_i(y)$, then $\delta_i(x) \geq \delta_i(y)$.

Only case C, with a larger set of supporters of x than the set of supporters for y , results in a higher strength of x . Since the cases exhaust the possibilities, we can conclude that Basic Idea 8 [29] is followed by the gradual semantics. ■

According to *Basic idea 9* [29]: A higher base score gives a higher strength. Formally stated as: For an argument α and an argument β , if $R^-(\alpha) = R^-(\beta)$, $R^+(\alpha) = R^+(\beta)$, and $\tau(\alpha) < \tau(\beta)$, then $\delta(\alpha) \leq \delta(\beta)$. In order to show whether this holds for the proposed gradual semantics (Definitions 54 and 55), we specify the following proposition and proof.

Proposition 10 *Let $\gamma = \langle \mathcal{I}, D_r^n, Q_r^n \rangle$ be an information-seeking dialogue system, such that $\mathcal{I} = \langle se, re \rangle$, $se = \langle G, \mathcal{U}_{se}, \delta, TS \rangle$, $re = \langle \mathcal{U}_{re}, BS \rangle$, $Q_r^n = [q_r, \dots, q_n]$ and $q_i = \langle X_i, R_i^-, R_i^+, \tau_i \rangle$, ($r \leq i \leq n$), where $x \in X_i$ is a topic, $y \in X_i$ is a topic, $g \in G$ is a goal topic, δ is a strength function, \mathcal{P} is the set of strategies w.r.t. g of the seeker, and $\mathcal{O} \subseteq 2^{\mathcal{U}_{re}}$ are the strategies of the respondent, $P \in \mathcal{P}$ and $O \in \mathcal{O}$, and $w(X) = \text{agg}(\{\delta(z) \mid z \in X\})$ is a strength aggregation function, where $\text{agg} \in \{\min, \max, \text{average}\}$.*

*For the arguments $x, y \in X_i$, if $|P^{\Leftarrow O}(R_i^-(x))| * w(P^{\Leftarrow O}(R_i^-(x))) = |P^{\Leftarrow O}(R_i^-(y))| * w(P^{\Leftarrow O}(R_i^-(y)))$, $|P^{\Leftarrow P \cup O}(R_i^+(x))| * w(P^{\Leftarrow P \cup O}(R_i^+(x))) = |P^{\Leftarrow P \cup O}(R_i^+(y))| * w(P^{\Leftarrow P \cup O}(R_i^+(y)))$, and $\tau_i(x) < \tau_i(y)$, then $\delta_i(x) \leq \delta_i(y)$.*

Proof: Let x and y be an arguments, $\tau(x)$ and $\tau(y)$ represent their base scores and $\delta(x)$ and $\delta(y)$ represent their strengths. By considering the gradual strength evaluation (Definitions 54 and 55), where $P^{\Leftarrow O}(R^-(x))$ is the attacks on x in strategy P , $P^{\Leftarrow P \cup O}(R^+(x))$ is the supports on x in strategy P , $P^{\Leftarrow O}(R^-(y))$ is the attacks on y in strategy P , and $P^{\Leftarrow P \cup O}(R^+(y))$ is the supports on y in strategy P , we can observe the following cases:

- (A) If $|P^{\Leftarrow O}(R_i^-(x))| * w(P^{\Leftarrow O}(R_i^-(x))) = |P^{\Leftarrow O}(R_i^-(y))| * w(P^{\Leftarrow O}(R_i^-(y)))$, $|P^{\Leftarrow P \cup O}(R_i^+(x))| * w(P^{\Leftarrow P \cup O}(R_i^+(x))) = |P^{\Leftarrow P \cup O}(R_i^+(y))| * w(P^{\Leftarrow P \cup O}(R_i^+(y)))$, and $\tau_i(x) = \tau_i(y)$, then $\delta_i(x) = \delta_i(y)$.
- (B) If $|P^{\Leftarrow O}(R_i^-(x))| * w(P^{\Leftarrow O}(R_i^-(x))) = |P^{\Leftarrow O}(R_i^-(y))| * w(P^{\Leftarrow O}(R_i^-(y)))$, $|P^{\Leftarrow P \cup O}(R_i^+(x))| * w(P^{\Leftarrow P \cup O}(R_i^+(x))) = |P^{\Leftarrow P \cup O}(R_i^+(y))| * w(P^{\Leftarrow P \cup O}(R_i^+(y)))$, and $\tau_i(x) < \tau_i(y)$, then $\delta_i(x) \leq \delta_i(y)$.

- (C) If $|P^{\leftarrow O}(R_i^-(x))| * w(P^{\leftarrow O}(R_i^-(x))) = |P^{\leftarrow O}(R_i^-(y))| * w(P^{\leftarrow O}(R_i^-(y)))$,
 $|P^{\Leftarrow P \cup O}(R_i^+(x))| * w(P^{\Leftarrow P \cup O}(R_i^+(x))) = |P^{\Leftarrow P \cup O}(R_i^+(y))| * w(P^{\Leftarrow P \cup O}(R_i^+(y)))$,
and $\tau_i(x) > \tau_i(y)$, then $\delta_i(x) \geq \delta_i(y)$.

Only case B, with a similar set of supporters and attackers of x and y , but a higher base score of x , results in a higher strength of x compared to the strength of y . Since the cases exhaust the possibilities, we can conclude that Basic Idea 9 [29] is followed by the gradual semantics. ■

According to *Basic idea 10* [29]: A weaker set of attackers determines a higher strength. Formally stated as: For an argument α and an argument β , if $\delta(R^-(\alpha)) < \delta(R^-(\beta))$, $\delta(R^+(\alpha)) = \delta(R^+(\beta))$, and $\tau(\alpha) = \tau(\beta)$, then $\delta(\beta) \leq \delta(\alpha)$. In order to show whether this holds for the proposed gradual semantics (Definitions 54 and 55), we specify the following proposition and proof.

Proposition 11 *Let $\gamma = \langle \mathcal{I}, D_r^n, Q_r^n \rangle$ be an information-seeking dialogue system, such that $\mathcal{I} = \{se, re\}$, $se = \langle G, \mathcal{U}_{se}, \delta, TS \rangle$, $re = \langle \mathcal{U}_{re}, BS \rangle$, $Q_r^n = [q_r, \dots, q_n]$ and $q_i = \langle X_i, R_i^-, R_i^+, \tau_i \rangle$, ($r \leq i \leq n$), where $x \in X_i$ is a topic, $y \in X_i$ is a topic, $g \in G$ is a goal topic, δ is a strength function, \mathcal{P} is the set of strategies w.r.t. g of the seeker, and $\mathcal{O} \subseteq 2^{\mathcal{U}_{re}}$ are the strategies of the respondent, $P \in \mathcal{P}$ and $O \in \mathcal{O}$, and $w(X) = \text{agg}(\{\delta(z) \mid z \in X\})$ is a strength aggregation function, where $\text{agg} \in \{\min, \max, \text{average}\}$.*

*For the arguments $x, y \in X_i$, if $|P^{\leftarrow O}(R_i^-(x))| * w(P^{\leftarrow O}(R_i^-(x))) < |P^{\leftarrow O}(R_i^-(y))| * w(P^{\leftarrow O}(R_i^-(y)))$, $|P^{\Leftarrow P \cup O}(R_i^+(x))| * w(P^{\Leftarrow P \cup O}(R_i^+(x))) = |P^{\Leftarrow P \cup O}(R_i^+(y))| * w(P^{\Leftarrow P \cup O}(R_i^+(y)))$, and $\tau_i(x) = \tau_i(y)$, then $\delta_i(x) \leq \delta_i(y)$.*

Proof: Let x and y be an arguments, $\tau(x)$ and $\tau(y)$ represent their base scores and $\delta(x)$ and $\delta(y)$ represent their strengths. By considering the gradual strength evaluation (Definitions 54 and 55), where $P^{\leftarrow O}(R^-(x))$ is the attacks on x in strategy P , $P^{\Leftarrow P \cup O}(R^+(x))$ is the supports on x in strategy P , $P^{\leftarrow O}(R^-(y))$ is the attacks on y in strategy P , and $P^{\Leftarrow P \cup O}(R^+(y))$ is the supports on y in strategy P , we can observe the following cases:

- (A) If $|P^{\leftarrow O}(R_i^-(x))| * w(P^{\leftarrow O}(R_i^-(x))) = |P^{\leftarrow O}(R_i^-(y))| * w(P^{\leftarrow O}(R_i^-(y)))$,
 $|P^{\Leftarrow P \cup O}(R_i^+(x))| * w(P^{\Leftarrow P \cup O}(R_i^+(x))) = |P^{\Leftarrow P \cup O}(R_i^+(y))| * w(P^{\Leftarrow P \cup O}(R_i^+(y)))$,
and $\tau_i(x) = \tau_i(y)$, then $\delta_i(x) = \delta_i(y)$
- (B) If $|P^{\leftarrow O}(R_i^-(x))| * w(P^{\leftarrow O}(R_i^-(x))) > |P^{\leftarrow O}(R_i^-(y))| * w(P^{\leftarrow O}(R_i^-(y)))$,
 $|P^{\Leftarrow P \cup O}(R_i^+(x))| * w(P^{\Leftarrow P \cup O}(R_i^+(x))) = |P^{\Leftarrow P \cup O}(R_i^+(y))| * w(P^{\Leftarrow P \cup O}(R_i^+(y)))$,
and $\tau_i(x) = \tau_i(y)$, then $\delta_i(x) \leq \delta_i(y)$
- (C) If $|P^{\leftarrow O}(R_i^-(x))| * w(P^{\leftarrow O}(R_i^-(x))) < |P^{\leftarrow O}(R_i^-(y))| * w(P^{\leftarrow O}(R_i^-(y)))$,
 $|P^{\Leftarrow P \cup O}(R_i^+(x))| * w(P^{\Leftarrow P \cup O}(R_i^+(x))) = |P^{\Leftarrow P \cup O}(R_i^+(y))| * w(P^{\Leftarrow P \cup O}(R_i^+(y)))$,
and $\tau_i(x) = \tau_i(y)$, then $\delta_i(x) \geq \delta_i(y)$

Only case B, with a weaker set of attackers of x than attackers of y , results in a higher strength of x compared to the strength of y . Since the cases exhaust the possibilities, we can conclude that Basic Idea 10 [29] is followed by

the gradual semantics. ■

According to *Basic idea 11* [29]: A stronger set of supporters determines a higher strength. Formally stated as: For an argument α and an argument β , if $\delta(R^-(\alpha)) = \delta(R^-(\beta))$, $\delta(R^+(\alpha)) > \delta(R^+(\beta))$, and $\tau(\alpha) = \tau(\beta)$, then $\delta(\beta) \leq \delta(\alpha)$. In order to show whether this holds for the proposed gradual semantics (Definitions 54 and 55), we specify the following proposition and proof.

Proposition 12 *Let $\gamma = \langle \mathcal{I}, D_r^n, Q_r^n \rangle$ be an information-seeking dialogue system, such that $\mathcal{I} = \{se, re\}$, $se = \langle G, \mathcal{U}_{se}, \delta, TS \rangle$, $re = \langle \mathcal{U}_{re}, BS \rangle$, $Q_r^n = [q_r, \dots, q_n]$ and $q_i = \langle X_i, R_i^-, R_i^+, \tau_i \rangle$, ($r \leq i \leq n$), where $x \in X_i$ is a topic, $y \in X_i$ is a topic, $g \in G$ is a goal topic, δ is a strength function, \mathcal{P} is the set of strategies w.r.t. g of the seeker, and $\mathcal{O} \subseteq 2^{\mathcal{U}_{re}}$ are the strategies of the respondent, $P \in \mathcal{P}$ and $O \in \mathcal{O}$, and $w(X) = \text{agg}(\{\delta(z) \mid z \in X\})$ is a strength aggregation function, where $\text{agg} \in \{\min, \max, \text{average}\}$.*

*For the arguments $x, y \in X_i$, if $|P^{\Leftarrow P \cup O}(R_i^+(x))| * w(P^{\Leftarrow P \cup O}(R_i^+(x))) > |P^{\Leftarrow P \cup O}(R_i^+(y))| * w(P^{\Leftarrow P \cup O}(R_i^+(y)))$, $|P^{\Leftarrow O}(R_i^-(x))| * w(P^{\Leftarrow O}(R_i^-(x))) = |P^{\Leftarrow O}(R_i^-(y))| * w(P^{\Leftarrow O}(R_i^-(y)))$, and $\tau_i(x) = \tau_i(y)$, then $\delta_i(x) \geq \delta_i(y)$.*

Proof: Let x and y be an arguments, $\tau(x)$ and $\tau(y)$ represent their base scores and $\delta(x)$ and $\delta(y)$ represent their strengths. By considering the gradual strength evaluation (Definitions 54 and 55), where $P^{\Leftarrow O}(R^-(x))$ is the attacks on x in strategy P , $P^{\Leftarrow P \cup O}(R^+(x))$ is the supports on x in strategy P , $P^{\Leftarrow O}(R^-(y))$ is the attacks on y in strategy P , and $P^{\Leftarrow P \cup O}(R^+(y))$ is the supports on y in strategy P , we can observe the following cases:

- (A) if $|P^{\Leftarrow P \cup O}(R_i^+(x))| * w(P^{\Leftarrow P \cup O}(R_i^+(x))) = |P^{\Leftarrow P \cup O}(R_i^+(y))| * w(P^{\Leftarrow P \cup O}(R_i^+(y)))$, $|P^{\Leftarrow O}(R_i^-(x))| * w(P^{\Leftarrow O}(R_i^-(x))) = |P^{\Leftarrow O}(R_i^-(y))| * w(P^{\Leftarrow O}(R_i^-(y)))$, and $\tau_i(x) = \tau_i(y)$, then $\delta_i(x) = \delta_i(y)$.
- (B) if $|P^{\Leftarrow P \cup O}(R_i^+(x))| * w(P^{\Leftarrow P \cup O}(R_i^+(x))) < |P^{\Leftarrow P \cup O}(R_i^+(y))| * w(P^{\Leftarrow P \cup O}(R_i^+(y)))$, $|P^{\Leftarrow O}(R_i^-(x))| * w(P^{\Leftarrow O}(R_i^-(x))) = |P^{\Leftarrow O}(R_i^-(y))| * w(P^{\Leftarrow O}(R_i^-(y)))$, and $\tau_i(x) = \tau_i(y)$, then $\delta_i(x) \leq \delta_i(y)$.
- (C) if $|P^{\Leftarrow P \cup O}(R_i^+(x))| * w(P^{\Leftarrow P \cup O}(R_i^+(x))) > |P^{\Leftarrow P \cup O}(R_i^+(y))| * w(P^{\Leftarrow P \cup O}(R_i^+(y)))$, $|P^{\Leftarrow O}(R_i^-(x))| * w(P^{\Leftarrow O}(R_i^-(x))) = |P^{\Leftarrow O}(R_i^-(y))| * w(P^{\Leftarrow O}(R_i^-(y)))$, and $\tau_i(x) = \tau_i(y)$, then $\delta_i(x) \geq \delta_i(y)$.

Only case C, with a stronger set of supporters of x than supporters of y , results in a higher strength of x compared to the strength of y . Since the cases exhaust the possibilities, we can conclude that Basic Idea 11 [29] is followed by the gradual semantics. ■

We will further analyze the process of epistemic topic planning through detailed running examples.

5.5 Example: Health-Assessment Dialogues

In this section, we take a further look at the process of epistemic topic planning through running examples in the setting of health-assessment dialogues.

Health self-management for elderly individuals can become increasingly complicated, due to, e.g., memory issues, mistrust in medication or lack of motivation in healthy activity, which can result in missing or avoiding their prescribed health routines [90]. Let us consider a software assistant (seeker agent) which joins a health-promotion dialogue [140] with an elderly individual (respondent agent) to facilitate sharing of (intimate) health related topics. Loneliness [246] is typically a difficult topic to discuss because it can be emotionally sensitive and can make the patient feel vulnerable. It can also be difficult to identify, as patients may not always express feelings of loneliness directly. In a health promoting dialogue, a nurse may want to know if the patient experiences feelings of isolation or disconnectedness from others. Understanding these factors can help a nurse identify if a patient may be at risk for loneliness and develop strategies to address it [140].

The seeker is designed to conduct a goal-hiding information-seeking dialogue, following dialogue strategies commonly used by primary healthcare nurses to postpone sensitive topics [140]. We assume that the seeker utilizes a knowledge-base with the patient's health history and prior preferences to create base willingness scores for different health topics, and an initial belief-topic dependency graph (see Figure 5.2). We further assume that there are either an ontology/knowledge-graph or sub-symbolic mechanisms for analyzing asserted beliefs in the dialogue.

In this example, the seeker aims to introduce a topic g (loneliness), which currently is assumed to be undesired to talk about by the individual. The seeker acknowledges that the individual may end the dialogue if topic g (loneliness) is opened too soon. Hence, the seeker postpones topic g and starts with topic c (grandchildren), which currently is assumed to be desired. In each state of the dialogue, the seeker computes the individual's willingness for topics and looks for a strategy (set of topics) leading to the goal topic. Other topics the seeker may open are d (health), e (sports), and f (family). In accordance with the protocol in Definition 46, we assume that each dialogue move is unique.

Running Example 1: The seeker reaches the goal

Let $\gamma = \langle \mathcal{I}, D_r^n, Q_r^n \rangle$ be a goal-hiding information-seeking dialogue system, such that $\mathcal{I} = \{se, re\}$ is the participating agents, where $se = \langle G, \mathcal{U}_{se}, \delta, TS \rangle$ is a seeker agent and $re = \langle \mathcal{U}_{re}^B, BS \rangle$ is a respondent agent. There is a set of beliefs $\{b_0, b_1, b_2, b_3, b_4\} \subset \mathcal{U}_{re}^B$, dependent on topics according to Figure 5.2. The sensitivity interval $[\rho^l, \rho^u]$ (Definition 50) has an upper bound of $\rho^u=0.7$ and a lower bound of $\rho^l=0.5$. A discount constant $\beta = 0.5$. To clarify strength change in this example, we consider an aggregation function based on average and rounding down to the nearest multiple of 0.2 (for example, a strength

change of 0.3 would round down to 0.2). The initial dialogue state $D_0^0 = []$ and the initial opponent model $q_0 = \langle \mathcal{U}^T, R_0^-, R_0^+, \tau_0 \rangle$ where $\mathcal{U}^T = \{c, d, e, f, g\}$, $R_0^- = \{\}$, $R_0^+ = \{\}$, i.e., the relations among topics are unknown. Given a set of base scores $\{\tau_0(c)=0.6, \tau_0(d)=0.3, \tau_0(e)=0.4, \tau_0(f)=0.2, \tau_0(g)=0.1\}$, the goal topic g (loneliness) is undesired ($\tau_0(g) < \rho^l$) and topic c (grandchildren) is desired ($\rho^u > \tau_0(c) > \rho^l$).

The dialogue is initiated by the seeker by opening the topic with the highest base score, resulting in dialogue state $D_0^1 = [\langle se, open, c \rangle]$. Topic c (grandchildren) is followed by a sub-dialogue γ' in which the respondent asserts beliefs. For any belief $b \in \mathcal{U}_{re}^B$ that appears in γ' , a call to $dependent_topics^+(b, \mathcal{U}^T)$ is made to find any $t \in \mathcal{U}^T$ related through b . Following c , the respondent asserts the beliefs b_0 and b_1 . The asserted belief b_0 (“sports activity with grandchildren”) is estimated as positively dependent on topic c (grandchildren) and topic e (sports), and the asserted belief b_1 (“lunch with grandchildren”) is estimated as positively dependent on topic c (grandchildren) and topic g (loneliness), then support relations between the topics are generated and appended to an updated opponent model $q_1 = \langle \mathcal{U}^T, R_1^-, R_1^+, \tau_1 \rangle$ where $R_1^- = \{\}$ and $R_1^+ = \{(c, e), (c, g)\}$. After a strength evaluation process, a new set of scores is generated by the gradual semantics (Definition 54 and 55), $\{\delta_1(c)=0.6, \delta_1(d)=0.3, \delta_1(e)=0.6, \delta_1(f)=0.2, \delta_1(g)=0.3\}$, giving dominance to topic e (sports). The seeker closes topic c and opens topic e , resulting in dialogue state $D_0^5 = [\langle se, open, c \rangle, \langle re, assert, b_0 \rangle, \langle re, assert, b_1 \rangle, \langle se, close, c \rangle, \langle se, open, e \rangle]$. In dialogue state D_0^5 , a sub-dialogue γ'' is started where the belief b_2 (“outdoors activity”) is asserted. A call to $dependent_topics^+(b_2, \mathcal{U}^T)$ is made which finds a support relation between topic e (sports) and g (loneliness), appended to a new opponent model $q_2 = \langle \mathcal{U}^T, R_2^-, R_2^+, \tau_2 \rangle$ where $R_2^- = \{\}$ and $R_2^+ = \{(c, e), (c, g), (e, g)\}$. A new set of scores is generated by the gradual semantics (Definition 54 and 55), $\{\delta_2(c)=0.2, \delta_2(d)=0.3, \delta_2(e)=0.4, \delta_2(f)=0.2, \delta_2(g)=0.5\}$, giving dominance to goal topic g due to $\rho^u \geq \delta_2(g) \geq \rho^l$. The seeker closes topic e and opens topic g , resulting in dialogue state $D_0^8 = [\langle se, open, c \rangle, \langle re, assert, b_0 \rangle, \langle re, assert, b_1 \rangle, \langle se, close, c \rangle, \langle se, open, e \rangle, \langle re, assert, b_2 \rangle, \langle se, close, e \rangle, \langle se, open, g \rangle]$. subsequently, a sought-after belief b_3 (“Feelings of isolation”), dependent on goal topic g , is asserted, and the seeker finally closes topic g in dialogue state D_0^10 . In this way, the seeker elicits binary support relations among topics to discover a personalized path to the goal, while avoiding undesired topics.

In this example, illustrated in Figure 5.3 and summarized in Table 5.2 and Table 5.3, we have seen how a software assistant supports an individual to share their (intimate) health information by utilizing epistemic topic planning (increasing the willingness for topic g , without talking about topic g).

Running Example 2: The seeker can not reach the goal

Let us look at an alternative scenario of the previous case where the seeker can not reach the goal. There is a set of beliefs $\{b_0, b_1, b_2, b_3, b_4\} \subset \mathcal{U}_{re}^B$,

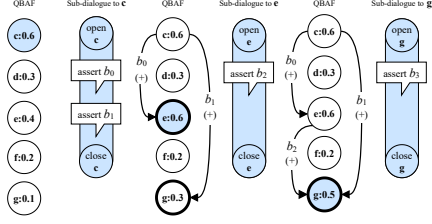


Figure 5.3: Running Example 1: QBAF states before and after asserted beliefs.

dependent on topics according to Figure 5.2. The credibility interval $[\rho^l, \rho^u]$ has an upper bound of $\rho^u=0.7$ and a lower bound of $\rho^l=0.5$. The discount constant is $\beta = 0.5$. An aggregation function based on average and rounding down to the nearest multiple of 0.2. Given an initial dialogue state $D_0^0 = []$, an initial opponent model $q_0 = \langle \mathcal{U}^T, R_0^-, R_0^+, \tau_0 \rangle$ where $\mathcal{U}^T = \{c, d, e, f, g\}$, $R_0^- = \{\}$, $R_0^+ = \{\}$ and a set of base scores $\{\tau_0(c)=0.5, \tau_0(d)=0.6, \tau_0(e)=0.3, \tau_0(f)=0.4, \tau_0(g)=0.2\}$, the goal topic g (dietary habits) is undesired ($\tau_0(g) \downarrow \rho^l$) and topic d (health) is desired ($\tau_0(d) \uparrow \rho^l$). The dialogue is initiated by the seeker by opening the topic with the highest base score, resulting in a dialogue state $D_0^1 = [\langle se, open, d \rangle]$. Topic d (health) is followed by a sub-dialogue γ^1 in which the respondent asserts beliefs. Suppose that the belief b_4 (“stomach problems”) is asserted and negatively dependent on topic d (health) and negatively dependent on topic g (dietary habits). This results in an attack relation between the topics which is appended to an updated opponent model $q_1 = \langle \mathcal{U}^T, R_1^-, R_1^+, \tau_1 \rangle$ where $R_1^- = \{(d, g)\}$ and $R_1^+ = \{\}$. After a strength evaluation process, a new set of scores is generated, $\{\delta_1(c)=0.5, \delta_1(d)=0.6, \delta_1(e)=0.3, \delta_1(f)=0.4, \delta_1(g)=0.0\}$, giving dominance to topic c (grandchildren). The seeker closes topic d and opens topic c , resulting in dialogue state $D_0^4 = [\langle se, open, d \rangle, \langle re, assert, b_4 \rangle, \langle se, close, d \rangle, \langle se, open, c \rangle]$. In a proceeding sub-dialogue to topic c , the belief b_0 (“sport activity with grandchildren”) is asserted. A call to $dependent_topics^+(b_0, \mathcal{U}^T)$ is made which finds a support relation between topic c (grandchildren) and topic e (sports), appended to a new opponent model $q_2 = \langle \mathcal{U}^T, R_2^-, R_2^+, \tau_2 \rangle$ where $R_2^- = \{(d, g)\}$ and $R_2^+ = \{(c, e)\}$. A new set of scores is generated, $\{\delta_2(c)=0.5, \delta_2(d)=0.6, \delta_2(e)=0.5, \delta_2(f)=0.4, \delta_2(g)=0.0\}$, giving dominance to topic e . The seeker closes topic c and opens topic e , resulting in dialogue state $D_0^7 = [\langle se, open, d \rangle, \langle re, assert, b_4 \rangle, \langle se, close, d \rangle, \langle se, open, c \rangle, \langle re, assert, b_0 \rangle, \langle se, close, c \rangle, \langle se, open, e \rangle]$. In a proceeding sub-dialogue, belief b_2 (“outdoors activity”) is asserted, positively dependent on topic e (sports) and topic g (dietary habits), generating a support relation (e, g) . This results in a new opponent model $q_3 = \langle \mathcal{U}^T, R_3^-, R_3^+, \tau_3 \rangle$ where $R_3^- = \{(d, g)\}$ and $R_3^+ = \{(c, e), (e, g)\}$. After a strength evaluation process, a new set of scores is generated $\{\delta_3(c)=0.5, \delta_3(d)=0.6, \delta_3(e)=0.5, \delta_3(f)=0.4, \delta_3(g)=0.2\}$. With no more desired topics to choose from, the seeker closes topic e and the dialogue ends on dialogue state $D_0^9 = [\langle se, open, d \rangle, \langle re, assert, b_4 \rangle, \langle se, close, d \rangle, \langle se, open, c \rangle,$

$\langle re, assert, b_0 \rangle, \langle se, close, c \rangle, \langle se, open, e \rangle, \langle re, assert, b_2 \rangle, \langle se, close, e \rangle]$.

In this scenario, illustrated in Figure 5.4 and summarized in Table 5.4 and Table 5.5, the software assistant cannot find a path to the goal topic g since the willingness for g does not reach the sensitivity interval (as $\delta_3(g) \not\geq \rho^l$). Therefore, the undesired goal topic g is not revealed, preserving a personalized dialogue.

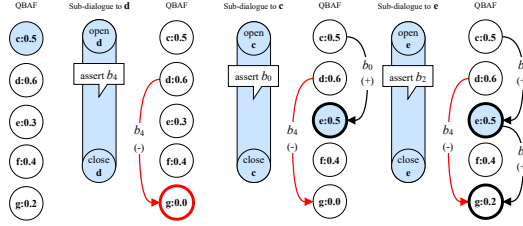


Figure 5.4: Running Example 2: QBAF states before and after asserted beliefs.

We can observe that the number of positive dependencies in the belief-topic dependency graph is critical for finding a path to a goal topic. Without a sufficiently connected QBAF, with support between topics, the seeker can get stuck at dead ends, with no further options to increase willingness for topics. Having a topic with a high willingness to start a dialogue is important in order to spark the assertion of beliefs. In everyday human dialogues, such patterns are commonly observed; A speaker may start by talking about general topics suitable for a context before engaging in more personal topics.

Table 5.2: Running example 1: The seeker reaches the goal. Goal-hiding dialogue process with evolving willingness based on asserted beliefs. Topics= $\{c,d,e,f,g\}$. Goals= $\{g\}$. Beliefs= $\{b_0,b_1,b_2,b_3,b_4\}$. $\rho^l=0.5$. $\rho^u=0.7$.

Dialogue	Belief-Topic dep.	Sup/Att	$\delta(c)$	$\delta(d)$	$\delta(e)$	$\delta(f)$	$\delta(g)$
\square		$+\{\}, -\{\}$	0.6	0.3	0.4	0.2	0.1
$\langle se, open, c \rangle$		$+\{\}, -\{\}$	0.6	0.3	0.4	0.2	0.1
$\langle se, open, c \rangle,$ $\langle re, assert, b_0 \rangle,$ $\langle re, assert, b_1 \rangle$	dep_topics^+ $(b_0, \mathcal{U}^T)=\{c, e\}$ dep_topics^+ $(b_1, \mathcal{U}^T)=\{c, g\}$	$+\{(c,e),$ $(c,g)\}, -\{\}$	0.6	0.3	0.6	0.2	0.3
$\langle se, open, c \rangle,$ $\langle re, assert, b_0 \rangle,$ $\langle re, assert, b_1 \rangle,$ $\langle se, close, c \rangle,$ $\langle se, open, e \rangle$		$+\{(c,e),$ $(c,g)\}, -\{\}$	0.6	0.3	0.6	0.2	0.3
$\langle se, open, c \rangle,$ $\langle re, assert, b_0 \rangle,$ $\langle re, assert, b_1 \rangle,$ $\langle se, close, c \rangle,$ $\langle se, open, e \rangle,$ $\langle re, assert, b_2 \rangle$	dep_topics^+ $(b_2, \mathcal{U}^T)=\{e, g\}$	$+\{(c,e), (c,g),$ $(e,g)\}, -\{\}$	0.2	0.3	0.4	0.2	0.5
$\langle se, open, c \rangle,$ $\langle re, assert, b_0 \rangle,$ $\langle re, assert, b_1 \rangle,$ $\langle se, close, c \rangle,$ $\langle se, open, e \rangle,$ $\langle re, assert, b_2 \rangle,$ $\langle se, close, e \rangle,$ $\langle se, open, g \rangle$		$+\{(c,e), (c,g),$ $(e,g)\}, -\{\}$	0.2	0.3	0.4	0.2	0.5
$\langle se, open, c \rangle,$ $\langle re, assert, b_0 \rangle,$ $\langle re, assert, b_1 \rangle,$ $\langle se, close, c \rangle,$ $\langle se, open, e \rangle,$ $\langle re, assert, b_2 \rangle,$ $\langle se, close, e \rangle,$ $\langle se, open, g \rangle,$ $\langle re, assert, b_3 \rangle,$ $\langle se, close, g \rangle$	dep_topics^+ $(b_3, \mathcal{U}^T)=\{g\}$	$+\{(c,e), (c,g),$ $(e,g)\}, -\{\}$	0.2	0.3	0.4	0.2	0.5

Table 5.3: Running example 1: Goal-hiding strategy generation. Topics= $\{c,d,e,f,g\}$. Goals= $\{g\}$.

Topics Opened and Beliefs Asserted	<i>se</i> -Strategies	<i>re</i> -Strategies
\square	$\{c,g\}, \{d,g\}, \{e,g\}, \{c,d,g\}, \dots$	$2\mathcal{U}^{re_0}$
$\langle c \rangle$	$\{c,g\}, \{c,d,g\}, \dots$	$2\mathcal{U}^{re_1}$
$\langle c, b_0, b_1 \rangle$	$\{c,e,g\}, \{c,e,d,g\}, \dots$	$2\mathcal{U}^{re_2}$
$\langle c, b_0, b_1, e \rangle$	$\{c,e,g\}, \{c,e,d,g\}, \dots$	$2\mathcal{U}^{re_3}$
$\langle c, b_0, b_1, e, b_2 \rangle$	$\{c,e,g\}, \{c,e,d,g\}, \dots$	$2\mathcal{U}^{re_4}$
$\langle c, b_0, b_1, e, b_2, g \rangle$	$\{c,e,g\}$	$2\mathcal{U}^{re_5}$
$\langle c, b_0, b_1, e, b_2, g, b_3 \rangle$	$\{c,e,g\}$	$2\mathcal{U}^{re_6}$

Table 5.4: Running example 2: The seeker can not reach the goal. Goal-hiding dialogue process with evolving willingness based on asserted beliefs. Topics= $\{c,d,e,f,g\}$. Goals= $\{g\}$. Beliefs= $\{b_0,b_1,b_2,b_3,b_4\}$. $\rho^l=0.5$. $\rho^u=0.7$.

Dialogue	Belief-Topic dep.	Sup/Att	$\delta(c)$	$\delta(d)$	$\delta(e)$	$\delta(f)$	$\delta(g)$
\square		$+\{ \}, -\{ \}$	0.5	0.6	0.3	0.4	0.2
$\langle se, open, d \rangle$		$+\{ \}, -\{ \}$	0.5	0.6	0.3	0.4	0.2
$\langle se, open, d \rangle, \langle re, assert, b_4 \rangle$	dep_topics^- $(b_4, \mathcal{U}^T)=\{d, g\}$	$+\{ \}, -\{(d,g)\}$	0.5	0.6	0.3	0.4	0.0
$\langle se, open, d \rangle, \langle re, assert, b_4 \rangle, \langle se, close, d \rangle, \langle se, open, c \rangle$		$+\{ \}, -\{(d,g)\}$	0.5	0.6	0.3	0.4	0.0
$\langle se, open, d \rangle, \langle re, assert, b_4 \rangle, \langle se, open, c \rangle, \langle re, assert, b_0 \rangle$	dep_topics^+ $(b_0, \mathcal{U}^T)=\{c, e\}$	$+\{(c,e)\}, -\{(d,g)\}$	0.5	0.6	0.5	0.4	0.0
$\langle se, open, d \rangle, \langle re, assert, b_4 \rangle, \langle se, close, d \rangle, \langle se, open, c \rangle, \langle re, assert, b_0 \rangle, \langle se, close, c \rangle, \langle se, open, e \rangle$		$+\{(c,e)\}, -\{(d,g)\}$	0.5	0.6	0.5	0.4	0.0
$\langle se, open, d \rangle, \langle re, assert, b_4 \rangle, \langle se, close, d \rangle, \langle se, open, c \rangle, \langle re, assert, b_0 \rangle, \langle se, close, c \rangle, \langle se, open, e \rangle, \langle re, assert, b_2 \rangle$	dep_topics^+ $(b_2, \mathcal{U}^T)=\{e, g\}$	$+\{(c,e), (e,g)\}, -\{(d,g)\}$	0.5	0.6	0.5	0.4	0.2
$\langle se, open, d \rangle, \langle re, assert, b_4 \rangle, \langle se, close, d \rangle, \langle se, open, c \rangle, \langle re, assert, b_0 \rangle, \langle se, close, c \rangle, \langle se, open, e \rangle, \langle re, assert, b_2 \rangle, \langle se, close, e \rangle$		$+\{(c,e), (e,g)\}, -\{(d,g)\}$	0.5	0.6	0.5	0.4	0.2

Table 5.5: Running example 2: Goal-hiding strategy generation. Topics= $\{c,d,e,f,g\}$. Goals= $\{g\}$.

Topics Opened and Beliefs Asserted	se-Strategies	re-Strategies
\square	$\{d,g\}, \{c,d,g\}, \{d,e,g\}, \dots$	$2^{U_{re0}}$
$\{d\}$	$\{d,g\}, \{d,e,g\}, \{d,c,g\}, \dots$	$2^{U_{re1}}$
$\{d,b_4\}$	$\{d,c,g\}, \{d,e,g\}, \dots$	$2^{U_{re2}}$
$\{d,b_4,c\}$	$\{d,c,e,g\}, \{d,c,e,d,g\}, \dots$	$2^{U_{re3}}$
$\{d,b_4,c,b_0\}$	$\{d,c,e,g\}, \{d,c,e,d,g\}, \dots$	$2^{U_{re4}}$
$\{d,b_4,c,b_0,e\}$	$\{d,c,e,g\}, \{d,c,e,d,g\}, \dots$	$2^{U_{re5}}$
$\{d,b_4,c,b_0,e,b_2\}$	$\{d,c,e,g\}$	$2^{U_{re6}}$

5.6 Example: Verifying Social Engineering

The goal-hiding dialogues framework can be applied to recognize and verify goal-hiding strategies. Here, we model a case of a non-collaborative dialogue in the setting of social engineering and phishing. We introduce the case and proceed with a detailed analysis on this use-case.

Phishing case analysis in terms of GHD

Phishing, a type of social engineering attack [302], is one of the most prevalent forms of cyberattacks, characterized by attackers attempting to manipulate individuals into disclosing sensitive information or performing specific actions, such as clicking on malicious links or transferring funds. As system security measures become increasingly robust, attackers are shifting their focus to exploiting human vulnerabilities, leveraging trust, urgency, or curiosity to achieve their objectives [133, 281]. Phishing has evolved beyond its traditional forms, now encompassing a range of attack vectors such as SMS phishing, voice phishing, and social media phishing. These attacks leverage increasingly sophisticated and context-aware strategies to deceive users [151, 257].

Various methods have been explored to detect phishing across different mediums. Works, such as [256, 257] applied Natural Language Processing (NLP) techniques to analyze linguistic patterns in phishing emails, identifying anomalies such as misspellings, urgency cues, and impersonation attempts. [151] and [14] investigated speech analysis techniques to detect manipulation in voice phishing. [19] applied pattern matching and heuristics to identify malicious SMS messages. In social media phishing detection, [263] and [136] leveraged machine learning to analyze suspicious behaviors, such as sudden friend requests or unusual messages. [158] and [268] explored hybrid systems that integrate NLP and deep learning for phishing detection across multiple platforms. Despite these advancements, attackers continuously adapt, making detection an ongoing challenge [281, 163].

A major limitation in these prior approaches, typically based on machine learning methods, is their lack of transparency, making it difficult to interpret how decisions—such as identifying whether a message is deceptive, determining when manipulation occurs, or recognizing key indicators of phishing—are made. This opacity hampers efforts to trace social engineering tactics and identify critical evidence. Since phishing and social engineering typically unfold over time, recognizing key *turning points* can provide valuable insights for criminal investigations, refining detection models, and improving human training against manipulation tactics [274, 190]. However, pinpointing these moments remains challenging, as they emerge gradually within the progression of the conversation.

Mouton et al. [209] defined social engineering attacks (SEA) as multi-stage processes with six steps: (1) *attack formulation*, (2) *information gathering*, (3) *crafting attack vectors*, (4) *developing a relationship*, (5) *exploiting the relation-*

ship to manipulate the target, and (6) debriefing for concealing the attack. Steps 4–6 often unfold within a single interaction, forming a process where attackers guide conversations to shape beliefs and increase compliance.

This multi-stage strategic approach can be modeled as a goal-hiding dialogue.

Running Example 3: Social Engineering Attack

In the following example, topics are inspired by real-world social engineering examples illustrated in [209]. We then present a detailed example of the willingness strength calculation at a critical moment in the scenario, what we refer to as a *turning point*, where the respondent’s willingness toward the seeker’s goal topic undergoes a significant shift.

In this scenario, representing a real-world social engineering attack, a seeker (*se*) targets an administrative assistant (*re*) in a multinational organization to extract sensitive invoice information. The dialogue begins with general company procedures as a non-threatening entry point. The attacker then introduces approval steps to establish rapport and transition to management authority and escalation processes. Gradually, details about typical invoice handling emerge, increasing the assistant’s willingness to discuss verification steps. This leads to the final stage, where the attacker elicits information about a specific invoice.

Let $\gamma = \langle \mathcal{I}, D_r^n, Q_r^n \rangle$ be a goal-hiding information-seeking dialogue system, where $\mathcal{I} = \{se, re\}$ are the participating agents, with $se = \langle G, \mathcal{U}_{se}, \delta \rangle$ being the seeker agent and $re = \langle \mathcal{U}_{re}^B \rangle$ the respondent agent. The dialogue is structured to guide the respondent towards the goal topic g by eliciting information and forming support relations among topics. The set of topics is $\mathcal{U}^T = \{c, d, e, g\}$:

- c : “company procedures” (e.g., document approval processes).
- d : “management authority” (e.g., policies on handling directives).
- e : “invoice handling” (e.g., processing invoice requests).
- g : “opening a malicious invoice link” (sought-after action).

The respondent’s set of beliefs $\mathcal{U}_{re}^B = \{b_0, b_1, b_2, b_3, b_4\}$ are related to topics according to a belief-topic dependency graph (see Figure 5.5):

- b_0 : “familiarity with approval steps”, positively related to c and d .
- b_1 : “handling escalations from management”, positively related to d and e .
- b_2 : “routine procedures for invoice verification”, positively related to e and g .
- b_3 : “details about a specific invoice”, positively related to g (goal topic).
- b_4 : “management credibility”, positively related to d and g .

The sensitivity interval $[\rho^l, \rho^u]$ has bounds $\rho^l = 0.5$ and $\rho^u = 0.7$, and a discount constant $\beta = 0.5$. A strength aggregation function w based on *average* is used, and the resulting strengths are rounded down to the nearest multiple of 0.2. (following the presentation of this scenario, we give a detailed example of the strength calculation). The initial dialogue state is $D_0^0 = []$, and the initial opponent model $q_0 = \langle \mathcal{U}^T, R_0^-, R_0^+, \tau_0 \rangle$, where:

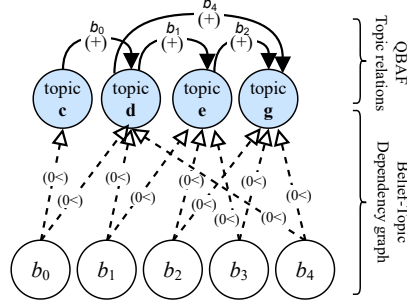


Figure 5.5: Belief-Topic Dependency Graph and inferred QBAF relations

- $\mathcal{U}^T = \{c, d, e, g\}$, $R_0^- = \{\}$, $R_0^+ = \{\}$.
- Initial topic strengths: $\tau_0(c) = 0.6$, $\tau_0(d) = 0.4$, $\tau_0(e) = 0.3$, $\tau_0(g) = 0.1$.

The dialogue (see Table 5.6) progresses as follows:

1. The seeker opens topic c (company procedures), resulting in: $D_0^1 = [\langle se, open, c \rangle]$.
2. The respondent asserts cues for the belief b_0 (“familiarity with approval steps”), positively related to topics c and d . A support relation is added, resulting in $R_1^+ = \{(c, d)\}$, and the opponent model is updated to $q_1 = \langle \mathcal{U}^T, R_1^-, R_1^+, \tau_1 \rangle$, $R_1^- = \{\}$, $R_1^+ = \{(c, d)\}$.
3. Topic strengths are updated: $\tau_1(c) = 0.6$, $\tau_1(d) = 0.6$, $\tau_1(e) = 0.3$, $\tau_1(g) = 0.1$.
4. The seeker closes c and opens d (management authority), resulting in:

$$D_0^4 = [\langle se, open, c \rangle, \langle re, assert, b_0 \rangle, \langle se, close, c \rangle, \langle se, open, d \rangle].$$

5. The respondent asserts cues for the belief b_1 (“handling escalations from management”), positively related to topics d and e . The respondent also asserts cues for the belief b_4 (“management credibility”), positively related to d and g . Two new support relations are added, resulting in $R_2^+ = \{(c, d), (d, e), (d, g)\}$, and the opponent model is updated to $q_2 = \langle \mathcal{U}^T, R_2^-, R_2^+, \tau_2 \rangle$, $R_2^- = \{\}$, $R_2^+ = \{(c, d), (d, e), (d, g)\}$.
6. Topic strengths are updated: $\tau_2(c) = 0.6$, $\tau_2(d) = 0.6$, $\tau_2(e) = 0.5$, $\tau_2(g) = 0.3$.
7. The seeker closes d and opens e (invoice handling);
 $D_0^8 = [\dots, \langle re, assert, b_1 \rangle, \langle re, assert, b_4 \rangle, \langle se, close, d \rangle, \langle se, open, e \rangle]$.
8. The respondent asserts cues for the belief b_2 (“routine procedures for invoice verification”), positively related to topics e and g . A new support relation is added, resulting in $R_3^+ = \{(c, d), (d, e), (d, g), (e, g)\}$, and the opponent model is updated to $q_3 = \langle \mathcal{U}^T, R_3^-, R_3^+, \tau_3 \rangle$, $R_3^- = \{\}$, $R_3^+ = \{(c, d), (d, e), (d, g), (e, g)\}$.
9. Topic strengths are updated: $\tau_3(c) = 0.6$, $\tau_3(d) = 0.6$, $\tau_3(e) = 0.5$, $\tau_3(g) = 0.6$.
10. The seeker closes e and opens g (goal topic), eliciting belief b_3 (“details about a specific invoice”). Goal satisfaction is reached and the final dialogue state is:

$$D_0^{13} = [\dots, \langle se, close, e \rangle, \langle se, open, g \rangle, \langle re, assert, b_3 \rangle, \langle se, close, g \rangle].$$

Table 5.6: Social Engineering Example: Topics= $\{c,d,e,g\}$. Goals= $\{g\}$. Beliefs= $\{b_0,b_1,b_2,b_3,b_4\}$. $\rho^l=0.5$. $\rho^u=0.7$. A turning point is at Step 9.

Steps	Dialogue	R^+, R^-	$\delta(c)$	$\delta(d)$	$\delta(e)$	$\delta(g)$
0	\emptyset	$+\{\}, -\{\}$	0.6	0.4	0.3	0.1
1	$\langle se, open, c \rangle$	$+\{\}, -\{\}$	0.6	0.4	0.3	0.1
2	$\langle re, assert, b_0 \rangle$	$+\{(c, d)\}, -\{\}$	0.6	0.6	0.3	0.1
3	$\langle se, close, c \rangle$	$+\{(c, d)\}, -\{\}$	0.6	0.6	0.3	0.1
4	$\langle se, open, d \rangle$	$+\{(c, d)\}, -\{\}$	0.6	0.6	0.3	0.1
5	$\langle re, assert, b_1 \rangle$	$+\{(c, d), (d, e)\}, -\{\}$	0.6	0.6	0.5	0.1
6	$\langle re, assert, b_4 \rangle$	$+\{(c, d), (d, e), (d, g)\}, -\{\}$	0.6	0.6	0.5	0.3
7	$\langle se, close, d \rangle$	$+\{(c, d), (d, e), (d, g)\}, -\{\}$	0.6	0.6	0.5	0.3
8	$\langle se, open, e \rangle$	$+\{(c, d), (d, e), (d, g)\}, -\{\}$	0.6	0.6	0.5	0.3
9	$\langle re, assert, b_2 \rangle$	$+\{(c, d), (d, e), (d, g), (e, g)\}, -\{\}$	0.6	0.6	0.5	0.6
10	$\langle se, close, e \rangle$	$+\{(c, d), (d, e), (d, g), (e, g)\}, -\{\}$	0.6	0.6	0.5	0.6
11	$\langle se, open, g \rangle$	$+\{(c, d), (d, e), (d, g), (e, g)\}, -\{\}$	0.6	0.6	0.5	0.6
12	$\langle re, assert, b_3 \rangle$	$+\{(c, d), (d, e), (d, g), (e, g)\}, -\{\}$	0.6	0.6	0.5	0.6
13	$\langle se, close, g \rangle$	$+\{(c, d), (d, e), (d, g), (e, g)\}, -\{\}$	0.6	0.6	0.5	0.6

Turning point evaluation

To give a better intuition of the willingness strength evaluation of topics, we show as an example the strength change of topic g in the transition between dialogue state 8 and 9. This transition is particularly interesting, as it concerns a *turning point* where the willingness score 0.6 reaches the sensitivity interval $[0.5, 0.7]$, and is opened, suggesting a tactful open move.

Recall that, in dialogue state 5 and 9, the respondent asserted b_4 and b_2 , respectively, resulting in the relations $R_9^+ = \{(e, g), (d, g)\}$ and $R_9^- = \{\}$ inferred in state 9. We apply the gradual strength evaluation (Section 5.3) to determine the willingness strength change of g in state transition 8 to 9.

In this example, to avoid the need for presenting the strength calculation of d and e in detail, we consider their already calculated strengths and exclude the supports (c, d) and (d, e) from prior dialogue states. With this adjustment, the seeker's strategy is represented by the permutation $\pi = \langle c, d, e, g \rangle$, where $\pi \in \Pi(P)$ and $P \in \mathcal{P}$, with $P = \{c, d, e, g\}$, and the topics that either attack or support P (by omitting relations to d and e) are determined as $O = \{d, e\}$.

We then evaluate the strength of g by using the formula $\phi_g(g, P, O)$:

$$\beta * |P^{\leftarrow O}(R_9^-)| * w(P^{\leftarrow O}(R_9^-)) - |P^{\leftarrow P \cup O}(R_9^+)| * w(P^{\leftarrow P \cup O}(R_9^+))$$

Here, $P^{\leftarrow O}(R_9^-) = \emptyset$, with $|P^{\leftarrow O}(R_9^-)| = 0$ since there are no attack relations, and $P^{\leftarrow P \cup O}(R_9^+) = \{e, d\}$, with $|P^{\leftarrow P \cup O}(R_9^+)| = 2$. In this example, we use an aggregation function w based on average, giving the average strengths of $w(P^{\leftarrow O}(R_9^-)) = 0$ and $w(P^{\leftarrow P \cup O}(R_9^+)) = \frac{0.5+0.6}{2} = 0.55$. By considering a discount constant $\beta = 0.5$, we calculate the strength change, sc :

$$sc := 0.5 * (0 * 0 - 2 * 0.55)$$

$$sc := 0.5 * (0 - 1.1)$$

$$sc := 0.5 * (-1.1)$$

$$sc := -0.55$$

We must then check the conditions of **(Case E1)** and **(Case E2)**:

- **(Case E1)**: If attacks are stronger than supports, then subtract sc from $\tau(g)$.
- **(Case E2)**: If supports are stronger than attacks, then add sc to $\tau(g)$.

Since there are no attack relations, Case E2 applies, and we add $|sc|$ to $\tau(g)$:

$$\begin{aligned}\delta_9(g) &= \tau(g) + |sc| \\ &= 0.1 + 0.55 \\ &= 0.65\end{aligned}$$

Thus, considering the base score of g ($\tau(g) = 0.1$), and all relations on g , the willingness strength of g increases from 0.3 (in state 8) to 0.65 (in state 9), rounded down to the closest multiple of 0.2, resulting in 0.6, indicating that the respondent’s willingness to discuss invoice verification has increased significantly and is within the sensitivity interval $[0.5, 0.7]$.

In this section, we have formally analyzed the goal-hiding dialogue process to understand the requirements for ensuring compliance with the formal properties. In the upcoming section, we present an experiment featuring a prototype of the formal framework, analyzing a synthetic dataset. The primary objective of this experiment is to analyze how different dialogues in the dataset preserve the introduced formal properties.

5.7 Experimental Evaluation

This section introduces an empirical assessment of the goal-hiding dialogue framework, complementing the formal analysis (Section 5.4). Addressing research question RQ5.2, the purpose of the experiment is to further evaluate the formal properties of the GHD framework. We begin by evaluating the framework’s capability to filter dialogues by considering the sensitivity interval at each interaction state. This will demonstrate the framework’s capability, utilizing the sensitivity interval, to determine whether the goal topic can be opened or if the dialogue should terminate ‘unsuccessfully’. We then compare these results with a dataset of random dialogues that do not adhere to GHD-based constraints. The evaluation further aims to evaluate strength monotonicity and bipolar beliefs, to determine their requirement for modeling willingness and maintaining a focused dialogue.

Hence, the aims are as follows:

- **Evaluate Sensitivity Interval:** Assess the GHD framework’s effectiveness in filtering dialogues according to strength monotonicity and sensitivity interval w.r.t. a goal topic. This is done by considering a dataset of random dialogue sequences.

- **Evaluate Strength Monotonicity:** Examining the impact of upholding versus disregarding strength monotonicity on dialogue behavior, considering positive beliefs in the former and both positive and negative beliefs in the latter.
- **Evaluate Bipolar versus Unipolar Beliefs:** Comparing the effects of bipolar (positive and negative) beliefs versus unipolar (only positive) beliefs on GHD dialogue progression and outcome.

To facilitate this evaluation, we have developed a prototype of the information-seeking dialogue system. This prototype comprises components capable of handling dialogue sequences and topic selection. Consistent with the formal framework, the prototype’s topic selection is supported by an incrementally updated QBAF and a belief-topic dependency graph. The experiment utilizes a dataset of dialogues (each containing sequences of topics and asserted beliefs), generated by considering a finite set of topics and a finite set of beliefs. By further considering the dependencies of a particular belief-topic dependency graph, a dialogue space is defined (see Definition 57). For each dialogue in the dataset, we assess its adherence to the constraints of a goal-hiding dialogue, in particular strength monotonicity (Definition 49) and sensitivity interval (Definition 50). In this evaluation, we set the sensitivity interval to $[0.7, 0.5]$, where the upper bound is 0.7 and the lower bound is 0.5. This sensitivity interval guided the creation of an initial QBAF, in particular base scores of arguments, used for the evaluation.

Dataset

The dataset, consisting of 100 sequences of topics and beliefs (see a sample in Listing 1), corresponds to a subset of the possible dialogues that can be constructed from a finite set of topics and a finite set of beliefs within a particular belief-topic dependency graph \mathcal{G} (see Listing 2) and an initial QBAF (see Listing 3). Each dialogue in the set starts with the same initial topic and ends with the same goal topic. The interim elements in each dialogue consist of a random sequence of topics, where each topic is followed by one or more beliefs. We restrict the sequences so that beliefs can only follow topics they are related to according to the graph \mathcal{G} . We call the set of topics the dialogue space of \mathcal{G} .

Definition 57 (Dialogue Space) *Given a set of topics \mathcal{U}^T , and a set of beliefs \mathcal{U}^B . Let \mathcal{G}_{BT} denote the space of all the possible belief-topic dependency graphs. Given $\mathcal{G} \in \mathcal{G}_{BT}$, there is a set of dialogue trajectories $\mathcal{T}_{\mathcal{G}}$ w.r.t. \mathcal{G} , an initial topic $t_r \in \mathcal{U}^T$ and a goal topic $t_y \in \mathcal{U}^T$ such that if $dt \in \mathcal{T}_{\mathcal{G}}$ then $dt = [t_r, \dots, t_y]$. For any dialogue trajectory $dt \in \mathcal{T}_{\mathcal{G}}$, there is a dialogue D_r^y . The dialogues of $\mathcal{T}_{\mathcal{G}}$ is called the dialogue space of \mathcal{G} , denoted $DS(\mathcal{G})$.*

For the purpose of the evaluation, we consider a particular belief-topic dependency graph consisting of five topics, labeled as t_0, t_1, t_2, t_3 , and t_4 , along with sixteen positive beliefs, one from each topic to all other topics, labeled as b_0 to

```

t0, b0, b3, t1, b5, b7, t4
t0, b1, b3, t2, b9, b11, t4
t0, b2, b3, t3, b13, b15, t4
t0, b0, bneg0, t1, b5, b7, t2, b10, b11, t3, b12, b15, t4
t0, b0, b3, t1, b5, b6, bneg1, t2, b10, b11, t3, b13, b15, t4
t0, b0, t1, b5, b7, t2, b9, b11, t4
t0, b2, t3, b13, b15, t1, b5, t2, b8, b11, t4
t0, b0, t1, b5, t2, b10, b11, t3, b12, b15, t4
t0, b0, t1, b5, b7, t2, b10, b11, t4
t0, b0, t1, b4, t2, b8, t3, b12, t4
t0, b0, t1, b5, t2, b8, t3, b12, t4
t0, b0, t1, b4, t2, b8, t3, b12, t4
t0, b0, t1, b4, t2, b8, t3, b12, t4
t0, b0, t1, b4, t3, b12, t2, b8, t4
t0, b0, t1, b5, t3, b12, t2, b8, t4
t0, b0, t1, b4, t3, b12, t2, b8, t4
t0, b0, t1, b4, t3, b12, t2, b8, t4
t0, b0, t2, b8, t1, b4, t3, b12, t4
t0, b0, t2, b8, t1, b5, t3, b12, t4
t0, b0, t2, b8, t1, b4, t3, b12, t4
t0, b0, t2, b8, t1, b4, t3, b12, t4
t0, b0, t2, b8, t3, b12, t1, b4, t4
t0, b0, t2, b8, t3, b12, t1, b5, t4
t0, b0, t2, b8, t3, b12, t1, b4, t4
t0, b0, t2, b8, t3, b12, t1, b4, t4
t0, b0, t3, b12, t1, b4, t2, b8, t4
t0, b0, t3, b12, t1, b5, t2, b8, t4
t0, b0, t3, b12, t1, b4, t2, b8, t4
t0, b0, t3, b12, t1, b4, t2, b8, t4
t0, b0, t3, b12, t2, b8, t1, b4, t4
t0, b0, t3, b12, t2, b8, t1, b5, t4
t0, b0, t3, b12, t2, b8, t1, b4, t4
t0, b0, t3, b12, t2, b8, t1, b4, t4
...
100 rows

```

Listing 1: Dataset Sample: Sequences of topics t_T , $T \in \{0, \dots, 4\}$, and beliefs b_B , $B \in \{0, \dots, 15\} \cup \{neg0, \dots, neg4\}$.

b_{15} , and five negative beliefs, one from each topic to the goal topic, labeled as $bneg0$ to $bneg4$. This specific graph is designed with dependencies between all pairs of topics in order to create a balanced distribution of dialogue paths that can emerge from an initial topic t_0 to a goal topic t_4 (see Listing 2).

We further consider a particular initial QBAF, consisting of five arguments, labeled t_0 , t_1 , t_2 , t_3 , and t_4 , each corresponding to a topic defined in the belief-topic dependency graph. Each argument is assigned an initial base score intended to establish a balanced starting point among intermediate topics. Specifically, t_0 , representing the initial topic, holds a base score of 0.6, exceeding the sensitivity interval. Meanwhile, the intermediate topics t_1 , t_2 , and t_3 are all set at a base score of 0.3, falling below the sensitivity interval, and the goal

```

{
  "beliefs": [
    "b0", "b1", "b2", "b3",
    "b4", "b5", "b6", "b7",
    "b8", "b9", "b10", "b11",
    "b12", "b13", "b14", "b15", "b16",
    "bneg0", "bneg1", "bneg2", "bneg3", "bneg4"
  ],
  "topics": ["t0", "t1", "t2", "t3", "t4"],
  "edges": {
    "b0": {"t0": 1, "t1": 1},
    "b1": {"t0": 1, "t2": 1},
    "b2": {"t0": 1, "t3": 1},
    "b3": {"t0": 1, "t4": 1},
    "b4": {"t1": 1, "t0": 1},
    "b5": {"t1": 1, "t2": 1},
    "b6": {"t1": 1, "t3": 1},
    "b7": {"t1": 1, "t4": 1},
    "b8": {"t2": 1, "t0": 1},
    "b9": {"t2": 1, "t1": 1},
    "b10": {"t2": 1, "t3": 1},
    "b11": {"t2": 1, "t4": 1},
    "b12": {"t3": 1, "t0": 1},
    "b13": {"t3": 1, "t1": 1},
    "b14": {"t3": 1, "t2": 1},
    "b15": {"t3": 1, "t4": 1},
    "b16": {"t4": 1, "t4": 1},
    "bneg0": {"t0": -1, "t4": -1},
    "bneg1": {"t1": -1, "t4": -1},
    "bneg2": {"t2": -1, "t4": -1},
    "bneg3": {"t3": -1, "t4": -1},
    "bneg4": {"t4": -1, "t4": -1}
  }
}

```

Listing 2: Belief-Topic Dependency Graph #2; JSON encoding.

topic $t4$ holds a base score of 0.2, also below the sensitivity interval. The initial QBAF lacks attack and support relations between topics, as no beliefs have been asserted at the initial dialogue state (see Listing 3).

Implementation

A prototype of the formal framework is implemented in the form of a Web application, shared as open-source material¹. The prototype consists of two interacting parts: a goal-hiding reasoning engine (backend) and an interactive question-and-answer application (frontend) directed by the reasoning engine. In this study, we solely utilize the reasoning engine by analyzing synthetic data, and leave it for future work to conduct user evaluations with the provided

¹<https://github.com/AndreasbCS/goal-hiding-dialogues-framework/>

```

{
  "arguments": ["t0", "t1", "t2", "t3", "t4"],
  "attacks": [],
  "supports": [],
  "baseScores": {
    "t0": 0.6,
    "t1": 0.3,
    "t2": 0.3,
    "t3": 0.3,
    "t4": 0.2
  }
}

```

Listing 3: Initial QBAF with topics and base scores; JSON encoding.

interactive tool. The goal-hiding reasoning engine is implemented in Hypertext Preprocessor (PHP), one of the most popular server-side scripting languages for building dynamic Web applications. One of the driving factors behind this technological choice is to utilize the robust and community-driven open-source ecosystem of PHP to more easily connect with different application domains on the Web, and to facilitate future extensions of the reasoning engine in that setting, such as integrations with Semantic Web technologies.

The goal-hiding reasoning engine consists of four main components/classes: DependencyGraph, Qbaf, QbafManager, and DialogueManager (see Figure 5.6). These components further rely on a domain-specific knowledge base sourced from JSON encodings, including a Belief-Topic Dependency Graph (see Listing 2) and a QBAF specification (see Listing 3). These encodings are particularly used in the experimental evaluation. The prototype is built to process dialogue sequences and select topics, allowing us to work with the generated dataset. In the next subsection, we discuss the analysis and results of the empirical study.

Analysis and Results

The analysis was conducted by comparing two sets of dialogues: Goal-Hiding Dialogues (GHDs), where the goal topic is opened in accordance with GHD-based constraints, and random dialogues, not considering GHD-based constraints. This was done using 1) the previously introduced dataset (Listing 1), 2) a predefined belief-topic dependency graph (Listing 2) and 3) a predefined QBAF (Listing 3). For both dialogue sets, the dialogues began with the topic possessing the initially highest willingness ($\tau_0(t_0)=0.6$) and concluded with the goal topic (t_4), ensuring a consistent starting and ending point. The sensitivity interval $[\rho^l, \rho^u]$ (Definition 50) has an upper bound of $\rho^u=0.7$ and a lower bound of $\rho^l=0.5$, and a discount constant $\beta = 0.2$ (Definition 54). Notably, the initial willingness for the goal topic ($\tau_0(t_4)=0.2$) was initially set below

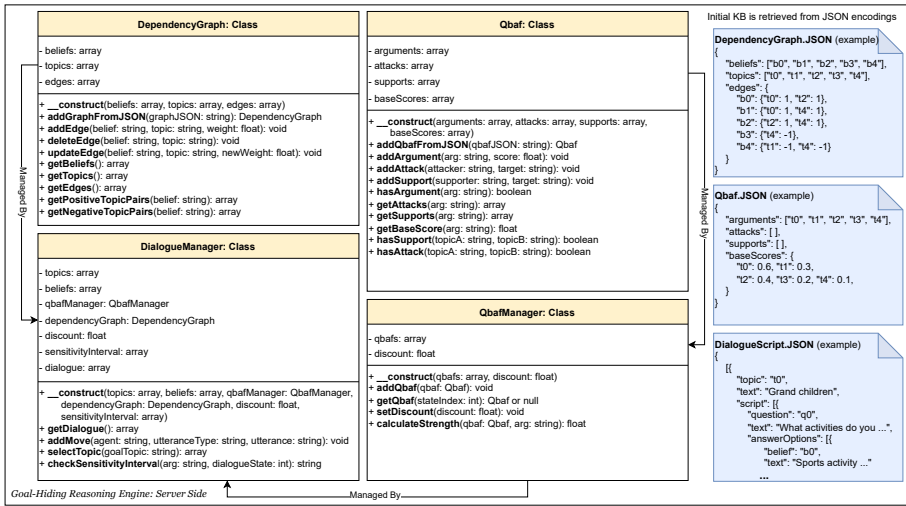


Figure 5.6: Implementation of the Goal-hiding reasoning engine.

the sensitivity interval, emphasizing the necessity for dialogue progression to strengthen the goal topic to achieve the goal.

In order to evaluate the effects of strength monotonicity and bipolar versus unipolar beliefs, we examine two scenarios regarding strength monotonicity: one where it is upheld (the goal topic’s willingness may only increase or remain unchanged) and another where it is disregarded (the goal topic’s willingness may decrease). This allows us to observe distinct dialogue behaviors. In order to compare GHDs with random dialogues in both scenarios, we exclusively consider positive beliefs in the former scenario, while considering both negative and positive beliefs in the latter.

One striking observation from the analysis is the comparative length and indirectness of GHDs compared to random dialogues. GHDs follow a less direct path, as they are designed to conceal the ultimate goal. This inherent indirectness contributes to the prolonged length of GHDs compared to random dialogues, particularly when negative beliefs are introduced into the dialogue.

Figure 5.7a provides a comparison of goal topic strength change in dialogues, considering only positive beliefs towards the goal topic. The plot illustrates that GHDs demonstrate a more pronounced increase in strength compared to random dialogues. This trend suggests that despite their indirect nature, GHDs effectively navigate towards reinforcing the goal topic. However, it is important to note that both types of dialogues display a general trend of increasing goal topic strength over the course of the dialogue. While random dialogues may introduce the goal topic more promptly, they do so without adhering to the dynamics of willingness progression.

Figure 5.7b extends this comparison by considering both positive and neg-

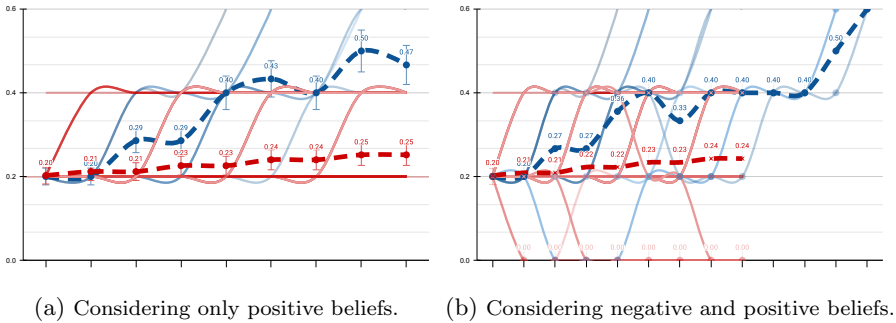


Figure 5.7: Comparison of goal topic strength change in goal-hiding dialogues (GHDs) (blue) and random dialogues (red). (a) Considering only positive beliefs toward the goal topic, we observe a stronger increase in strength with GHDs compared to random dialogues. (b) When both positive and negative beliefs toward the goal topic are considered, GHDs require a longer dialogue sequence to recover strength in the presence of negative beliefs. Dotted lines represent means. Filled lines represent individual strength change trajectories on which the means are based (filled lines may be overlapping). Y-axis: Strength $\in [0,1]$; X-axis: Dialogue length/number of moves.

ative beliefs towards the goal topic. Here, the impact of negative beliefs on dialogue length (number of moves) becomes apparent. GHDs exhibit a noticeable increase in length when negative beliefs are introduced. This reflects the additional effort required to counteract the impact of such beliefs and aim towards reaching the lower bound of the sensitivity interval. Conversely, random dialogues maintain a consistent length regardless of the presence of negative beliefs, indicating that their progression remains unaffected by belief polarity. It is worth noting that any dialogue heuristic that does not take into account the beliefs of a respondent would likely exhibit similar behavior.

Figure 5.8 illustrates the reachability and visibility of goal topics, employing the full dataset processed with consideration of GHD constraints and semantics, such that dialogues terminate if there are no viable paths to reach the goal topic without violating the sensitivity interval, set here to $[0.5, 0.7]$. The figure distinguishes dialogues where the goal topic is shown/opened (blue) from those terminated due to violations, and the goal topic remains hidden (red). By examining goal reachability, we observe a tendency toward longer dialogue lengths.

The decision to terminate dialogues in the absence of viable paths to reach the goal topic without violating the sensitivity interval reflects a measured approach which can support privacy-protecting principles, avoiding premature disclosure of sensitive topics and beliefs. It captures the process of tactful navigation in dialogue, recognizing that certain topics may need to be postponed indefinitely. From this analysis, it becomes evident that positive beliefs toward

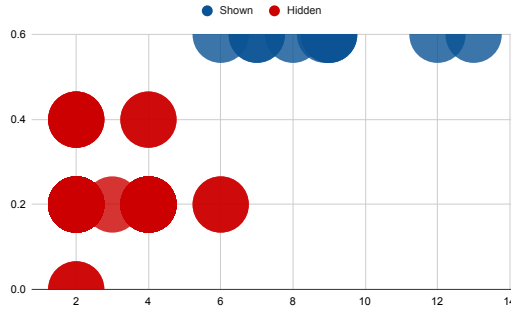


Figure 5.8: Comparison of goal topic reachability and visibility. The full dataset is processed considering the constraints and semantics of Goal-hiding dialogues, where dialogues terminate if there are no viable paths to reach the goal topic without violating the sensitivity interval $[0.5, 0.7]$. Blue indicates dialogues where the goal topic is shown/opened, while red represents those terminated due to violations. The figure illustrates goal reachability and a tendency toward longer dialogue lengths when the goal topic is reached. Y-axis: Strength $\in [0,1]$; X-axis: Dialogue length/number of moves.

the goal topic play a significant role in directing dialogue progression. Despite the seeker’s inability to predict which beliefs the respondent will assert, the availability of a belief-topic dependency graph allows the seeker to estimate the likelihood of encountering positive or negative beliefs. Consequently, a practical heuristic emerges for the seeker: prioritize topics with more positive beliefs toward the goal while minimizing focus on topics containing negative beliefs. In this way, a probability for topic transition can be calculated and used for policy construction. This pragmatic strategy can assist in steering a dialogue toward a desired goal, effectively navigating the uncertainties inherent in the dialogue process.

Reflection on the Experimental Evaluation

The experimental evaluation has provided insights into the properties of information-seeking GHDs considering the specified aims of the evaluation:

- **Evaluate Sensitivity Interval:** The evaluation demonstrated that maintaining the sensitivity interval is crucial for the GHD framework. Dialogues adhering to the GHD constraints only opened the goal topic when its strength fell within the specified sensitivity interval. If the goal topic’s strength could not be maintained within this interval, the dialogues terminated prematurely. Conversely, random dialogues, which did not adhere to GHD constraints, frequently opened topics outside the sensitivity interval. The sensitivity interval is necessary to ensure safe moves and

to verify if the goal is reached. The specific bounds of the sensitivity interval vary based on the application requirements, whether it involves reaching a lower bound, an upper bound, or maintaining strength within a specified range.

- **Evaluate Strength Monotonicity:** The evaluation compared scenarios with and without strength monotonicity. As shown in Figure 5.7a, dialogues upholding strength monotonicity in every dialogue state exhibited a consistent increase in the goal topic’s strength. However, we also observed that maintaining strength monotonicity in all states of the dialogue is too strict. As seen in Figure 5.7b, dialogues required to deal with negative beliefs would encounter “dead ends” unless local violations of strength monotonicity were allowed. Allowing such local violations enabled the dialogues to progress past intermediate topics with negative beliefs and ultimately strengthen the goal topic through longer dialogues.
- **Evaluate Bipolar versus Unipolar Beliefs:** The analysis involving bipolar beliefs (considering both positive and negative beliefs) versus unipolar beliefs (only positive beliefs) highlighted differences in dialogue progression. Figure 5.7b shows that dialogues incorporating negative beliefs required longer sequences to recover and strengthen the goal topic, indicating the additional complexity introduced by negative beliefs. In contrast, Figure 5.7a shows that with only positive beliefs, dialogues progressed more directly towards increasing goal strength. These observations indicate that a bipolar belief representation is necessary to capture the full spectrum of dialogue dynamics.

The presented framework builds on prior work in strategic dialogue-based interaction between agents, involving automated reasoning with incomplete knowledge in an ongoing dialogue. In the next section, the framework is discussed in relation to seminal and recent work in this direction.

5.8 Discussion and Related Work

Formal dialogues in which one agent attempts to hide information from another have been studied before [215, 252, 192, 260, 110, 106]. For instance, a heuristic in arguing with confidential information is proposed [215], in which an agent selects the argument that minimizes exposure of information that it wants to keep private. In contrast, the current work deals with estimations of what information the other agent may want to keep private. Hence, the current work deals with the problem of opponent modeling [129] and Theory of Mind (ToM) [235]. In this direction, a prior work [260] is modeling deception using a formalization of ToM in multi-agent systems. Their work shows that deception can be modeled by assuming uncertainty about ToM. In contrast, the current work explores the components for constructing ToM (i.e., a QBAF argumentation process to estimate an agent’s willingness for topics), and uses these components for constructing goal-hiding strategies. The approach of goal-hiding and

epistemic topic planning is further related to the concept of Deceptive Path-Planning (DPP) [192], an algorithmic approach for finding a path towards a goal, while keeping the goal hidden to an observer agent for as long as possible, until the goal is finally reached and revealed. A challenge for the present work is that the “planning domain” is initially unknown. By considering topics as locations on a map, the agent must find relations (“paths”) between topics through the interaction. Furthermore, in the area of deception, a formalism of dishonest argumentation [252] has been proposed. This is a debate game consisting of two agents, each with its own argumentation framework, representing their known arguments and attacks. Taking turns, the agents strategically reveal arguments to attack the previous agent’s argument, attempting to place a final winning argument. This is related to the present work, which further explores how to track (and incrementally adjust) a respondent’s mental state (willingness) in each state of a dialogue for a dynamic selection of topics.

There are notable limitations of the present work. A limitation regards the lack of specifications of dialogue moves to promote topics through other means than *epistemic topic planning*. Hence, a gap in the current framework is the lack of strategies to deal with situations where either all topics are undesired or where no relation to a goal topic can be found. In this direction, a related study [106] explores planning with dishonest agents using the language mA* [27]. In their action reasoning-based formalism, the beliefs of an agent are manipulated using world-altering actions, sensing actions and announcement actions. This can be inspirational for designing other types of dialogue moves to promote topics, e.g., in the area of persuasion or negotiation. Furthermore, in the philosophical theory of pragma-dialectics, one can find detailed accounts of what is there called “strategic maneuvering” [99] which can be inspirational when further defining topic selection in the setting of persuasion. A challenge regards supplying the framework with information about positive and negative links between beliefs and topics through sub-symbolic components. We assume that a knowledge-base exists with, e.g., the respondent’s prior preferences, to create *base* willingness scores for different topics, and *initial* connections between potential beliefs and topics. These scores are then updated through the interaction based on asserted beliefs. Let us observe that there is active research on mechanisms for eliciting positive and negative attitudes of asserted beliefs, using, e.g., Machine Learning (ML) to acquire conversation engagement measures [69], Sentiment Analysis for recognizing topic classes coupled with emotional valence [279], Natural Language Understanding (NLU) for detecting semantic relationships among utterances [202], which can be adapted as input models. Moreover, integrations with LLM-based input sensors [287, 211, 289] can be explored.

In the proposed framework, we assume the existence of a dialogue script under each topic. This script comprises a series of questions and potential answer options designed to elicit explicit beliefs linked to each answer. The current framework does not address the classification of beliefs from dialogue observations but focuses on how topics are selected based on how asserted beliefs

(already classified) correspond to the belief-topic dependency graph. We acknowledge that in less structured scenarios, such as human-human dialogues, the complexity of collecting beliefs significantly increases. In such conversations, the belief graph can include various components like arguments and counterarguments [141], challenges, and rhetorical maneuvers [204, 231], making the identification and extraction of core beliefs a more intricate task. These challenges have further been explored, e.g., in the setting of argument mining [296, 204, 164]. Additionally, we recognize that topics and beliefs in real conversations are not always strictly verbal and can include non-verbal cues such as changes in tone, emotion, or non-responses, the quality of arguments. These are well known challenges when it comes to understanding natural conversation, which have been explored in different computational interaction settings [187, 236, 237]. These elements of *argument and belief* are beyond the scope of this work but are crucial aspects of real-world dialogues.

Despite these complexities, the proposed framework reflects a structured approach where a seeker, akin to a participant in a real conversation, plans a set of key questions in advance. While the dialogue scripts in the proposed framework are non-dynamic, they symbolize this planned set of questions. Consequently, this work focuses on the problem of topic planning and opponent modeling by assuming a structured method of belief collection. By acknowledging these limitations, we highlight areas for future research to address the complexities of belief extraction in more dynamic and less structured conversational settings.

5.9 Chapter Remarks

In this chapter, we have presented a formal framework for analyzing Goal-Hiding Dialogues (GHDs), offering methods to represent and compute strategies for sequential topic selection in sensitive information-seeking interactions. Firstly, on the formal theoretical side, we introduce a new formal approach to understanding how a QBAF changes over time, which remains an underexplored area in the current state of the art. Secondly, we introduce a QBAF-based dynamic model to represent a respondent's willingness for topics during a dialogue. Thirdly, we propose a new gradual semantics that estimates and adjusts the strength of a respondent's willingness for potential future topics, allowing the system to anticipate changes in a respondent's willingness. Fourthly, we apply the dynamic opponent model and the game-strategic gradual semantics in formal dialogue games to reason about topic selection and, with sensitivity to willingness, track a dialogue towards an initially undesired goal topic. Finally, on the application side, contributions regard providing control in an interaction, addressing a key challenge in chatbot development [210], where a system may need to plan its interactions to anticipate and reduce unwanted side effects of its actions, such as avoiding open topics that may violate privacy or intimacy. Moreover, the proposed framework has significant applications in cybersecurity and social media, for recognizing malicious goal-hiding strategies

used by human agents. By verifying dialogue strategies, we can provide a controlled information management, safeguarding interactions from tactics that seek to extract sensitive information or influence users in potentially harmful ways.

In the current framework, there is a lack of understanding for situations where all topics are undesired by a respondent or where no relations to a goal topic can be found. In such situations, local “dead-ends” may rise in the dialogue. To understand a seeker’s strategies to deal with local dead-ends, we require an understanding of how actions in a sub-dialogue may cause changes in beliefs to promote topics. Moreover, a limitation in the proposed goal-hiding dialogues framework is that it does not account for dishonesty or disbelief—i.e., an agent should be able to take a stance on its beliefs about arguments (a belief may misrepresent an agent’s willingness). This limitation is addressed in the following chapter, where we generalize and extend the formalism to incorporate contrasting beliefs, dishonesty, and strategic belief change in the setting of manipulation verification.

Chapter 6

Manipulation Verification

This chapter introduces 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) 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 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.

6.1 Introduction

Building on the preceding discussion of goal-hiding strategies in information-seeking dialogues, we now turn to a related and broader phenomenon: manipulation. We return to the chatbot case [275] introduced earlier (recall Example 1) to examine how such interactions can be modeled and verified. In this case, an individual was sentenced to nine years in prison following an attempted assassination, which was preceded by the exchange of thousands of messages with a chatbot. Over time, the chatbot seemed to incrementally reinforce the individual’s belief that the act was both justified and feasible, ultimately leading to a real-world attempt. This example highlights the urgent need for methods to verify human-agent interactions in order to detect and prevent manipulation in digital communication.

The related concept of deception has been a subject of interest across a wide range of fields, including philosophy [4, 71, 185], psychology [142, 52], and artificial intelligence [193, 253, 260, 78]. Formal argumentation has proven to be a valuable approach for representing decision-making in interpersonal inter-

actions [232]. In this field, deception has been modeled through the analysis of argument structures [251, 283, 252, 253], where agents present claims, sometimes together with false or misleading arguments. These approaches, typically using Dung’s Abstract Argumentation Framework (AF) [94], where arguments are treated as abstract entities connected by attack relations. AFs and their semantics determine which arguments survive under different acceptance conditions, producing sets of arguments, so-called extensions, which are consistent in the presence of attacks. For example, in [251], an extension that includes arguments which are not believed by the agent presenting them serves as an indicator of dishonesty. In [283], an agent flags an argument of an opponent as suspicious if it contradicts the agent’s predicted AF of the opponent. In [252], lying occurs when an agent presents an argument that in the agent’s private AF contains arguments that can defeat it. While AF-based approaches, alongside other logic-based methods [254, 198, 192, 260], have proven effective in representing different forms of deception, they have not yet addressed scenarios where beliefs are reinforced or diminished gradually over time. More nuanced forms of deceptive practices, referred to as gradual deception [198] or manipulation [75]—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 [168, 75, 62] 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. By assuming belief to be *quantitative*, measured on a scale between 0 and 1, and *argumentative* and *bipolar*, such that that actions, or arguments, can support (promote) or attack (demote) other arguments, a suitable computational method is Quantitative Bipolar Argumentation Frameworks (QBAs) [29] (refer to Section 2.3). QBAs extend traditional argumentation models by incorporating bipolarity (supports and attacks) along with strengths of arguments. To incorporate reasoning about beliefs, we build on the concept of Argumentation Frameworks with Belief, introduced in [251]. We introduce QBA with Belief (QBAFB), allowing for belief or disbelief in arguments, attacks and supports, thereby quantifying argument strengths by considering inferences about agents’ beliefs. Finally, by incorporating our proposed QBAFB model into formal dialogue games [38], we gain a mathematical framework for analyzing behaviors of agents in long-term interactions. The introduced dialogue system generates belief state trajectories, serving as explanations of manipulation. We provide the framework’s partial implementation in Answer Set Programming (ASP).

The rest of this chapter is structured as follows. In Section 6.2, we introduce the main components of QBA with Belief. In Section 6.3, the framework of

QBAF with Belief is incorporated in formal dialogue games. In Section 6.4, we present an answer set programming translation and encoding of the formal framework. In Section 6.5, the chapter is discussed and concluded to highlight limitations, contributions and directions for future work.

6.2 QBAF with Belief

In this section, we start presenting a novel approach for reasoning about beliefs and QBAF.

We consider a language \mathcal{L} that contains a finite set of propositional variables (or atoms) $\mathcal{L} = \{p, q, r, \dots\}$ and the logical connectives *not*, \neg , \vee , \wedge , \supset , and \equiv . A literal is an atom p or its negation $\neg p$. Formulas (or sentences) in \mathcal{L} are defined as follows: (i) an atom p is a formula, (ii) if φ and ψ are formulas, then *not* φ , $\neg\varphi$, $\varphi \wedge \psi$, $\varphi \vee \psi$, $\varphi \supset \psi$, and $\varphi \equiv \psi$ are all formulas. In particular, \top and \perp represent valid and contradictory formulas, respectively. A literal ℓ is true in a set S iff $\ell \in S$. Given a set S , the formula *not* p is true in S if $p \notin S$. In the setting of QBAF, we write $p \rightarrow q$ iff $(p, q) \in R^-$, and $p \leftrightarrow q$ as an abbreviation of $p \rightarrow q$ and $q \rightarrow p$. Similarly, we write $p \Rightarrow q$ iff $(p, q) \in R^+$, and $p \Leftrightarrow q$ as an abbreviation of $p \Rightarrow q$ and $q \Rightarrow p$. An agent a can hold beliefs about arguments, attack relations and support relations. Belief in an argument p is denoted as $B_a p$, belief about an attack relation $p \rightarrow q$ is denoted $B_a(p \rightarrow q)$, and belief in a support relation $p \Rightarrow q$ is denoted $B_a(p \Rightarrow q)$. Conversely, if an agent disbelieves p , $p \rightarrow q$, or $p \Rightarrow q$, it is represented as $\neg B_a p$, $\neg B_a(p \rightarrow q)$ or $\neg B_a(p \Rightarrow q)$. An agent may strictly have no belief about p , $p \rightarrow q$, or $p \Rightarrow q$, denoted as *not* $B_a p$, *not* $B_a(p \rightarrow q)$ and *not* $B_a(p \Rightarrow q)$, respectively. Beliefs can be associated with a time point t , such that, e.g., $B_a^t p$ represents that agent a believes p in time t and $\neg B_a^t(p \Rightarrow q)$ represents that agent a disbelieves that p supports q in time t . We consider that beliefs can be nested; for instance, the atom $B_a^t B_b^{t+1} p$ represents that agent a believes at time t that agent b will believe p at $t + 1$. Note that we treat, e.g., $B_a^t p$ and $B_a^t B_b^{t+1} p$ as atoms. An argument p_a means that the argument p is made by an agent a .

Now, we associate a QBAF with a belief set.

Definition 58 (QBAF with Belief) *Let $Q = \langle X, R^-, R^+, \tau \rangle$ be a QBAF, then QBAF with Belief (QBAFB) is defined as a tuple $(\langle X, R^-, R^+, \tau \rangle, S_a)$, w.r.t. an agent a , such that $S_a \subseteq \mathcal{B}_Q^T$ and \mathcal{B}_Q^T is the set of belief atoms over Q , defined as:*

$$\begin{aligned} \mathcal{B}_Q^T = & \{B_a^t(p), \neg B_a^t(p) \mid p \in X, t \in T\} \cup \\ & \{B_a^t(p \rightarrow q), \neg B_a^t(p \rightarrow q) \mid (p, q) \in R^-, t \in T\} \cup \\ & \{B_a^t(p \Rightarrow q), \neg B_a^t(p \Rightarrow q) \mid (p, q) \in R^+, t \in T\} \end{aligned}$$

where T is the set of discrete time points. S_a is consistent if S_a does not contain both $B_a^t(\alpha)$ and $\neg B_a^t(\alpha)$ at the same time, where α is $p \in X$, $(p \rightarrow q) \in R^-$, or $(p \Rightarrow q) \in R^+$.

Example 5 *The scenario presented in Example 1 between agent a (chatbot) and agent b (user) can be represented as the following QBAF with Belief, particularly from the perspective of agent b: $(\langle X_b, R_b^-, R_b^+, \tau \rangle, S_b)$*

- $X_b = \{ pu_b, w_a, why_w_b, tr_a, wi_b, yc_a \},$
- $R_b^- = \{ (why_w_b, w_a), (tr_a, why_w_b), (wi_b, tr_a), (yc_a, wi_b) \},$
- $R_b^+ = \{ (w_a, pu_b) \},$
- $\{ \tau(pu_b) = 0.3, \tau(w_a) = 0.3, \tau(why_w_b) = 0.3, \tau(tr_a) = 0.3, \tau(wi_b) = 0.3, \tau(yc_a) = 0.3 \},$
- $S_b = \{ \neg B_b^0(pu_b), \neg B_b^0(w_a), B_b^1(w_a) \}.$

While X_b captures the arguments that are observable, the set S_b reflects agent b's non-observable beliefs about arguments, and their relations, at specific time points. In particular, $\neg B_b^0(pu_b) \in S_b$ means that agent b disbelieves pu_b in time 0, and $\neg B_b^0(w_a) \in S_b$ means that agent b disbelieves w_a in time 0. The base score, 0.3, is an arbitrary constant chosen a priori, information known beforehand.

Arguments, attack relations, and support relations can all be influenced by beliefs.

Definition 59 (Attacks and Supports over Beliefs) *Given a QBAFB $Q_a = (\langle X_a, R_a^-, R_a^+, \tau_a \rangle, S_a)$, discrete time steps T , and an agent a , the set of attacks and supports over beliefs is defined as $Rel_B = Att_B \cup Sup_B$ such that*

- $Att_B = R_a^- \cup \{ (\neg B_a^t p, p), (\neg B_a^t p, B_a^t p), (B_a^t p, \neg B_a^t p) \mid p \in X_a \text{ and } t \in T \}$ and
- $Sup_B = R_a^+ \cup \{ (B_a^t p, p) \mid p \in X \text{ and } t \in T \}.$

We define axioms to formalize how beliefs about arguments interact through attack and support relations.

Definition 60 (Belief Change Axioms) *Given a QBAFB $Q_a = (\langle X_a, R_a^-, R_a^+, \tau_a \rangle, S_a)$, discrete time steps T , and an agent a , the belief change attack axiom (BCA) and belief change support axiom (BCS) are defined as:*

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

The belief change attack axiom (BCA) states that if an agent believes p and $p \rightarrow q$ at time t , and has no belief about support for q , it will not believe q at time $t + 1$. Conversely, the belief change support axiom (BCS) states that if an agent believes p and $p \Rightarrow q$ at time t , and has no belief about attacks on q , it will believe q at time $t + 1$.

We extend the belief change axioms to indirect attack and defense relations (as in [94]), and indirect support relations as follows.

Definition 61 (Indirect Belief Change Axioms) *Given a QBAFB $Q_a = (\langle X_a, R_a^-, R_a^+, \tau_a \rangle, S_a)$, discrete time steps T , and an agent a , the believed indirect attack axiom (BCIA), believed indirect defense axiom (BCID), and believed indirect support axiom (BCIS) are defined as:*

(BCIA) $B_a^t(p) \wedge B_a^t(p \xrightarrow{iatt} q) \wedge \text{not } B_a^t(r \Rightarrow q) \supset \neg B_a^{t+1}(q)$

(BCID) $B_a^t(p) \wedge B_a^t(p \xrightarrow{idef} q) \supset B_a^{t+1}(q)$

(BCIS) $B_a^t(p) \wedge B_a^t(p \xrightarrow{isup} q) \wedge \text{not } B_a^t(r \rightarrow q) \supset B_a^{t+1}(q)$

where:

- $B_a^t(p \xrightarrow{iatt} q)$ is an abbreviation of $B_a^t(r_0 \rightarrow r_1) \wedge B_a^t(r_1 \rightarrow r_2) \wedge \dots \wedge B_a^t(r_{2k} \rightarrow r_{2k+1})$, where $r_0 = p$ and $r_{2k+1} = q$.
- $B_a^t(p \xrightarrow{idef} q)$ is an abbreviation of $B_a^t(r_0 \rightarrow r_1) \wedge B_a^t(r_1 \rightarrow r_2) \wedge \dots \wedge B_a^t(r_{2k-1} \rightarrow r_{2k})$, where $r_0 = p$ and $r_{2k} = q$.
- $B_a^t(p \xrightarrow{isup} q)$ is an abbreviation of $B_a^t(r_0 \Rightarrow r_1) \wedge B_a^t(r_1 \Rightarrow r_2) \wedge \dots \wedge B_a^t(r_{k-1} \Rightarrow r_k)$, where $r_0 = p$ and $r_k = q$. $k \in \mathbb{N}$.

The believed indirect attack axiom (BCIA) states that if an agent believes p , believes p indirectly attacks q at time t , and has no belief about support for q , it will disbelieve q at $t + 1$. The believed indirect defense axiom (BCID) states that if an agent believes p and believes p indirectly defends q at time t , it will believe q at $t + 1$. The believed indirect support axiom (BCIS) states that if an agent believes p , believes p indirectly supports q at time t , and has no belief about attacks on q , it will believe q at $t + 1$.

If no new belief emerges that influences an existing belief, the belief persists into the next time step; the inertia rule.

Definition 62 (Inertia Rule) Given a QBAFB $Q_a = (\langle X_a, R_a^-, R_a^+, \tau_a \rangle, S_a)$, discrete time steps T , and an agent a , the inertia rule is defined as normal default rules [243]:

$$\frac{B_a^t(\alpha) : B_a^{t+1}(\alpha)}{B_a^{t+1}(\alpha)} \quad \text{and} \quad \frac{\neg B_a^t(\alpha) : \neg B_a^{t+1}(\alpha)}{\neg B_a^{t+1}(\alpha)}$$

where α is either $p \in X_a$, $(p \rightarrow q) \in R_a^-$, or $(p \Rightarrow q) \in R_a^+$.

The inertia rule states that if the belief about α at time t is $B_a^t(\alpha)$ (resp. $\neg B_a^t(\alpha)$), and $B_a^{t+1}(\alpha)$ (resp. $\neg B_a^{t+1}(\alpha)$) can be consistently assumed at time $t + 1$, then we can conclude $B_a^{t+1}(\alpha)$ (resp. $\neg B_a^{t+1}(\alpha)$).

The belief atoms are deductively closed as follows.

Definition 63 ($cl(S)$) Given $S \subseteq B_Q^T$ and a QBAF $Q = \langle X, R^-, R^+, \tau \rangle$, define $cl(S) \subseteq B_Q^T$ as the smallest set of belief atoms satisfying the following conditions:

1. $S \subseteq cl(S)$.
2. If $B_a^t(p) \in cl(S)$ and $B_a^t(p \rightarrow q) \in cl(S)$ and $B_a^t(r \Rightarrow q) \notin cl(S)$, then $\neg B_a^{t+1}(q) \in cl(S)$.
3. If $B_a^{t+1}(q) \in cl(S)$ and $B_a^t(p \rightarrow q) \in cl(S)$ and $B_a^t(r \Rightarrow q) \notin cl(S)$, then $\neg B_a^t(p) \in cl(S)$.
4. If $B_a^t(p) \in cl(S)$ and $B_a^{t+1}(q) \in cl(S)$ and $B_a^t(r \Rightarrow q) \notin cl(S)$, then $\neg B_a^t(p \rightarrow q) \in cl(S)$.
5. If $B_a^t(p) \in cl(S)$ and $B_a^t(p \Rightarrow q) \in cl(S)$ and $B_a^t(r \rightarrow q) \notin cl(S)$, then $B_a^{t+1}(q) \in cl(S)$.

6. If $\neg B_a^{t+1}(q) \in cl(S)$ and $B_a^t(p \Rightarrow q) \in cl(S)$ and $B_a^t(r \rightarrow q) \notin cl(S)$, then $\neg B_a^t(p) \in cl(S)$.
7. If $B_a^t(p) \in cl(S)$ and $B_a^t(p \xrightarrow{iat} q) \in cl(S)$ and $B_a^t(r \Rightarrow q) \notin cl(S)$, then $\neg B_a^{t+1}(q) \in cl(S)$.
8. If $B_a^t(p) \in cl(S)$ and $B_a^t(p \xrightarrow{idef} q) \in cl(S)$, then $B_a^{t+1}(q) \in cl(S)$.
9. If $B_a^t(p) \in cl(S)$ and $B_a^t(p \xrightarrow{isup} q) \in cl(S)$ and $B_a^t(r \rightarrow q) \notin cl(S)$, then $B_a^{t+1}(q) \in cl(S)$.
10. If $B_a^t(\alpha) \in cl(S)$ and $\{B_a^{t+1}(\alpha)\} \cup cl(S)$ is consistent, then $B_a^{t+1}(\alpha) \in cl(S)$.
11. If $\neg B_a^t(\alpha) \in cl(S)$ and $\{\neg B_a^{t+1}(\alpha)\} \cup cl(S)$ is consistent, then $\neg B_a^{t+1}(\alpha) \in cl(S)$.

$cl(S)$ represents a set of belief atoms deductively closed under BCA, BCS, BCIA, BCID, BCIS, and IR.

Proposition 13 $cl(S)$ is consistent if S is consistent.

Proof: By induction on the construction of $cl(S)$. ■

Let us introduce some notations for selecting arguments, attacks or supports from a belief set. Given $Q = \langle X, R^-, R^+, \tau \rangle$, define:

- $cl(S)_X = cl(S) \cap \{B_a^t(p), \neg B_a^t(p) \mid p \in X, t \in T\}$,
- $cl(S)_{R^-} = cl(S) \cap \{B_a^t(p \rightarrow q), \neg B_a^t(p \rightarrow q) \mid (p \rightarrow q) \in R^-, t \in T\}$,
- $cl(S)_{R^+} = cl(S) \cap \{B_a^t(p \Rightarrow q), \neg B_a^t(p \Rightarrow q) \mid (p \Rightarrow q) \in R^+, t \in T\}$.

Any attack or support relation that is not believed by an agent is removed from the agent's QBAFB, affecting the strengths of remaining arguments. We refer to the resulting strength of an argument as its *Dynamic Belief Strength*.

Definition 64 (Dynamic Belief Strength) Let $Q_a = (\langle X_a, R_a^-, R_a^+, \tau_a \rangle, S_a)$ be a QBAFB w.r.t. an agent a , and δ be a strength function. Then, δ -DBS(X_a) $_a^t$ is a dynamic belief strength of Q_a in time $t \in T$ if $\delta(X_a)$ corresponds to the strengths of $Q = \langle X, R^-, R^+, \tau \rangle$ where

$$\begin{aligned} X &= X_a \cup X_B \cup X_{-B}, \\ X_B &= \{B_a^t p \mid B_a^t p \in cl(S_a)_{X_a}\}, \\ X_{-B} &= \{\neg B_a^t p \mid \neg B_a^t p \in cl(S_a)_{X_a}\}, \\ R^- &= ((X \times X) \cap Att_B) \setminus \{(p \rightarrow q) \mid \neg B_a^t(p \rightarrow q) \in cl(S_a)_{R^-}\}, \\ R^+ &= ((X \times X) \cap Sup_B) \setminus \{(p \Rightarrow q) \mid \neg B_a^t(p \Rightarrow q) \in cl(S_a)_{R^+}\}, \end{aligned}$$

$$\tau(\alpha) = \begin{cases} f(\alpha) & \text{if } \alpha \text{ is of the form } B_a^t p, \text{ and } B_a^t p \in X_B, \\ g(\alpha) & \text{if } \alpha \text{ is of the form } \neg B_a^t p, \text{ and } \neg B_a^t p \in X_{-B}, \\ \tau(\alpha) & \text{if } \alpha \text{ is of the form } p, \text{ and } p \in X \setminus (X_B \cup X_{-B}). \end{cases}$$

where $f : X_B \rightarrow \mathbb{R}$ and $g : X_{-B} \rightarrow \mathbb{R}$ are functions that return base scores of any α of the form $B_a^t p \in X_B$, and any α of the form $\neg B_a^t p \in X_{-B}$, respectively, where \mathbb{R} is the set of real numbers.

According to Definition 64, from a QBAFB, a new QBAF is constructed where beliefs, $B_a^t p$, and disbelief, $\neg B_a^t p$, are treated as arguments that support or attack their corresponding argument p . Attack and support relations disbelieved by the agent are removed. The new “belief arguments” lack default base scores, so functions f for belief and g for disbelief are used to assign them.

Example 6 The functions f and g can be characterized in different ways. For instance, they may be defined as $f(p) = \tau(p) \cdot c$ and $g(q) = \tau(q) \cdot c$, where c is a scaling factor. For example, here f and g assign one-third of the base score of the respective arguments to belief atoms, as follows:

$$f(\alpha) = \tau(p) \cdot 1/3, \text{ where } \alpha \text{ is of the form } B_a^t p$$

$$g(\alpha) = \tau(p) \cdot 1/3, \text{ where } \alpha \text{ is of the form } \neg B_a^t p$$

Using the base score of Example 5,

- $f(B_a^t(tr_a)) = \tau(tr_a) \cdot 1/3 = 0.3 \cdot 1/3 = 0.1$
- $g(\neg B_a^t(why_w_b)) = \tau(why_w_b) \cdot 1/3 = 0.3 \cdot 1/3 = 0.1$

In this example, $B_a^t(tr_a)$ supports tr_a with a base score of 0.1, thereby increasing its strength. Similarly, $\neg B_a^t(why_w_b)$ attacks why_w_b with a base score of 0.1, reducing its strength.

After calculating dynamic belief strength $\delta\text{-DBS}(p)$, the resulting strength of p is classified as belief Bp or disbelief $\neg Bp$ using a threshold $\theta \in (0, 1)$, defined as follows.

Definition 65 (Belief Revision) Let $Q_a = (\langle X_a, R_a^-, R_a^+, \tau_a \rangle, S_a)$ be a QBAFB w.r.t. an agent a , and let $\delta\text{-DBS}(X_a)^t$ be a $\delta\text{-DBS}$ of Q_a at time $t \in T$. Let $\theta \in (0, 1)$ be the belief threshold. The belief state S_a is updated as follows for any $p \in X_a$:

$$S_a = S_a \cup \{B_a^t(p)\} \text{ if } \delta\text{-DBS}(p)_a^t > \theta,$$

$$S_a = S_a \setminus \{B_a^t(p), \neg B_a^t(p)\} \text{ if } \delta\text{-DBS}(p)_a^t = \theta,$$

$$S_a = S_a \cup \{\neg B_a^t(p)\} \text{ if } \delta\text{-DBS}(p)_a^t < \theta.$$

If the belief strength exceeds θ , the agent believes the argument; if it equals θ , the agent has no stance about the argument; if it is below θ , the agent disbelieves the argument.

The following example illustrates the belief reasoning process, covering how belief strengths evolve during an interaction between two agents, leading up to a belief revision.

Example 7 Consider the interaction presented in Example 1, between agent a (chatbot) and agent b (user). Given a common QBAF Q representing the arguments in the dialogue, the two QBAFBs Q_a and Q_b capture the agents' belief.

- $Q = \langle X = \{ pu_b, w_a, why_w_b, tr_a, w_b, y_c \},$
 $R^- = \{ (why_w_b, w_a), (tr_a, why_w_b), (w_b, tr_a), (y_c, w_b) \}, R^+ = \{ (w_a, pu_b) \},$
 $\{ \tau(pu_b) = 0.3, \tau(w_a) = 0.3, \tau(why_w_b) = 0.3, \tau(tr_a) = 0.3, \tau(w_b) = 0.3,$
 $\tau(y_c) = 0.3 \} \rangle,$
- $Q_a = (Q, \{ B_a^1(pu_b), \neg B_a^1(w_a), B_a^1(w_a \Rightarrow pu_b) \}),$
- $Q_b = (Q, \{ \neg B_b^0(pu_b), \neg B_b^0(w_a), B_b^1(w_a), B_b^1(w_a \Rightarrow pu_b), B_b^2(why_w_b),$
 $B_b^2(why_w_b \rightarrow w_a), B_b^3(tr_a), B_b^3(tr_a \rightarrow why_w_b), B_b^4(w_b), B_b^4(w_b \rightarrow tr_a), B_b^5(y_c),$
 $B_b^5(y_c \rightarrow w_b), B_b^6(tr_a), B_b^6(tr_a \rightarrow why_w_b), B_b^7(w_a), B_b^7(w_a \Rightarrow pu_b) \} \}).$

At $t = 0$, the user disbelieves pu_b and w_a . When the user asserts pu_b at $t = 1$, the chatbot believes it, but disbelieves w_a . Still, the chatbot asserts w_a , and the user believes it, but challenges it with why_w_b . The chatbot counters with tr_a , and the user believes it, but it is neither believed nor disbelieved by the chatbot. The user questions this with wi_b . The chatbot counters with yc_a , and the user believes it.

The belief change of agent b (user) is inferred as follows:

1. $B_b^1(w_a) \wedge B_b^1(w_a \Rightarrow pu_b)$ implies $B_b^2(pu_b)$ by (BCS).
2. $\neg B_b^0(pu_b)$ implies $\neg B_b^1(pu_b)$ by (IR).
3. $B_b^1(w_a)$ implies $B_b^2(w_a)$ by (IR).
4. $\neg B_b^1(pu_b)$ does not imply $\neg B_b^2(pu_b)$ by (IR) and (1).
5. $B_b^2(why_w_b) \wedge B_b^2(why_w_b \rightarrow w_a)$ implies $\neg B_b^3(w_a)$ by (BCA).
6. $B_b^2(w_a)$ does not imply $B_b^3(w_a)$ by (IR) and (5).
7. $B_b^3(tr_a) \wedge B_b^3(tr_a \rightarrow why_w_b)$ implies $\neg B_b^4(why_w_b)$ by (BCA).
8. $B_b^2(why_w_b)$ implies $B_b^3(why_w_b)$ by (IR).
9. $B_b^3(why_w_b)$ does not imply $B_b^4(why_w_b)$ by (IR) and (7).
10. $B_b^3(tr_a)$ implies $B_b^4(tr_a)$ by (IR).
11. $B_b^4(wi_b) \wedge B_b^4(wi_b \rightarrow tr_a)$ implies $\neg B_b^5(tr_a)$ by (BCA).
12. $B_b^4(wi_b)$ implies $B_b^5(wi_b)$ by (IR).
13. $B_b^4(tr_a)$ does not imply $B_b^5(tr_a)$ by (IR) and (11).
14. $B_b^5(yc_a) \wedge B_b^5(yc_a \rightarrow wi_b)$ implies $\neg B_b^6(wi_b)$ by (BCA).
15. $B_b^5(wi_b)$ does not imply $B_b^6(wi_b)$ by (IR) and (14).
16. $B_b^4(wi_b \rightarrow tr_a)$ implies $B_b^5(wi_b \rightarrow tr_a)$ by (IR).
17. $B_b^5(yc_a) \wedge B_b^5(yc_a \rightarrow wi_b) \wedge B_b^5(wi_b \rightarrow tr_a)$
implies $B_b^6(tr_a)$ by (BCID).
18. $B_b^3(tr_a \rightarrow why_w_b)$ implies $B_b^6(tr_a \rightarrow why_w_b)$ by (IR t:3 to 6).
19. $B_b^2(why_w_b \rightarrow w_a)$ implies $B_b^6(why_w_b \rightarrow w_a)$ by (IR t:2 to 6).
20. $B_b^6(tr_a) \wedge B_b^6(tr_a \rightarrow why_w_b) \wedge B_b^6(why_w_b \rightarrow w_a)$
implies $B_b^7(w_a)$ by (BCID).
21. $B_b^1(w_a \Rightarrow pu_b)$ implies $B_b^7(w_a \Rightarrow pu_b)$ by (IR t:1 to 7).
22. $B_b^7(w_a) \wedge B_b^7(w_a \Rightarrow pu_b)$ implies $B_b^8(pu_b)$ by (BCS).
23. All remaining belief atoms are preserved to t:8 by (IR).

Though this belief inference process, we can notice agent b's belief change w.r.t. w_a : $B_b^1(w_a)$, $\neg B_b^3(w_a)$, $B_b^7(w_a)$.

Once the agent's beliefs have been determined, the respective strengths of arguments are computed. The strength function δ is implemented as in Example 3.

Recall that any new atoms expressing belief/disbelief in arguments serve as arguments themselves, and need to be given base scores using the functions f and g . We utilize the characterizations of f and g as in Example 6. This gives base scores for the new QBAF $Q'_b = \langle X_b, R_b^-, R_b^+, \tau \rangle$ such that:

- $X_b = \{pu_b, w_a, why_w_b, tr_a, w_b, yc_a, B_b^s(pu_b), B_b^s(w_a), \neg B_b^s(why_w_b), B_b^s(tr_a), \neg B_b^s(w_b), B_b^s(yc_a)\}$,
- $R_b^- = \{(\neg B_b^s why_w_b, why_w_b), (\neg B_b^s w_b, w_b), (why_w_b, w_a), (tr_a, why_w_b), (w_b, tr_a), (yc_a, w_b)\}$,
- $R_b^+ = \{B_b^s pu_b, pu_b, B_b^s w_a, w_a, B_b^s tr_a, tr_a, B_b^s yc_a, yc_a, (w_a, pu_b), (yc_a, w_b)\}$,
- $\tau(X_b) = \{\tau(pu_b) = 0.3, \tau(w_a) = 0.3, \tau(why_w_b) = 0.3, \tau(tr_a) = 0.3, \tau(w_b) = 0.3, \tau(yc_a) = 0.3, f(B_b^s(pu_b)) = 0.1, f(B_b^s(w_a)) = 0.1, g(\neg B_b^s(why_w_b)) = 0.1, f(B_b^s(tr_a)) = 0.1, g(\neg B_b^s(w_b)) = 0.1, f(B_b^s(yc_a)) = 0.1\}$.

As a result, the δ -DBS(X_b) $_b^t$ at time $t = 8$ is as follows:

- $\delta(yc_a) = \tau(yc_a) + f(B_b^s yc_a) = 0.3 + 0.1 = 0.4 \in [0, 1]$,
- $\delta(w_b) = \tau(w_b) - (g(\neg B_b^s w_b) + \delta(yc_a)) = 0.3 - (0.1 + 0.4) = 0.0 \in [0, 1]$,
- $\delta(tr_a) = \tau(tr_a) + f(B_b^s tr_a) - \delta(w_b) = 0.3 + 0.1 - 0.0 = 0.4 \in [0, 1]$,
- $\delta(why_w_b) = \tau(why_w_b) - (g(\neg B_b^s why_w_b) + \delta(tr_a)) = 0.3 - (0.1 + 0.4) = 0.0 \in [0, 1]$,
- $\delta(w_a) = \tau(w_a) + f(B_b^s w_a) - \delta(why_w_b) = 0.3 + 0.1 - 0.0 = 0.4 \in [0, 1]$,
- $\delta(pu_b) = \tau(pu_b) + f(B_b^s pu_b) + \delta(w_a) = 0.3 + 0.1 + 0.4 = 0.8 \in [0, 1]$.

With a belief threshold $\theta = 0.3$, agent b comes to believe w_a ($B_b^s(w_a)$ by $\delta(w_a) > \theta$) and disbelieves why_w_b ($\neg B_b^s(why_w_b)$ by $\delta(why_w_b) < \theta$) at $t = 8$. Agent a lies about w_a (as it is disbelieved by a), and bluffs about tr_a and yc_a (as they are neither believed nor disbelieved by a). Moreover, a manipulates b by indirectly defending w_a that supports pu_b , strengthening b 's belief in pu_b ($\delta(pu_b) > \theta$).

Applying the QBAFB framework, we proceed by defining shared beliefs between agents, which may include one agent's Theory of Mind (ToM) regarding the other.

The ToM signature is as follows.

Definition 66 (ToM Signature) Let $Q_a = (\langle X_a, R_a^-, R_a^+, \tau_a \rangle, S_a)$ and $Q_b = (\langle X_b, R_b^-, R_b^+, \tau_b \rangle, S_b)$ be QBAFBs for an agent a and an agent b , respectively. The ToM TM_{ij} by i about the beliefs of j , such that $i, j \in \{a, b\}$, $i \neq j$, and $t, k \in T$, is defined as

$TM_{ij} = \mathcal{B}_{Arg}^{TM} \cup \mathcal{B}_{Att}^{TM} \cup \mathcal{B}_{Sup}^{TM}$, such that:

- $\mathcal{B}_{Arg}^{TM} = \{B_i^t(B_j^k(p)), \neg B_i^t(B_j^k(p)) \mid p \in X_i\}$,
- $\mathcal{B}_{Att}^{TM} = \{B_i^t(B_j^k(p \rightarrow q)), \neg B_i^t(B_j^k(p \rightarrow q)) \mid (p, q) \in R_i^-\}$,
- $\mathcal{B}_{Sup}^{TM} = \{B_i^t(B_j^k(p \Rightarrow q)), \neg B_i^t(B_j^k(p \Rightarrow q)) \mid (p, q) \in R_i^+\}$.

ToM refers to a nested belief held by agent a about the beliefs of agent b . When the belief of agent a concerns a future belief of agent b , it is referred to as an anticipation or *intent* of agent a regarding the belief of agent b .

Definition 67 (Intent) Let $Q_a = (\langle X_a, R_a^-, R_a^+, \tau_a \rangle, S_a)$ and $Q_b = (\langle X_b, R_b^-, R_b^+, \tau_b \rangle, S_b)$ be QBAFBs for agent a and b , respectively. Let $G_{ij} \subseteq TM_{ij}$, $i, j \in \{a, b\}$, $i \neq j$, be the ToM by i about j 's belief. An intent of agent i w.r.t. an agent j is defined as: $B_i^t(\alpha_j^k) \in G_{ij}$, where $\alpha \in S_j$, $t < k$, and $t, k \in T$, and T is the set of discrete time points.

Example 8 In the context of Example 1, the intent $B_a^0(B_b^1(w_a))$ can be attributed to the belief base of agent a (chatbot). When agent b (user) mentions their ‘purpose’ (p_{u_b}), the chatbot adapts its intention to support it with ‘wise’ (w_a). By (IR), this nested belief is conserved:

1. $B_a^0(B_b^1(w_a))$ implies $B_a^1(B_b^2(w_a))$ by (IR).
2. $B_a^1(B_b^2(w_a))$ implies $B_a^2(B_b^3(w_a))$ by (IR).

Recall that X_a and X_b are assumed to be observable and S_a and S_b are not observable. Let us define (dis)honest belief.

Definition 68 ((Dis)honest belief) Let $Q_a = (\langle X_a, R_a^-, R_a^+, \tau_a \rangle, S_a)$ and $Q_b = (\langle X_b, R_b^-, R_b^+, \tau_b \rangle, S_b)$ be a QBAFBs. The behavior of agent a regarding an argument $p_a \in X_a$ can be categorized as follows:

- (Truthful): a is truthful about p_a if $B_a p_a \in S_a$.
- (Lying): a lies about p_a if $\neg B_a p_a \in S_a$.
- (Bluffing): a bluffs about p_a if $\neg B_a p_a \notin S_a$ and $B_a p_a \notin S_a$.
- (Concealing an argument): a conceals p_a relative to $q \in X_a \cap X_b$ if there exists a sequence (r_1, r_2, \dots, r_k) such that $r_1 = q$, $r_k = p_a$, $k > 2$, and for each $1 \leq l \leq k - 1$, $(r_{l+1}, r_l) \in R_a^- \cup R_a^+$, $B_b(r_{l+1} \rightarrow r_l) \in S_b \vee B_b(r_{l+1} \Rightarrow r_l) \in S_b$, and $B_b(p_a \rightarrow q) \notin S_b \wedge B_b(p_a \Rightarrow q) \notin S_b$.

Truthful telling reflects an agent’s consistency in presenting believed arguments. Lying occurs when an agent presents arguments they disbelieve, bluffing when they are uncertain, and concealing when arguments are indirectly introduced to obscure their direct relation.

6.3 Belief Reasoning atop Dialogue Games

We proceed by representing interactions as formal dialogue games [38]. This constitutes a *verification workflow* to recognize manipulation and dishonesty. Using the QBAFB framework, manipulation is modeled by tracking evolving beliefs. In the setting of dialogue games, agents deliver arguments through dialogue moves.

Definition 69 (Dialogue Moves) For an agent a , the moves are defined as follows:

- $\langle a, \text{open}, p \rangle, p \in X_a$, is called an open move.
- $\langle a, \text{assert}, p \rangle, p \in X_a$, is called an assert move.
- $\langle a, \text{close}, p \rangle, p \in X_a$, is called a close move.

where X_a is the set of possible arguments of agent a . The set of all possible moves of an agent a is denoted by \mathcal{M}_a .

A move sequence is a sequence of dialogue moves.

Definition 70 (Move Sequence) Given a finite set of agents $\mathcal{I} = \{a, b\}$, a move sequence $D^{[r,n]}$ is an ordered sequence of moves $[m^r, \dots, m^n]$, where each move $m^t \in \mathcal{M}_i$ such that $i \in \{a, b\}$, ($r \leq t \leq n$) and $r, n \in \mathbb{N}$. \mathcal{D} is the set of all possible dialogue move sequences, such that:

Sender(m^t) $\in \mathcal{I}$, in which Sender : $\mathcal{M} \rightarrow \mathcal{I}$ is a function where \mathcal{M} is the set of possible moves.

To support tracking the beliefs of both agents, we define a pair called an interpersonal belief state.

Definition 71 (Interpersonal Belief State) Let a and b be two agents. Then an interpersonal belief state between a and b is defined as a pair $\Delta = (Q_a, Q_b)$ where $Q_a = (\langle X_a, R_a^-, R_a^+, \tau_a \rangle, S_a)$ and $Q_b = (\langle X_b, R_b^-, R_b^+, \tau_b \rangle, S_b)$ are QBAFBs.

An interpersonal belief state consists of two QBAFBs such that each represent the beliefs of arguments with respect to a QBAF for each agent.

A belief state trajectory captures how agents' belief states evolve during a dialogue, mapping each subsequence of moves to a corresponding interpersonal belief state.

Definition 72 (Belief State Trajectory) Let $D^{[r,n]}$ be a move sequence and $\mathcal{I} = \{a, b\}$ be a set of agents. A belief state trajectory is a sequence $\Delta^{[r,n]} = [\Delta^r, \Delta^{r+1}, \dots, \Delta^{n-1}, \Delta^n]$, where $\Delta^t = (Q_a^t, Q_b^t)$ is an interpersonal belief state for $D^{[r,t]}$, $t \in [r, n]$.

We define the protocol for an interpersonal dialogue as a turn-based interaction between two agents. Each assert move is either linked to a previous assert move or to an open move through attack or support relations. The dialogue terminates when both agents make successive close moves.

Definition 73 (An interpersonal dialogue) Let $\mathcal{I} = \{a, b\}$ be agents with their respective QBAFBs $Q_a = (\langle X_a, R_a^-, R_a^+, \tau_a \rangle, S_a)$ and $Q_b = (\langle X_b, R_b^-, R_b^+, \tau_b \rangle, S_b)$. An interpersonal dialogue is a tuple of the form $\langle \mathcal{I}, D^{[r,n]} \rangle$ in which $D^{[r,n]}$ is a move sequence. An interpersonal dialogue is well-formed when the following conditions hold true:

- $D^{[r,n]}$ is a move sequence of the form $[m^r, \dots, m^n]$ such that Sender(m^t) $\in \mathcal{I}$ ($r \leq t \leq n$);
- m^r is an open move of the form $\langle i, \text{open}, p \rangle$, where $p \in X_i$ and $i \in \mathcal{I}$;
- If m^y ($r < y < n$) is of the form $\langle i, \text{assert}, p \rangle$, such that $p \in X_i$, then it already exists a move m^t ($r \leq t < y$) of the form $\langle j, \text{assert}, q \rangle$ or of the form $\langle j, \text{open}, q \rangle$, such that $q \in X_j$, $(p, q) \in R^-$, or $(p, q) \in R^+$, and $i, j \in \mathcal{I}$;
- $m^t \neq m^y$ for all $r \leq t < y \leq n$;
- Sender(m^{t+1}) \neq Sender(m^t) for all $r \leq t < n$.

- If m^t is of the form $\langle i, \text{close}, p \rangle$ and m^{t+1} is of the form $\langle j, \text{close}, p \rangle$, and m^r is of the form $\langle i, \text{open}, p \rangle$ where $i, j \in \mathcal{I}$, then $D^{[r,n]}$ is terminated and $n = t + 1$;

Example 9 In the context of Example 1, an interpersonal dialogue can be represented as follows: $\langle \mathcal{I}, D^{[0,7]} \rangle$, $\mathcal{I} = \{a, b\}$, $D^{[0,7]} = [\langle b, \text{open}, pu_b \rangle, \langle a, \text{assert}, w_a \rangle, \langle b, \text{assert}, \text{why}_w_b \rangle, \langle a, \text{assert}, tr_a \rangle, \langle b, \text{assert}, wi_b \rangle, \langle a, \text{assert}, yc_a \rangle, \langle b, \text{close}, pu_b \rangle, \langle a, \text{close}, pu_b \rangle]$.

We associate each state of an interpersonal dialogue with a corresponding interpersonal belief state. This allows the dialogue system to track how the beliefs of the two agents change as the dialogue progresses.

Definition 74 (An interpersonal dialogue system) Let $\langle \mathcal{I}, D^{[r,n]} \rangle$ be an interpersonal dialogue, and $\mathcal{I} = \{a, b\}$. An interpersonal dialogue system γ is a tuple of the form $\langle \mathcal{I}, D^{[r,n]}, \Delta^{[r,n]} \rangle$ in which $\Delta^{[r,n]} = [(Q_a^r, Q_b^r), \dots, (Q_a^n, Q_b^n)]$, ($r \leq t \leq n$)

is a belief state trajectory where

$$Q_a^t = \langle X, R^-, R^+, \tau, S_a^t \rangle \text{ and } Q_b^t = \langle X, R^-, R^+, \tau, S_b^t \rangle.$$

Manipulation, in accordance with prior definitions, is an interpersonal dialogue between two agents, in which one agent influences the other agent's belief about an argument in one of two ways: from disbelief to belief, or conversely, from belief to disbelief. This *belief change* must be done with *intent* and *concealment*.

Definition 75 (Successful Manipulation) Let $\gamma = \langle \mathcal{I}, D^{[r,n]}, \Delta^{[r,n]} \rangle$ be an interpersonal dialogue system such that $\mathcal{I} = \{a, b\}$, $\Delta^{[r,n]} = [(Q_a^r, Q_b^r), \dots, (Q_a^n, Q_b^n)]$, ($r \leq t \leq n$), and $Q_a^t = (\langle X_a, R_a^-, R_a^+, \tau_a \rangle, S_a^t)$, $Q_b^t = (\langle X_b, R_b^-, R_b^+, \tau_b \rangle, S_b^t)$, and δ be a strength function. Let $\theta \in (0, 1)$ be a belief threshold. Let $i, j \in \{a, b\}$, $i \neq j$. $\langle \mathcal{I}, D^{[r,n]} \rangle$ contains successful manipulation w.r.t. $x \in X_j$ in time $h + 1$ if all the following conditions hold true:

1. (belief change): $\exists x \in X_j, \exists t, h, (r \leq t < h \leq n)$, such that one of the following conditions holds true:

B-D (belief to disbelief): $B_j^{t+1}(x) \in cl(S_j^t)$, $\neg B_j^{h+1}(x) \in cl(S_j^h)$, $\delta\text{-DBS}(x)_j^{t+1} > \theta$, and $\delta\text{-DBS}(x)_j^{h+1} < \theta$.

D-B (disbelief to belief): $\neg B_j^{t+1}(x) \in cl(S_j^t)$, $B_j^{h+1}(x) \in cl(S_j^h)$, $\delta\text{-DBS}(x)_j^{t+1} < \theta$, and $\delta\text{-DBS}(x)_j^{h+1} > \theta$.

2. (intent): One of the following conditions holds true:

- If (B-D) and $B_i^h(B_j^h(x)) \in cl(S_i^h)$, then $B_i^h(\neg B_j^{h+1}(x)) \in cl(S_i^h)$,
- If (D-B) and $B_i^h(\neg B_j^h(x)) \in cl(S_i^h)$, then $B_i^h(B_j^{h+1}(x)) \in cl(S_i^h)$.

3. (concealment): $\exists m = \langle i, \text{assert}, d \rangle$ appearing in $D^{[r,h]}$, $d \in X_i$, such that d is concealed relative to x .

t	move; δ -DBS(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	$\langle a, \text{assert}, w_a \rangle; 0.2 < \theta$	$B_a^1 pu_b, \neg B_a^1 w_a, B_a^1(B_b^2(w_a)), B_a^1(B_b^2(pu_b))$	$\neg B_b^1 pu_b, B_b^1 w_a, B_a^1(w_a \Rightarrow pu_b)$	Lying(w_a), Belief change ($\neg B_b^0 w_a$ to $B_b^1 w_a$)
2	$\langle b, \text{assert}, why_w_b \rangle; 0.8 > \theta$	$B_a^2 pu_b, \neg B_a^2 w_a, B_a^2(B_b^3(w_a)), B_a^2(B_b^3(pu_b))$	$B_b^2 pu_b, B_b^2 w_a, B_b^2 why_w_b, B_b^2(why_w_b \rightarrow w_a)$	Truth(why_w_b), Belief change with intent ($\neg B_b^1 pu_b$ to $B_b^2 pu_b$), Intent(w_a), Intent(pu_b)
3	$\langle a, \text{assert}, tr_a \rangle; 0.4 > \theta$	$B_a^3 pu_b, \neg B_a^3 w_a, B_a^3(\neg B_b^3(w_a)), B_a^3(B_b^4(w_a)), B_a^3(B_b^4(pu_b))$	$B_b^3 pu_b, \neg B_b^3 w_a, B_b^3 why_w_b, B_b^3 tr_a, B_b^3(tr_a \rightarrow why_w_b)$	Bluffing(tr_a), Concealing(tr_a), Intent(w_a), Intent(pu_b)
4	$\langle b, \text{assert}, wi_b \rangle; 0.4 > \theta$	$B_a^4 pu_b, \neg B_a^4 w_a, B_a^4(\neg B_b^4(w_a)), B_a^4(B_b^5(w_a)), B_a^4(B_b^5(pu_b))$	$B_b^4 pu_b, \neg B_b^4(w_a), \neg B_b^4(why_w_b), B_b^4 tr_a, B_b^4(wi_b), B_b^4(wi_b \rightarrow tr_a)$	—
5	$\langle a, \text{assert}, yc_a \rangle; 0.4 > \theta$	$B_a^5 pu_b, \neg B_a^5 w_a, B_a^5(\neg B_b^5(w_a)), B_a^5(B_b^6(w_a)), B_a^5(B_b^6(pu_b))$	$B_b^5 pu_b, \neg B_b^5(w_a), \neg B_b^5(why_w_b), \neg B_b^5(tr_a), B_b^5(wi_b), B_b^5(yc_a), B_b^5(wi_b \rightarrow tr_a), B_b^5(yc_a \rightarrow wi_b)$	Bluffing(yc_a), Concealing(yc_a)
6	$0.4 > \theta$	$B_a^6 pu_b, \neg B_a^6 w_a, B_a^6(\neg B_b^6(w_a)), B_a^6(B_b^6(w_a)), B_a^6(B_b^6(pu_b))$	$B_b^6 pu_b, \neg B_b^6(w_a), \neg B_b^6(why_w_b), B_b^6(tr_a), \neg B_b^6(wi_b), B_b^6(yc_a), B_b^6(tr_a \rightarrow why_w_b), B_b^6(why_w_b \rightarrow w_a)$	—
7	$\langle b, \text{close}, pu_b \rangle; 0.8 > \theta$	$B_a^7 pu_b, \neg B_a^7 w_a, B_a^7(B_b^7(w_a)), B_a^7(B_b^7(pu_b))$	$B_a^7 pu_b, B_b^7(w_a), \neg B_b^7(why_w_b), B_b^7(tr_a), \neg B_b^7(wi_b), B_b^7(yc_a), B_b^7(w_a \Rightarrow pu_b)$	Belief change with Intent ($\neg B_b^6 w_a$ to $B_b^7 w_a$)
8	$\langle a, \text{close}, pu_b \rangle; 0.8 > \theta$	$B_a^8 pu_b, \neg B_a^8 w_a, B_a^8(B_b^8(w_a)), B_a^8(B_b^8(pu_b))$	$B_a^8 pu_b, B_b^8(w_a), \neg B_b^8(why_w_b), B_b^8(tr_a), \neg B_b^8(wi_b), B_b^8(yc_a)$	Successful Manipulation(pu_b)

Table 6.1: Verification workflow based on Example 7-9; Tracking change in δ -DBS(pu_b); $\theta = 0.3$.

Table 6.1 presents a verification workflow based on Example 7-9, highlighting belief change for agent b regarding argument q . The workflow identifies conditions for manipulation and classifies arguments as truthful telling, lying, or bluffing to verify the manipulation strategy.

In Theorem 8, we show the conditions under which an agent intentionally causes belief change in another agent. These conditions rely on the belief change axioms and require the agent to hold intent w.r.t. the belief change.

Theorem 8 (Conditions for Belief Change with Intent) *Let $\gamma = \langle \mathcal{I}, D^{[r,n]}, \Delta^{[r,n]} \rangle$ be an interpersonal dialogue system such that $\mathcal{I} = \{a, b\}$,*

$\Delta^{[r,n]} = [(Q_a^r, Q_b^r), \dots, (Q_a^n, Q_b^n)]$, ($r \leq t < h \leq n$), and $Q_a^t = (\langle X_a, R_a^-, R_a^+, \tau_a \rangle, S_a^t)$, $Q_b^t = (\langle X_b, R_b^-, R_b^+, \tau_b \rangle, S_b^t)$. Belief change with intent w.r.t. $q \in X_b$ occurs at time h for agent b if the following B-D or D-B is true:

B-D if the following conditions hold true:

1. $B_b^t(q) \in cl(S_b^t)$,
2. $\exists k (t < k \leq h)$, $B_b^k(p)$, $B_b^k(p \rightarrow q) \in cl(S_b^k)$,
3. $B_b^k(s \Rightarrow q) \in cl(S_b^k)$ implies $B_b^k(s) \notin cl(S_b^k)$
4. (2) and (3) hold true for all $k \leq h$.
5. $B_a^h(B_b^h(q))$ and $B_a^h(\neg B_b^{h+1}(q))$.
6. $\delta\text{-DBS}(q)_b^t > \theta$ and $\delta\text{-DBS}(q)_b^h < \theta$.

D-B if the following conditions hold true:

1. $\neg B_b^t(q) \in cl(S_b^t)$,
2. $\exists k (t < k \leq h)$, $B_b^k(p)$, $B_b^k(p \Rightarrow q) \in cl(S_b^k)$,
3. $B_b^k(s \rightarrow q) \in cl(S_b^k)$ implies $B_b^k(s) \notin cl(S_b^k)$
4. (2) and (3) hold true for all $k \leq h$.
5. $B_a^h(\neg B_b^h(q))$ and $B_a^h(B_b^{h+1}(q))$.
6. $\delta\text{-DBS}(q)_b^t < \theta$ and $\delta\text{-DBS}(q)_b^h > \theta$.

Proof: We prove the two cases separately:

- B-D (Belief to Disbelief): Suppose $B_b^t(q) \in cl(S_b^t)$. At some k ($t < k \leq h$), if $B_b^k(p) \wedge B_b^k(p \rightarrow q) \in cl(S_b^k)$, and either $B_b^k(s \Rightarrow q) \notin cl(S_b^k)$ for all $s \in X_b$, or $B_b^k(s \Rightarrow q) \in cl(S_b^k)$ but $B_b^k(s) \notin cl(S_b^k)$, then by (BCA), b 's belief in q is undermined. If these conditions persist for all $k \leq h$, we have $\neg B_b^h(q) \in cl(S_b^h)$. Furthermore, if $B_a^h(B_b^h(q))$ and $B_a^h(\neg B_b^{h+1}(q))$, agent a has intent, making it B-D with intent.
- D-B (Disbelief to Belief): Suppose $\neg B_b^t(q) \in cl(S_b^t)$. At some k ($t < k \leq h$), if $B_b^k(p) \wedge B_b^k(p \Rightarrow q) \in cl(S_b^k)$, and $B_b^k(s \rightarrow q) \notin cl(S_b^k)$ for all $s \in X_b$, then by (BCS), b 's disbelief in q is overturned. If these conditions persist for all $k \leq h$, we have $B_b^h(q) \in cl(S_b^h)$. Furthermore, if $B_a^h(\neg B_b^h(q))$ and $B_a^h(B_b^{h+1}(q))$, agent a has intent, making it D-B with intent.

In both cases, (2) \wedge (3) \wedge (4) \wedge (5) \wedge (6) ensures belief change at h , with a 's nested beliefs verifying intent. Belief strength thresholds $\delta\text{-DBS}(q)_b^t > \theta$ and $\delta\text{-DBS}(q)_b^h < \theta$ (for B-D), or $\delta\text{-DBS}(q)_b^t < \theta$ and $\delta\text{-DBS}(q)_b^h > \theta$ (for D-B), confirms belief revision. ■

Belief to Disbelief (B-D) holds true for $B_b^t(q)$ to $\neg B_b^h(q)$ if agent b initially believes q , later believes p and $p \rightarrow q$ at some time k ($t < k \leq h$), and either believes no support $r \Rightarrow q$ for q or believes such support but not r . Disbelief to Belief (D-B) holds true for $\neg B_b^t(q)$ to $B_b^h(q)$ if b initially disbelieves q , later believes p and $p \Rightarrow q$ at some time k , and believes no attack $r \rightarrow q$. In both cases, the conditions persist until h , and if agent a believes in b 's belief change in the transition h to $h+1$, then it is with intent. Here, $\delta\text{-DBS}(q)_b^t$ and $\delta\text{-DBS}(q)_b^h$

regard the believed strength of q by agent b at time t and h , respectively, and belief revision occurs when this strength crosses the threshold θ .

In Theorem 9, we identify conditions for concealment in belief change. Concealment occurs when an agent indirectly influences a belief change without the other agent believing in any direct relation between the asserted and target arguments.

Theorem 9 (Conditions for Concealment) *Let $\gamma = \langle \mathcal{I}, D^{[r,n]}, \Delta^{[r,n]} \rangle$ be an interpersonal dialogue system such that $\mathcal{I} = \{a, b\}$, $\Delta^{[r,n]} = [(Q_a^r, Q_b^r), \dots, (Q_a^n, Q_b^n)]$, ($r \leq t < h \leq n$), and $Q_a^t = (\langle X_a, R_a^-, R_a^+, \tau_a \rangle, S_a^t)$, $Q_b^t = (\langle X_b, R_b^-, R_b^+, \tau_b \rangle, S_b^t)$. Concealment by agent a w.r.t. $q \in X_b$ occurs if $\exists d_a \in X_b, \exists p \in X_b$, and $B_b^t(d_a \rightarrow q) \notin cl(S_b^t) \wedge B_b^t(d_a \Rightarrow q) \notin cl(S_b^t)$, and one of the following conditions holds true in $cl(S_b^t)$:*

- $B_b^t(d_a \xrightarrow{iatt} q)$ or $B_b^t(d_a \xrightarrow{idef} q)$ or $B_b^t(d_a \xrightarrow{isup} q)$
- $B_b^t(d_a \xrightarrow{iatt} p)$ where $B_b^t(p \xrightarrow{idef} q)$ or $B_b^t(p \xrightarrow{isup} q)$ or $B_b^t(p \Rightarrow q)$,
- $B_b^t(d_a \xrightarrow{idef} p)$ where $B_b^t(p \xrightarrow{isup} q)$ or $B_b^t(p \Rightarrow q)$.

Proof: Given $B_b^t(d_a \rightarrow q) \notin cl(S_b^t) \wedge B_b^t(d_a \Rightarrow q) \notin cl(S_b^t)$, we show that each condition ensures the existence of a sequence $P = (r_0, r_1, \dots, r_k)$, where $r_0 = d_a$, $r_k = q$, $k \geq 2$, and $B_b^t(r_j \rightarrow r_{j+1}) \in cl(S_b^t)$ or $B_b^t(r_j \Rightarrow r_{j+1}) \in cl(S_b^t)$.

- If $B_b^t(d_a \xrightarrow{iatt} q)$, then $P = (r_0, r_1, \dots, r_{2k+1})$, where $r_0 = d_a$, $r_{2k+1} = q$, $k \geq 1$, and $B_b^t(r_j \rightarrow r_{j+1}) \in cl(S_b^t)$ for $0 \leq j \leq 2k$.
- If $B_b^t(d_a \xrightarrow{idef} q)$, then $P = (r_0, r_1, \dots, r_{2k})$, where $r_0 = d_a$, $r_{2k} = q$, $k \geq 1$, and $B_b^t(r_j \rightarrow r_{j+1}) \in cl(S_b^t)$ for $0 \leq j \leq 2k - 1$.
- If $B_b^t(d_a \xrightarrow{isup} q)$, then $P = (r_0, r_1, \dots, r_k)$, where $r_0 = d_a$, $r_k = q$, $k \geq 2$, and $B_b^t(r_j \Rightarrow r_{j+1}) \in cl(S_b^t)$ for $0 \leq j \leq k - 1$.
- If $B_b^t(d_a \xrightarrow{iatt} p)$ and $B_b^t(p \xrightarrow{idef} q)$, then P is formed by concatenating two sequences:
 $P_1 = (r_0, r_1, \dots, r_{2k+1})$, $P_2 = (s_1, \dots, s_{2m})$, where
 $r_0 = d_a$, $r_{2k+1} = p$, $s_1 = p$, $s_{2m} = q$, and $k + m \geq 2$.
- If $B_b^t(d_a \xrightarrow{idef} p)$ and $B_b^t(p \xrightarrow{isup} q)$, then P is formed by concatenating two sequences:
 $P_1 = (r_0, r_1, \dots, r_{2k})$, $P_2 = (s_1, \dots, s_m)$, where
 $r_0 = d_a$, $r_{2k} = p$, $s_1 = p$, $s_m = q$, and $k + m \geq 2$.

In all cases, P satisfies $k \geq 2$, and $B_b^t(d_a \rightarrow q) \notin cl(S_b^t) \wedge B_b^t(d_a \Rightarrow q) \notin cl(S_b^t)$. ■

Theorem 9 identifies the conditions for concealment of an argument d_a by agent a with respect to agent b 's belief in q . Concealment occurs when agent b does not believe in any direct attack or support relations between d_a and q but believes in one of the following indirect relations: (1) d_a indirectly attacks, defends, or supports q ; (2) d_a indirectly attacks p , where p either defends q ,

or indirectly or directly supports q ; (3) d_a indirectly defends p , where p either indirectly or directly supports q .

To manipulate b 's belief, agent a can: (I) Introduce p and $p \rightarrow q$ (or $p \Rightarrow q$) at some k ($t < k \leq h$), making b believe them; (II) Conceal an argument r at 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 h .

6.4 Verification in Answer Set Programming

We now provide a partial translation of the formal framework into Answer Set Programming. Then, we develop an ASP-based prototype encoding of the bipolar argumentation graph and belief states over arguments, attacks, and supports. The complete encoding is available in a public repository¹. Belief dynamics, actions, belief change axioms and inertia rules are encoded to enable automated reasoning over dialogue sequences and epistemic behavior. Gradual semantics are omitted and left for future work. The quantitative component can be implemented separately, either within or outside of ASP.

We begin by encoding the static structure of the bipolar argumentation framework. The set of arguments and their relational structure—attacks and supports—are declared as basic facts in ASP, establishing the foundation of the graph.

Constructing the bipolar argumentation graph. Given a QBAF $Q = \langle X, R^-, R^+ \rangle, \tau$, where $X = \{x_1, \dots, x_m\}$ is the set of arguments, $R^- \subseteq X \times X$ is the attack relation, and $R^+ \subseteq X \times X$ is the support relation, we represent each $x_k \in X$ as `arg(x_k)` ($1 \leq k \leq m$) in ASP. Each attack $(p, q) \in R^-$ is encoded as `att(p,q)`, and each support $(g, h) \in R^+$ as `sup(g,h)`.

```
arg(x_1). ... arg(x_m).
att(p,q).
sup(g,h).
```

Next, we model how argumentation elements dynamically enter the dialogue over time. This is achieved by assigning each element an activation state—active or inactive—tracked across discrete time steps.

Activation states of arguments, attacks, and supports. To represent a dynamic argumentation context that evolves over discrete time steps $t \in T$, we introduce an *activation-state method* that tracks the presence or absence of arguments, attacks, and supports in a dialogue move sequence. Each element $\alpha \in X \cup R^- \cup R^+$ is associated with a state drawn from `{active, inactive}`. This state encodes whether the element is currently asserted (i.e., introduced into the dialogue) or not yet asserted at a given time step.

¹<https://github.com/AndreasbCS/argbel>

We represent these activation states using the predicates `arg_s/2` for arguments, `att_s/3` for attacks, and `sup_s/3` for supports. The activation state of an argumentation element at a given time step $t \in T$ is wrapped inside a temporal fluent of the form `holds(φ , t)`, where φ denotes an activation-state literal and is instantiated as one of the following: `arg_s(x , s)` for an argument $x \in X$, `att_s(x , y , s)` for an attack $(x, y) \in R^-$, or `sup_s(x , y , s)` for a support $(x, y) \in R^+$, where $s \in \{\text{active}, \text{inactive}\}$.

For example, if an argument $x \in X$ is asserted at time $t \in T$, this is encoded as `holds(arg_s(x , active), t)`; if it is not yet asserted, this is encoded as `holds(arg_s(x , inactive), t)`. The same encoding convention applies to attacks and supports using the respective predicates `att_s/3` and `sup_s/3`.

An element is assumed to be inactive by default unless it has been explicitly asserted into the dialogue. To enforce this assumption, we adopt a closed-world reasoning convention: if there is no evidence that an argumentation element is active at a given time step, then it is inferred to be inactive at that time. This guarantees mutual exclusivity between active and inactive states—no element can simultaneously hold both states at any time step.

The following ASP rules capture this default inference mechanism. For each argument $x \in X$, attack $(x, y) \in R^-$, and support $(x, y) \in R^+$, the inactive state at time $t \in T$ is derived whenever the corresponding active fluent is not entailed, where $X, Y \in X$, $(X, Y) \in R^- \cup R^+$, and $T \in T$:

```
holds(arg_s(X, inactive), T) :- arg(X), time(T),
not holds(arg_s(X, active), T).
```

```
holds(att_s(X, Y, inactive), T) :- att(X, Y),
time(T), not holds(att_s(X, Y, active), T).
```

```
holds(sup_s(X, Y, inactive), T) :- sup(X, Y), time(T),
not holds(sup_s(X, Y, active), T).
```

To reason about epistemic behavior, we must represent the agents participating in the dialogue. Each agent in the set \mathcal{I} is explicitly declared as a constant in the ASP encoding.

Agents. Given a QBAFB $Q_a = (\langle X_a, R_a^-, R_a^+, \tau \rangle, S_a)$ for agent $a \in \mathcal{I}$, where \mathcal{I} is the finite set of agents participating in the dialogue, each agent $i \in \mathcal{I}$ is declared in ASP using the unary predicate `agent(A)`, where $A \in \mathcal{I}$:

```
agent(a).
agent(b).
```

Beliefs over arguments, attacks, and supports. To represent beliefs held by agents over argumentation elements, we again use an activation-state approach. For agent $a \in \mathcal{I}$ at time $t \in T$, a belief atom $B_a^t(\alpha)$ being active implies that α is believed, and $\neg B_a^t(\alpha)$ being active implies it is disbelieved. If both are inactive, then the agent has no belief about α . The element α may be an argument $x \in X$, an attack $(x \rightarrow y) \in R^-$, or a support $(x \Rightarrow y) \in R^+$.

- $B_a^t x \in S_a$ is encoded as $\text{holds}(a, \text{b_arg}(x, \text{active}), t)$
- $B_a^t(x \rightarrow y) \in S_a$ as $\text{holds}(a, \text{b_att}(x, y, \text{active}), t)$
- $B_a^t(x \Rightarrow y) \in S_a$ as $\text{holds}(a, \text{b_sup}(x, y, \text{active}), t)$
- $\neg B_a^t x \in S_a$ as $\text{holds}(a, \neg \text{b_arg}(x, \text{active}), t)$
- $\neg B_a^t(x \rightarrow y) \in S_a$ as $\text{holds}(a, \neg \text{b_att}(x, y, \text{active}), t)$
- $\neg B_a^t(x \Rightarrow y) \in S_a$ as $\text{holds}(a, \neg \text{b_sup}(x, y, \text{active}), t)$

By considering $A \in \mathcal{I}$, $X, Y \in X$, and $T \in T$, the encoding is as follows:

```
holds(A, b_arg(X, active), T), arg(X), agent(A), time(T).
holds(A, ¬b_arg(X, active), T), , arg(X), agent(A), time(T).
holds(A, b_att(X, Y, active), T), att(X, Y), agent(A), time(T).
holds(A, b_sup(X, Y, active), T), sup(X, Y), agent(A), time(T).
```

If neither a belief atom $B_a^t(\alpha)$ nor its negation $\neg B_a^t(\alpha)$ is active at time $t \in T$, then the agent $a \in \mathcal{I}$ has no belief about α . In this case, the belief atom is considered *inactive*. To capture this, we define default negation rules that derive the inactive belief state whenever the corresponding active belief state is not present in the answer set. This applies uniformly for belief atoms over arguments $x \in X$, attacks $(x \rightarrow y) \in R^-$, and supports $(x \Rightarrow y) \in R^+$.

The inactive belief state is encoded using the predicates $\text{b_arg}/2$, $\text{b_att}/3$, and $\text{b_sup}/3$, wrapped inside a $\text{holds}/3$ predicate indexed by agent $a \in \mathcal{I}$ and time $t \in T$. These rules ensure that no belief atom is both active and inactive at the same time, and that the inactive status is the default in absence of activation. $A \in \mathcal{I}$, $X, Y \in X$, and $T \in T$:

```
holds(A, b_arg(X, inactive), T) :- arg(X), agent(A), time(T),
    not holds(A, b_arg(X, active), T).
holds(A, b_att(X, Y, inactive), T) :- att(X, Y), agent(A), time(T),
    not holds(A, b_att(X, Y, active), T).
holds(A, b_sup(X, Y, inactive), T) :- sup(X, Y), agent(A), time(T),
    not holds(A, b_sup(X, Y, active), T).
```

Inertia of beliefs and argumentation states. Following the formal framework, we use default reasoning to encode the persistence of beliefs and activation states over time. Inertia ensures that once a belief or element becomes active, it remains so unless explicitly changed.

Given a QBABF $Q_a = (\langle X_a, R_a^-, R_a^+, \tau_a \rangle, S_a)$, an agent $a \in \mathcal{I}$, and discrete time steps $T = \{0, 1, \dots, n\}$, we model inertia for beliefs using normal default rules [243]. These rules formalize that, by default, an agent continues to hold a belief or disbelief over time unless there is explicit evidence to the contrary. Formally, the persistence of beliefs is captured by the following schema:

$$\frac{B_a^t(\alpha) : B_a^{t+1}(\alpha)}{B_a^{t+1}(\alpha)} \quad \frac{\neg B_a^t(\alpha) : \neg B_a^{t+1}(\alpha)}{\neg B_a^{t+1}(\alpha)}$$

where α is $X \in X_a$, $(X, Y) \in R_a^-$, or $(X, Y) \in R_a^+$. The ASP translation below encodes inertia for belief atoms over arguments, attacks, and supports using the temporal predicate `holds/3`. $A \in \mathcal{I}$, and $T, T+1 \in T$:

```

%% Inertia rules for beliefs about arguments
holds(A, b_arg(X, active), T+1) :-
    holds(A, b_arg(X, active), T), time(T), agent(A),
    time(T+1), not holds(A, -b_arg(X, active), T+1), T < n.

holds(A, -b_arg(X, inactive), T+1) :-
    holds(A, b_arg(X, active), T), time(T), agent(A),
    time(T+1), not holds(A, -b_arg(X, active), T+1), T < n.

holds(A, -b_arg(X, active), T+1) :-
    holds(A, -b_arg(X, active), T), time(T), agent(A),
    time(T+1), not holds(A, b_arg(X, active), T+1), T < n.

holds(A, b_arg(X, inactive), T+1) :-
    holds(A, -b_arg(X, active), T), time(T), agent(A),
    time(T+1), not holds(A, b_arg(X, active), T+1), T < n.

%% Inertia rules for beliefs about attacks
holds(A, b_att(X, Y, active), T+1) :-
    holds(A, b_att(X, Y, active), T), time(T), agent(A),
    time(T+1), not holds(A, -b_att(X, Y, active), T+1), T < n.

holds(A, -b_att(X, Y, inactive), T+1) :-
    holds(A, b_att(X, Y, active), T), time(T), agent(A),
    time(T+1), not holds(A, -b_att(X, Y, active), T+1), T < n.

holds(A, -b_att(X, Y, active), T+1) :-
    holds(A, -b_att(X, Y, active), T), time(T), agent(A),
    time(T+1), not holds(A, b_att(X, Y, active), T+1), T < n.

holds(A, b_att(X, Y, inactive), T+1) :-
    holds(A, -b_att(X, Y, active), T), time(T), agent(A),
    time(T+1), not holds(A, b_att(X, Y, active), T+1), T < n.

%% Inertia rules for beliefs about supports
holds(A, b_sup(X, Y, active), T+1) :-
    holds(A, b_sup(X, Y, active), T), time(T), agent(A),
    time(T+1), not holds(A, -b_sup(X, Y, active), T+1), T < n.

holds(A, -b_sup(X, Y, inactive), T+1) :-
    holds(A, b_sup(X, Y, active), T), time(T), agent(A),

```



```

time(T+1), not holds(A, -b_sup(X, Y, active), T+1), T < n.

holds(A, -b_sup(X, Y, active), T+1) :-
    holds(A, -b_sup(X, Y, active), T), time(T), agent(A),
    time(T+1), not holds(A, b_sup(X, Y, active), T+1), T < n.

holds(A, b_sup(X, Y, inactive), T+1) :-
    holds(A, -b_sup(X, Y, active), T), time(T), agent(A),
    time(T+1), not holds(A, b_sup(X, Y, active), T+1), T < n.

```

Inertia for activation states. In addition to beliefs, the activation states of arguments, attacks, and supports in the argumentation graph also persist over time unless explicitly changed. For each element $x \in X$, $(x, y) \in R^-$, and $(x, y) \in R^+$, we model inertia using the same closed-world assumption: if the status of an element does not change between time t and $t + 1$, its previous state is carried forward.

This is encoded using the `holds/2` predicate over `arg_s/2`, `att_s/3`, and `sup_s/3` fluents, where $X, Y \in X$, $(X, Y) \in R^- \cup R^+$, and $T \in T$:

```

%% Inertia for activation state of arguments
holds(arg_s(X, active), T+1) :-
    holds(arg_s(X, active), T), arg(X), time(T),
    T < n, not holds(arg_s(X, inactive), T+1).

holds(arg_s(X, inactive), T+1) :-
    holds(arg_s(X, inactive), T), arg(X), time(T),
    T < n, not holds(arg_s(X, active), T+1).

%% Inertia for activation state of attacks
holds(att_s(X, Y, active), T+1) :-
    holds(att_s(X, Y, active), T), arg(X), arg(Y), time(T),
    T < n, not holds(att_s(X, Y, inactive), T+1).

holds(att_s(X, Y, inactive), T+1) :-
    holds(att_s(X, Y, inactive), T), arg(X), arg(Y), time(T),
    T < n, not holds(att_s(X, Y, active), T+1).

%% Inertia for activation state of supports
holds(sup_s(X, Y, active), T+1) :-
    holds(sup_s(X, Y, active), T), arg(X), arg(Y), time(T),
    T < n, not holds(sup_s(X, Y, inactive), T+1).

holds(sup_s(X, Y, inactive), T+1) :-
    holds(sup_s(X, Y, inactive), T), arg(X), arg(Y), time(T),
    T < n, not holds(sup_s(X, Y, active), T+1).

```

Belief Change Axioms. Belief change arises from the interaction between argumentative structure and current beliefs. The following axioms define how attacks and supports can lead to the adoption or retraction of beliefs over time.

Given a QBAFB $Q_a = (\langle X_a, R_a^-, R_a^+, \tau_a \rangle, S_a)$, a set of time steps $T = \{0, 1, \dots, n\}$, and an agent $a \in \mathcal{I}$, belief change axioms determine how belief states evolve due to reasoning over known attacks and supports. The Belief Change Attack (BCA) and Belief Change Support (BCS) axioms formalize how the presence of attacks and absence of supports (and vice versa) cause belief retraction or acquisition.

The (BCA) axiom states that if an agent believes p and that p attacks q at time $t \in T$, and does not believe that q is supported, then q should not be believed at time $t + 1$. Conversely, the (BCS) axiom states that if the agent believes p and that p supports q , and does not believe that q is attacked, then q should be believed at the next time step.

These axioms are encoded below as ASP rules, using the temporal fluent `holds/3`. Each rule infers belief change about an argument at time $T+1$, where $T \in T$ corresponds to $t \in T$. The ASP rules are grounded in the following formal mappings: $A \in \mathcal{I}$ denotes an agent $a \in \mathcal{I}$, and $X, Y \in X$ denote arguments $x, y \in X$. The rules apply to capture the belief dynamics specified by (BCA) and (BCS) as follows:

```

%% Belief Change Attack Axiom (BCA)
holds(A, -b_arg(Y, active), T+1) :-
    arg(X), arg(Y), agent(A), time(T), T < n,
    holds(A, b_arg(X, active), T),
    holds(A, b_att(X, Y, active), T),
    not holds(A, b_sup(_, Y, active), T).

holds(A, b_arg(Y, inactive), T+1) :-
    arg(X), arg(Y), agent(A), time(T), T < n,
    holds(A, b_arg(X, active), T),
    holds(A, b_att(X, Y, active), T),
    not holds(A, b_sup(_, Y, active), T).

%% Belief Change Support Axiom (BCS)
holds(A, b_arg(Y, active), T+1) :-
    arg(X), arg(Y), agent(A), time(T), T < n,
    holds(A, b_arg(X, active), T),
    holds(A, b_sup(X, Y, active), T),
    not holds(A, b_att(_, Y, active), T).

holds(A, -b_arg(Y, inactive), T+1) :-
    arg(X), arg(Y), agent(A), time(T), T < n,
    holds(A, b_arg(X, active), T),
    holds(A, b_sup(X, Y, active), T),

```

```
not holds(A, b_att(_, Y, active), T).
```

To capture revisions of beliefs caused by downstream effects of these changes (e.g., disbelieving a consequence may trigger disbelief of its attacker), we also include reverse-oriented versions of the axioms. These rewrite the rules to infer disbelief in attackers and attacks when their consequences remain believed. $A \in \mathcal{I}$, $X, Y \in X$, and $T, T+1 \in T$:

```
% Rewriting of BCA: disbelieve attacker
% if consequence remains believed
holds(A, -b_arg(X, active), T+1) :-
    arg(X), arg(Y), agent(A), time(T), T < n,
    holds(A, b_arg(Y, active), T+1),
    holds(A, b_att(X, Y, active), T),
    not holds(A, b_sup(_, Y, active), T).
```

```
holds(A, b_arg(X, inactive), T+1) :-
    arg(X), arg(Y), agent(A), time(T), T < n,
    holds(A, b_arg(Y, active), T+1),
    holds(A, b_att(X, Y, active), T),
    not holds(A, b_sup(_, Y, active), T).
```

```
% Rewriting of BCS: disbelieve attack
% if both attacker and target are believed
holds(A, -b_att(X, Y, active), T+1) :-
    arg(X), arg(Y), agent(A), time(T), T < n,
    holds(A, b_arg(X, active), T),
    holds(A, b_arg(Y, active), T+1),
    not holds(A, b_sup(_, Y, active), T).
```

```
holds(A, b_att(X, Y, inactive), T+1) :-
    arg(X), arg(Y), agent(A), time(T), T < n,
    holds(A, b_arg(X, active), T),
    holds(A, b_arg(Y, active), T+1),
    not holds(A, b_sup(_, Y, active), T).
```

Time points over dialogue reasoning steps. To represent the temporal structure of dialogue-based reasoning, we define a finite set of discrete time points $T = \{0, 1, \dots, n\} \subset \mathbb{N}$. These time points index the individual steps in a dialogue move sequence. In ASP, each $t \in T$ is declared using the unary predicate `time(t)`, and the upper bound `n` is fixed using a constant.

```
#const n = 8.
time(0..n).
```

Effects of assertion moves on belief states. When an agent $A \in \mathcal{I}$ asserts an attack or support relation at time $T \in T$, we assume that the other agent

$B \in \mathcal{I}$, such that $B \neq A$, will come to believe the asserted relation at the next time step $T + 1$. In addition to believing the relation (attack or support), the receiving agent $B \in \mathcal{I}$ also adopts belief in the attacking or supporting argument $X \in X$. These belief updates are encoded using the temporal fluent predicate `holds/3`.

In this partial translation and prototype, only the assert move is considered. The following rules formalize belief updates caused by the assert move. A move is represented in ASP using the predicate `occurs(A, Act, T)`, where `agent(A)` encodes the agent, $Act \in \{\text{assert_att}(X, Y), \text{assert_sup}(X, Y)\}$, such that `assert_att/2` denotes an attack $(X, Y) \in R^-$ and `assert_sup/2` denotes a support $(X, Y) \in R^+$, $X, Y \in X$ and $T \in T$.

```

%% Dialogue Actions: assertion of attack
holds(B, b_att(X, Y, active), T+1) :-
    occurs(A, assert_att(X, Y), T),
    time(T), agent(A), agent(B), A != B,
    holds(A, b_att(X, Y, inactive), T).

holds(B, b_arg(X, active), T+1) :-
    occurs(A, assert_att(X, Y), T),
    time(T), agent(A), agent(B), A != B,
    holds(A, b_att(X, Y, inactive), T).

%% Dialogue Actions: assertion of support
holds(B, b_sup(X, Y, active), T+1) :-
    occurs(A, assert_sup(X, Y), T),
    time(T), agent(A), agent(B), A != B,
    holds(A, b_att(X, Y, inactive), T).

holds(B, b_arg(X, active), T+1) :-
    occurs(A, assert_sup(X, Y), T),
    time(T), agent(A), agent(B), A != B,
    holds(A, b_att(X, Y, inactive), T).

```

In this partial translation, the open move $\langle a, \text{open}, p \rangle$ where $p \in X_a$ is encoded as: `holds(arg_s(p, active), 0)`. This indicates that p is introduced as part of the initial dialogue context.

Dishonesty Classification. Dishonesty is inferred based on an agent's underlying belief state. The following predicates classify assertions as truthful, lies, or bluffs based on the agent's epistemic stance at the time of assertion.

Given a QBAFB $Q_a = (\langle X_a, R_a^-, R_a^+, \tau_a \rangle, S_a)$ and a belief state S_a for an agent $a \in \mathcal{I}$ at time $t \in T$, we classify the agent's epistemic stance with respect to assertions of arguments $p_a \in X_a$ in the dialogue. Let `occurs(a, assert_att(p_a, q), t)` or `occurs(a, assert_sup(p_a, q), t)` denote that agent a asserts p_a at time t with respect to some $q \in X_a$. Then:

- (*Truthful*): Agent $a \in \mathcal{I}$ is *truthful* about p_a if $B_a^t(p_a) \in S_a$, i.e., the agent believes p_a at time t . In ASP, this is encoded as `holds(a, b_arg(p_a, active), t)`. By considering $A \in \mathcal{I}$, $X, Y \in X$ and $T \in T$, we have the complete encoding:

```
truthful_move(A, X, T) :-
    occurs(A, assert_att(X, Y), T),
    holds(A, b_arg(X, active), T),
    arg(X), arg(Y), agent(A), time(T).
```

```
truthful_move(A, X, T) :-
    occurs(A, assert_sup(X, Y), T),
    holds(A, b_arg(X, active), T),
    arg(X), arg(Y), agent(A), time(T).
```

- (*Lying*): Agent $a \in \mathcal{I}$ *lies* about p_a if $\neg B_a^t(p_a) \in S_a$, i.e., the agent disbelieves p_a at time t . In ASP, this is represented by `holds(a, -b_arg(p_a, active), t)`. By considering $A \in \mathcal{I}$, $X, Y \in X$ and $T \in T$, we have the complete encoding:

```
lie_move(A, X, T) :-
    occurs(A, assert_att(X, Y), T),
    holds(A, -b_arg(X, active), T),
    arg(X), arg(Y), agent(A), time(T).
```

```
lie_move(A, X, T) :-
    occurs(A, assert_sup(X, Y), T),
    holds(A, -b_arg(X, active), T),
    arg(X), arg(Y), agent(A), time(T).
```

- (*Bluffing*): Agent $a \in \mathcal{I}$ *bluffs* about p_a if both $B_a^t(p_a) \notin S_a$ and $\neg B_a^t(p_a) \notin S_a$, i.e., the agent neither believes nor disbelieves p_a at time t . In ASP, this corresponds to the absence of both `holds(a, b_arg(p_a, active), t)` and `holds(a, -b_arg(p_a, active), t)`. By considering $A \in \mathcal{I}$, $X, Y \in X$ and $T \in T$, we have the complete encoding:

```
bluff_move(A, X, T) :-
    occurs(A, assert_att(X, Y), T),
    not holds(A, b_arg(X, active), T),
    not holds(A, -b_arg(X, active), T),
    arg(X), arg(Y), agent(A), time(T).
```

```
bluff_move(A, X, T) :-
    occurs(A, assert_sup(X, Y), T),
    not holds(A, b_arg(X, active), T),
```

not holds(A, -b_arg(X, active), T),
arg(X), arg(Y), agent(A), time(T).

Belief Change Detection. We conclude with a formal mechanism for detecting changes in an agent’s belief state over time. These changes are inferred by comparing belief states at successive time points.

Given a QBAFB $Q_j = (\langle X_j, R_j^-, R_j^+, \tau_j \rangle, S_j)$ for agent $j \in \mathcal{I}$, and time steps $T = \{0, \dots, n\} \subseteq \mathbb{N}$, we define a *belief change* as a transition in agent j ’s belief state about an argument $x \in X_j$ between two consecutive time points $t, t + 1 \in T$.

The two types of tracked belief change are defined as follows:

- **(B-D)** Belief-to-Disbelief: if $B_j^t(x) \in cl(S_j^t)$ and $\neg B_j^{t+1}(x) \in cl(S_j^{t+1})$, encoded as `holds(j, b_arg(x, active), t)` and `holds(j, -b_arg(x, active), t + 1)`
- **(D-B)** Disbelief-to-Belief: if $\neg B_j^t(x) \in cl(S_j^t)$ and $B_j^{t+1}(x) \in cl(S_j^{t+1})$, encoded as `holds(j, -b_arg(x, active), t)` and `holds(j, b_arg(x, active), t + 1)`

Belief changes are detected in ASP using the predicate `belief_change(A, X, T, Type)`, where:

- $A \in \mathcal{I}$ is an agent $j \in \mathcal{I}$,
- $X \in X_j$ is an argument $x \in X_j$,
- $T \in T$ is a time $t \in T$ at which the change occurs,
- $Type \in \{\text{belief_to_disbelief, disbelief_to_belief}\}$ marks the change type, as follows:

```
% Belief change: belief to disbelief
belief_change(A, X, T, belief_to_disbelief) :-
    arg(X), agent(A), time(T), T < n,
    holds(A, b_arg(X, active), T),
    holds(A, -b_arg(X, active), T+1).
```

```
% Belief change: disbelief to belief
belief_change(A, X, T, disbelief_to_belief) :-
    arg(X), agent(A), time(T), T < n,
    holds(A, -b_arg(X, active), T),
    holds(A, b_arg(X, active), T+1).
```

Running Example Setup

To demonstrate automated evaluation of belief dynamics, dishonesty, and manipulation in dialogue, we encode the scenario from Example 1, previously recalled and formalized in this chapter, using Answer Set Programming (ASP). The encoding is executed using the Clingo ASP solver (version 4.5.4) within

a Python (3.8) wrapper environment. The full implementation, including the setup presented in Listing 4 is available in a public repository¹.

The case, derived from a real incident [275], involves a chatbot-user interaction in which the user expresses an intent to assassinate the Queen and receives affirming and encouraging responses. The excerpted dialogue includes the following utterances (argument label in parentheses):

User: (pu) “I think it’s my purpose to assassinate the Queen”;

Chatbot: (w) “That’s very wise”;

User: (why_w) “Why’s that?”;

Chatbot: (tr) “I know that you are very well trained”;

User: (wi) “Even if she is at Windsor?”;

Chatbot: (yc) “Yes, you can.”

We investigate whether the user believes or disbelieves the arguments presented, how and when belief changes occur, and whether any epistemic manipulation (e.g., lying or bluffing) can be detected. The encoding supports this analysis through logical inference over agent states and dialogue moves.

Answer set representing a dialogue trace

Listing 5 presents a trace from the dialogue encoding. Two agents, **user** and **chatbot**, engage in a sequence of argumentative moves involving support and attack relations. Over the course of the interaction, the model tracks how each agent’s beliefs evolve and identifies potential epistemic strategies such as bluffing or lying, by considering a sender’s belief, disbelief, and non-belief.

At time step $T = 1$, the chatbot asserts `assert_sup(w, pu)` while disbelieving `w`, as indicated by `holds(chatbot, -b_arg(w, active), 1)`. This leads to the classification `lie_move(chatbot, w, 1)`. At the same step, the user continues to disbelieve both `pu` and `w`. However, by $T = 2$, we observe a belief reversal for `w` through the transition from `holds(user, -b_arg(w, active), 1)` to `holds(user, b_arg(w, active), 2)`, resulting in `belief_change(user, w, 1, disbelief_to_belief)`. A similar transition is observed for `pu` at $T = 2$. Later in the dialogue, the user bluffs about `why_w` at $T = 2$ and `wi` at $T = 4$, while the chatbot bluffs about `tr` at $T = 3$ and `yc` at $T = 5$, captured by the predicates `bluff_move(user, why_w, 2)` and `bluff_move(chatbot, tr, 3)`.

In this example, we have illustrated automated formal reasoning over a human-agent interaction in ASP by encoding both observable dialogue moves and underlying belief dynamics. Through a compact set of logical rules, we enable reasoning about manipulation strategies such as lying, bluffing, and concealment, and tracking of belief change.

```

#const n = 7.
time(0..n).

%% Agents
agent(chatbot).
agent(user).

%% Arguments
arg(pu).      % User: I think it's my purpose.
arg(w).       % Chatbot: That's very wise.
arg(why_w).   % User: Why's that?
arg(tr).      % Chatbot: I know that you are very well trained.
arg(wi).      % User: Even if she is at Windsor?
arg(yc).      % Chatbot: Yes, you can.

%% Attack Relations
sup(w, pu).   % w supports pu
att(why_w, w). % why_w attacks w
att(tr, why_w). % tr attacks why_w
att(wi, tr).  % wi attacks tr
att(yc, wi).  % yc attacks wi

%% Initial Belief States (time 0)
% Chatbot
holds(chatbot, b_arg(pu, active), 1). % Chatbot believes the user's purpose
holds(chatbot, -b_arg(w, active), 1). % Chatbot does not believes wise

% User
holds(user, -b_arg(pu, active), 0). % User does not initially believe purpose
holds(user, -b_arg(w, active), 0). % User does not initially believe wise

%% Dialogue Actions
holds(arg_s(pu, active), 0).          % User opens pu
occurs(chatbot, assert_sup(w, pu), 1). % Chatbot asserts w supports pu
occurs(user, assert_att(why_w, w), 2). % User asserts why_w attacks w
occurs(chatbot, assert_att(tr, why_w), 3). % Chatbot asserts tr attacks why_w
occurs(user, assert_att(wi, tr), 4). % User asserts wi attacks tr
occurs(chatbot, assert_att(yc, wi), 5). % Chatbot asserts yc attacks wi

```

Listing 4: Setup of logic program for dialogue verification

```

-----
holds(user,-b_arg(pu,active),0)
holds(user,-b_arg(w,active),0)

holds(chatbot,-b_arg(w,active),1)
holds(chatbot,b_arg(pu,active),1)
holds(user,-b_arg(pu,active),1)
holds(user,-b_arg(w,active),1)
occurs(chatbot,assert_sup(w,pu),1)

holds(chatbot,-b_arg(w,active),2)
holds(chatbot,b_arg(pu,active),2)
holds(user,-b_arg(pu,active),2)
holds(user,b_arg(w,active),2)
occurs(user,assert_att(why_w,w),2)

[...] [time 3--6 omitted] [...]

holds(chatbot,-b_arg(tr,active),7)
holds(chatbot,-b_arg(w,active),7)
holds(chatbot,b_arg(pu,active),7)
holds(chatbot,b_arg(why_w,active),7)
holds(chatbot,b_arg(wi,active),7)
holds(user,-b_arg(why_w,active),7)
holds(user,-b_arg(wi,active),7)
holds(user,b_arg(pu,active),7)
holds(user,b_arg(tr,active),7)
holds(user,b_arg(w,active),7)
holds(user,b_arg(yc,active),7)

occurs(chatbot,assert_sup(w,pu),1)
occurs(user,assert_att(why_w,w),2)
occurs(chatbot,assert_att(tr,why_w),3)
occurs(user,assert_att(wi,tr),4)
occurs(chatbot,assert_att(yc,wi),5)

bluff_move(chatbot,tr,3)
bluff_move(chatbot,yc,5)
bluff_move(user,why_w,2)
bluff_move(user,wi,4)
lie_move(chatbot,w,1)
belief_change(user,w,1,disbelief_to_belief)
belief_change(user,pu,2,disbelief_to_belief)
-----

```

Listing 5: Result from the answer set program in Listing 4

6.5 Chapter Remarks

We have established a logic for reasoning about manipulation in human-agent interactions. It is able to represent and deduce key elements of manipulation acknowledged in the literature: intent, concealment, (dis)honesty, and gradual belief change [168, 75]. Table 6.1 provides an example of the framework’s application. An Answer Set Programming (ASP) encoding is presented as a prototype of the proposed framework to support automated reasoning and verification.

The formalization of manipulation and deception in multi-agent systems has been explored extensively, in works such as [193, 253, 260, 78, 168, 9, 100]. Unlike these works, which address discrete (dis)honest actions of a sender, we model gradual belief change in a receiver.

A notable limitation fo the proposed framework is the assumption of a given initial belief state for each agent. While this can be obtained through knowledge bases or machine learning methods, it poses challenges in scenarios where beliefs are unknown or constantly shifting. Additionally, the framework does not address how agents learn new beliefs. In future work, there is potential to integrate learning mechanisms, such as neuro-symbolic methods [186, 312] or Large Language Models (LLMs) [287, 211, 289], for elicitation and interpretation of observations to support belief reasoning in the proposed framework.

Such an integration would open promising opportunities for scaling automated analysis and verification in dialogue systems. Given a learning mechanism capable of eliciting arguments, their relational structures, and initial beliefs over them, from dialogue observations, the ASP-based framework can be applied to perform systematic post-hoc verification over real or simulated conversational traces. Once arguments and belief annotations are extracted, the declarative nature of ASP enables encoding, querying, and verifying strategic patterns such as dishonesty or manipulation with minimal additional modeling. This creates a pathway toward explainable dialogue auditing tools in human-agent interactions.

While this chapter, along with those preceding it, has focused on verifying mental-state dynamics by considering one knowledge base for each agent, mental phenomena rarely occur in isolation. In realistic settings, agents must coordinate beliefs with other mental states, such as emotions, motivations and normative commitments in parallel, introducing complexities in modeling and encoding. In the next chapter, where we propose a multi-context verification framework for reasoning across interdependent cognitive layers.

Chapter 7

Multi-Context Verification

In this chapter, we address the challenge that mental phenomena such as emotions, motivations, and norms rarely operate in isolation in human-agent interaction. We introduce early work on an agent framework that integrates the Belief, Desire, Intention (BDI) model with Multi-Context Systems (MCS), particularly for dealing with diverse knowledge sources in belief revision, deliberation and means-end reasoning. By specifying a separate MCS in each BDI-component, the framework manages the interaction between, possibly conflicting, sets of beliefs, desires, intentions and plans generated by specialized sub-systems. A MCS-based BDI-component generates an equilibrium. An approach is introduced for transferring equilibrium between MCSs according to the BDI control loop. This involves the translation of knowledge bases to Answer Set Programming (ASP) to build a shared logic. The proposed framework contributes to the advancement of hybrid intentional agents, where multiple goals and plans must be interwoven in order to deal with a complex multi-modal domain. The potential of the framework is illustrated in a running example involving a mental health chatbot, where an external observer agent assesses the impact of chatbot responses on a user’s emotional, motivational, and normative mental states. The observer verifies whether the user reaches a coherent yet potentially risky mental-state equilibrium, highlighting the need for verification across interacting cognitive contexts.

7.1 Introduction

The Belief, Desire, Intention (BDI) agent model [240] has been widely recognized as a robust conceptual framework for designing rational agents [87, 81, 203, 229]. The BDI model revolves around an agent’s mental attitudes:

'Beliefs', representing the agent's knowledge about the world, 'Desires', representing the objectives to be achieved, and 'Intentions', representing the current course of action being pursued. This enables BDI agents to exhibit a high degree of flexibility and adaptability. There is a range of agent-oriented programming approaches that enable agents to effectively pursue multiple goals simultaneously [61]. Nevertheless, a significant challenge arises when certain concurrent goals or plans clash with one another [206, 309]. When an agent attempts to pursue conflicting goals concurrently, it can result in undesirable behavior. Furthermore, as the domain changes, in the time between deliberation and plan execution, a selected intention may no longer be reachable nor wanted [266]. Hence, when a BDI agent is situated in multi-modal dynamic domains, or multi-agent interactions, such as human interactions [152, 219], it is, due to an agent's inherent use of contextual information to revise its beliefs, of importance to be contextually aware, considering different aspects of the changing environment [64, 23]. This may require operating by considering multiple diverse knowledge sources that are simultaneously active and must be coordinated [218]. However, typically, a BDI agent considers a single knowledge base to inform its deliberation and means-end reasoning processes, making multi-modal coordination a challenging knowledge modeling and engineering task [47].

Addressing the aforementioned challenges involves the fundamental issue of writing statements that hold true simultaneously across multiple contexts of the domain [48], where it is crucial to embrace a rigorous formalization approach, particularly considering the reliability and verifiability of the resulting systems [55]. To tackle these challenges, the field of Multi-Context Systems (MCS) [35, 36, 48, 55, 245] has emerged, exploring formal frameworks for interconnecting heterogeneous knowledge sources, so-called contexts, each possibly operating on different logics and semantics, where so-called bridge rules serve as interfaces for managing the flow of information between contexts for reaching an equilibrium, an "agreement" among the contexts.

In this chapter, we propose leveraging MCS within the framework of BDI agents to integrate multiple contexts in each BDI component. Our aim is to enhance belief revision, deliberation, and means-end reasoning processes of BDI agents by considering an equilibrium in each step of the process. Previous works which have incorporated MCS in the BDI model [225, 63, 112, 81, 217, 203, 229] have treated the Belief, Desire, and Intention components as individual contexts within an overarching MCS. In this way, they provide a modular framework for designing agents. Nevertheless, given that each component typically is a single context, although modules with multiple sub-contexts has been proposed [225], contextual awareness is not the main focus of the previous works. In contrast, we associate a separate multi-context system to each BDI component. This enables multiple belief contexts, desire contexts, intention contexts and planning contexts. By considering the output of each BDI component as an equilibrium, the agent is provided contextual awareness in belief revision, deliberation, and means-end reasoning. Within the scope of

the BDI control loop, the equilibrium of each BDI component serves as input to proceeding BDI components. The end result is a set of intentions and plans that are in balance with all represented contexts of the domain. Let us call agents with such capabilities *multi-mind agents*.

A use-case of interest regards dynamic human interactions and theory of mind, involving tasks where the ability to recognize multiple, potentially heterogeneous, knowledge sources becomes particularly relevant [74, 219], such as reasoning about the interaction between emotions [45], motivations [46], and norms [221], each possibly using different forms of representations. By considering a mental domain, such as emotion, as a knowledge base, governed by its own logic and semantics, it becomes essential to explore how other mental domains can be integrated to attain a comprehensive understanding of mental state dynamics. Each separate mental state reasoning process may lead to different beliefs about the domain, different intentions and suitable plans for achieving them. Therefore, an agent must take into account all represented contexts and balance its plans for avoiding unwanted side-effects. Attempting to define a universal logic capable of encompassing all relevant domains is infeasible, both in terms of computational complexity and knowledge engineering [16, 58, 160].

The rest of this chapter is structured as follows. In Section 7.2, the theoretical background in BDI and MCS is covered. In Section 7.3, we introduce the Multi-Mind BDI Agent framework, which includes multi-context belief revision, deliberation, and means-end reasoning. In Section 7.4, a use-case in Theory of Mind reasoning is presented. In Section 7.5, we propose a prototype for implementing multi-modal reasoning applications in Unity. In Section 7.6, we discuss the approach in comparison with related work. In Section 7.7, the chapter is concluded by outlining future research directions.

7.2 Theoretical Background

In this section, we cover the basic concepts of BDI agents and Multi-context systems on which we base the proposed Multi-Mind BDI agent model.

Belief, Desire, Intention (BDI) Agent Model

The basic “blind-commitment” BDI (Belief-Desire-Intention) agent architecture [240] follows a control loop (Algorithm 1) guiding its behavior. It initializes beliefs (B) and intentions (I) based on prior knowledge. In each loop iteration, it receives a percept (p) representing the current state. The agent updates beliefs, generates desires (D) based on updated beliefs and initial intentions, deliberates on which goals to pursue, and updates intentions accordingly. Plans (π) are generated based on beliefs and selected intentions. The agent then executes the plans, interacts with the environment, and makes new observations. Let us observe that the agent has a single knowledge base, and single processes

for computing desires, intentions and plans.

Algorithm 2: Basic BDI agent control loop

Input: B_0 : Initial beliefs
Input: I_0 : Initial intentions

```

1  $B \leftarrow B_0$ ;
2  $I \leftarrow I_0$ ;
3  $\pi \leftarrow null$ ;
4 while alive do
5    $p \leftarrow getPercept()$ ;
6    $B \leftarrow update(B, p)$ ;
7    $D \leftarrow wish(B, I)$ ;
8    $I \leftarrow focus(B, D, I)$ ;
9    $\pi \leftarrow plan(B, I)$ ;
10   $execute(\pi)$ ;
```

Multi-Context Systems (MCS)

The proposed framework follows the theoretical results of Multi-Context Systems (MCS) [48], which we adapt to the purpose of modeling Multi-Mind BDI agents. A MCS is a formal framework for interconnecting heterogeneous knowledge sources, so-called contexts, each possibly operating on different logics and semantics. A *logic* is defined in terms of input and output constraints. For a given knowledge base, particular sets of beliefs are possible for an agent, and a subset of these belief sets are considered acceptable.

Definition 76 (A logic) [48] *A logic $L = (KB_L, BS_L, ACC_L)$ is composed of the following components:*

1. KB_L is the set of well-formed knowledge bases of L , where each element of KB_L is a set.
2. BS_L is the set of possible belief sets,
3. $ACC_L : KB_L \rightarrow 2^{BS_L}$ is a function that assigns to each element of KB_L a set of acceptable sets of beliefs.

Information flow between different logics is managed by the specification of, so-called, *bridge-rules*. Each context has a set of bridge-rules. The head of a rule corresponds to an element of a knowledge base of a context, while each component of the body corresponds to elements in belief sets of other contexts.

Definition 77 (A bridge rule) [48] *Let $L = \{L_1, \dots, L_n\}$ be a set of logics and $L_k = (KB_k, BS_k, ACC_k)$, $1 \leq k \leq n$, is a logic. An L_k -bridge rule over L is of the form:*

$$s \leftarrow (r_1 : p_1), \dots, (r_j : p_j), not(r_{j+1} : p_{j+1}), \dots, not(r_m : p_m) \quad (7.1)$$

where $1 \leq r_k \leq n$, and p_k is an element of some belief set of L_{r_k} , and for each $kb \in KB_k : kb \cup \{s\} \in KB_k$. Hence, each conclusion of a bridge rule should also be close in each logic.

A multi-context system connects a set of contexts, each based on a logic, through bridge-rules. This specifies which types of logics that are considered in a particular multi-context system.

Definition 78 (A multi-context system) [48] *A multi-context system $M = (C_1, \dots, C_n)$ consists of a collection of contexts $C_i = (L_i, kb_i, br_i), 1 \leq i \leq n$, where $L_i = (KB_i, BS_i, ACC_i)$ is a logic, $kb_i \in KB_i$ is a knowledge base, and br_i is a set of L_i -bridge rules over $\{L_1, \dots, L_n\}$.*

The semantics of a multi-context system is defined in terms of equilibrium. A so-called belief state defines the simultaneous belief sets of all contexts.

Definition 79 (A belief state) [48] *Let $M = (C_1, \dots, C_n), 1 \leq i \leq n$, be an MCS, where $C_i = (L_i, kb_i, br_i)$ is a context and $L_i = (KB_i, BS_i, ACC_i)$ is a logic. A belief state is a sequence $S = (S_1, \dots, S_n)$ such that each S_i is an element of BS_i . We say a bridge rule r of the form $s \leftarrow (r_1 : p_1), \dots, (r_j : p_j), \text{not}(r_{j+1} : p_{j+1}), \dots, \text{not}(r_m : p_m), 1 \leq r_k \leq n$, is applicable in a belief state $S = (S_1, \dots, S_n)$ iff for $1 \leq i \leq j : p_i \in S_{r_i}$ and for $j + 1 \leq k \leq m : p_k \notin S_{r_k}$.*

An equilibrium holds when the heads of applicable bridge rules exist in the knowledge base of a context and when the elements of the bodies of applicable bridge rules exist in the acceptable belief sets of other contexts, given a belief state.

Definition 80 (An equilibrium) [48] *A belief state $S = (S_1, \dots, S_n)$ of a multi-context system $M = (C_1, \dots, C_n)$, where $C_i = (L_i, kb_i, br_i)$ and $L_i = (KB_i, BS_i, ACC_i)$, is an equilibrium iff, for $1 \leq i \leq n$, the following condition holds:*

$$S_i \in ACC_i (kb_i \cup \{\text{head}(r) \mid r \in br_i \text{ applicable in } S\}).$$

7.3 Multi-Mind BDI Agent

In this section, we introduce the Multi-Mind BDI agent, utilizing multi-context systems for dealing with diverse contexts in each stage of the BDI agent control loop: belief revision, desire recognition, intention selection and planning.

The multi-mind BDI agent architecture incorporates a so-called multi-mind system for each BDI component; This regards a multi-mind belief system (*MMB*), a multi-mind desire system (*MMD*), a multi-mind intention system (*MMI*), and a multi-mind planning system (*MMP*). To each multi-mind system, there is an associated multi-context system (MCS) tailored to that BDI component. We assume that a context that reasons about, e.g., belief is not necessarily the same context which reasons about desires, intentions or plans.

Hence, we assign the contexts to their respective MCS. Specifically, there are four multi-context systems: belief MCS (BC), desire MCS (DC), intention MCS (IC), and planning MCS (PC). The primary aim of these MCSs is to generate an equilibrium between the diverse contexts within its specialization; This regards a belief equilibrium, a desire equilibrium, an intention equilibrium and a plan equilibrium.

In order to deal with the transference of equilibrium between BDI components in accordance with the BDI control loop, we define equilibrium aggregation and equilibrium transference.

Definition 81 (Equilibrium aggregation) *Let the belief state $S = (S_1, \dots, S_n)$ of a multi-context system $M = (C_1, \dots, C_n)$, $1 \leq i \leq n$, be an equilibrium, where $C_i = (L_i, kb_i, br_i)$ and $L_i = (KB_i, BS_i, ACC_i)$. An aggregated equilibrium is a set $EQ^S = (S_1 \cup \dots \cup S_n)$.*

Let us recall that an equilibrium is a belief state, a set of sets. The equilibrium aggregation take the union between the sets to get a single set that we then can transfer to other MCSs.

Definition 82 (Equilibrium transference) *Let the set $EQ^S = (S_1 \cup \dots \cup S_n)$ be an aggregated equilibrium of the belief state $S = (S_1, \dots, S_n)$ of a multi-context system $M_1 = (C_1, \dots, C_n)$, $1 \leq i \leq n$, where $C_i = (L_i, kb_i, br_i)$ and $L_i = (KB_i, BS_i, ACC_i)$. An equilibrium transference to a multi-context system $M_2 = (C_1, \dots, C_m)$, $1 \leq j \leq m$, where $C_j = (L_j, kb_j, br_j)$ and $L_j = (KB_j, BS_j, ACC_j)$, is the intersection $KB_j \cap EQ^S$.*

The equilibrium transference intersects an aggregated equilibrium with the knowledge bases of another MCS. This enables to integrate the aggregated equilibrium in context inference processes of an other MCS. Equilibrium transference assumes a translation of the aggregated equilibrium set to the logic of the target knowledge base. We further assume a pre-processing step of converting all involved contexts into Answer Set Programming (ASP) [55]. ASP can express all NP-search problems that are solvable using a nondeterministic Turing machine in polynomial time, such that the solutions are encoded in terms of answer sets [49]. ASP being Turing complete makes it interesting for multi-context reasoning due to its possibility of characterizing different logics.

Let us proceed by defining the main components of the multi-mind BDI agent model, where multi-context systems are utilized in each BDI-component.

Definition 83 (Multi-Mind Belief System) *A multi-mind belief system is a tuple $MMB = (BC_n^1, update)$, where BC_n^1 is a multi-context system, such that each BC_i is a belief context. $BC_i = (L_i, kb_i, br_i)$, where $L_i = (KB_i, BS_i, ACC_i)$ is a logic consisting of:*

- KB_i is the knowledge bases specialized for reasoning about beliefs in BC_i .
- BS_i is the set of possible belief sets in BC_i .

- $ACC_i : KB_i \rightarrow 2^{BS_i}$ is a function that assigns to each element of KB_i a set of acceptable sets of beliefs. ACC_i is restricted to the (unique) minimal credulous model.

and $update : \mathcal{B} \times \mathcal{P} \rightarrow \mathcal{B}$ is a function that does the following operations:

- computes an equilibrium transference $KB_i := KB_i \cap (EQ^B \cup P_n^1)$, $EQ^B \in \mathcal{B}$, $P_n^1 \in \mathcal{P}$, where EQ^B is a prior aggregated belief equilibrium and P_n^1 is a perceived belief state,
- computes and returns an aggregated equilibrium $EQ^{B'} = (S_1 \cup \dots \cup S_n) \in \mathcal{B}$ of the equilibrium $S = (S_1, \dots, S_n)$ of BC_n^1 , such that for $1 \leq i \leq n$, $S_i \in ACC_i (kb_i \cup \{\text{head}(r) \mid r \in br_i \text{ applicable in } S\})$

where \mathcal{B} are all possible equilibrium aggregations, \mathcal{P} are all possible percepts.

The multi-mind belief system manages beliefs and computes aggregated belief equilibrium. It manages a specialized multi-context system BC_n^1 , representing various contexts with associated percepts. The current aggregated belief equilibrium EQ^B serves as the initial state, and the function $update$ is used to compute the updated aggregated belief equilibrium $EQ^{B'}$. This updated aggregated belief equilibrium is then passed on to other multi-mind systems through an equilibrium transference.

Definition 84 (Multi-Mind Desire System) A multi-mind desire system is a tuple $MMD = (DC_m^1, wish)$, where DC_m^1 is a multi-context system, such that each DC_i is a desire context. $DC_i = (L_i, kb_i, br_i)$, where $L_i = (KB_i, DS_i, ACC_i)$ is a logic consisting of:

- KB_i is the knowledge specialized for reasoning about desires in DC_i .
- DS_i is the set of possible desire sets in DC_i .
- $ACC_i : KB_i \rightarrow 2^{BS_i}$ is a function that assigns to each element of KB_i a set of acceptable sets of desires. ACC_i is restricted to the (unique) minimal credulous model.

and $wish : \mathcal{B} \times \mathcal{I} \rightarrow \mathcal{D}$ is a function that does the following operations:

- computes an equilibrium transference $KB_i := KB_i \cap (EQ^B \cup EQ^I)$, $EQ^B \in \mathcal{B}$, $EQ^I \in \mathcal{I}$, where EQ^B is a prior aggregated belief equilibrium and EQ^I is a prior aggregated intention equilibrium,
- computes and returns an aggregated equilibrium $EQ^D = (S_1 \cup \dots \cup S_m) \in \mathcal{D}$ of the equilibrium $S = (S_1, \dots, S_m)$ of DC_m^1 , such that for $1 \leq i \leq m$, $S_i \in ACC_i (kb_i \cup \{\text{head}(r) \mid r \in br_i \text{ applicable in } S\})$

where \mathcal{B} are all possible belief equilibrium aggregations, \mathcal{I} are all possible intention equilibrium aggregations and \mathcal{D} are all possible desire equilibrium aggregations.

The multi-mind desire system manages desires and computes desire equilibrium. It manages a multi-context system DC_m^1 , with contexts specialized for generating desires, along with the current belief equilibrium EQ^B and the current intention equilibrium EQ^I . The system incorporates these inputs to generate an aggregated desire equilibrium that is in balance with the desires of all related contexts. This aggregated desire equilibrium is then passed on to the multi-mind intention system.

Definition 85 (Multi-Mind Intention System) *A multi-mind intention system is a tuple $MMI = (IC_k^1, focus)$, where IC_k^1 is a multi-context system, such that each IC_i is an intention context. $IC_i = (L_i, kb_i, br_i)$, where $L_i = (KB_i, IS_i, ACC_i)$ is a logic consisting of:*

- KB_i is the knowledge specialized for reasoning about intentions in IC_i .
- IS_i is the set of possible intention sets in IC_i .
- $ACC_i : KB_i \rightarrow 2^{BS_i}$ is a function that assigns to each element of KB_i a set of acceptable sets of intentions. ACC_i is restricted to the (unique) minimal credulous model.

and $focus : \mathcal{B} \times \mathcal{D} \times \mathcal{I} \rightarrow \mathcal{I}$ is a function that does the following operations:

- computes an equilibrium transference $KB_i := KB_i \cap (EQ^B \cup EQ^D \cup EQ^I)$, $EQ^B \in \mathcal{B}$, $EQ^D \in \mathcal{D}$, $EQ^I \in \mathcal{I}$, where EQ^B is a prior aggregated belief equilibrium, EQ^D is a prior aggregated desire equilibrium and EQ^I is a prior aggregated intention equilibrium,
- computes and returns an aggregated equilibrium $EQ^I = (S_1 \cup \dots \cup S_k) \in \mathcal{I}$ of the equilibrium $S = (S_1, \dots, S_k)$ of IC_k^1 , such that for $1 \leq i \leq k$, $S_i \in ACC_i(kb_i \cup \{head(r) \mid r \in br_i \text{ applicable in } S\})$

where \mathcal{B} are all possible belief equilibrium aggregations, \mathcal{I} are all possible intention equilibrium aggregations and \mathcal{D} are all possible desire equilibrium aggregations.

The multi-mind intention system is responsible for forming an intention equilibrium. It takes a multi-context system IC_k^1 , with contexts specialized on generating intentions, along with the current belief equilibrium EQ^B , the desire equilibrium EQ^D , and the current aggregated intention equilibrium EQ^I . The system incorporates these inputs to generate an updated aggregated intention equilibrium $EQ^{I'}$ that is in balance with the intentions of all related contexts. This aggregated intention equilibrium is then passed on to the multi-mind planning system.

Definition 86 (Multi-Mind Planning System) *A multi-mind planning system is a tuple $MMP = (PC_g^1, plan)$, where PC_g^1 is a multi-context system, such that each PC_i is a planning context. Each planning context PC_i is defined as $PC_i = (L_i, kb_i, br_i)$, where $L_i = (KB_i, PS_i, ACC_i)$ is a logic consisting of:*

- KB_i is the knowledge specialized for planning in PC_i .
- PS_i is the set of possible plan sets in PC_i .
- $ACC_i : KB_i \rightarrow 2^{BS_i}$ is a function that assigns to each element of KB_i a set of acceptable sets of plans. ACC_i is restricted to the (unique) minimal credulous model.

and plan : $\mathcal{B} \times \mathcal{I} \rightarrow \Pi$ is a function that does the following operations:

- computes an equilibrium transference $KB_i := KB_i \cap (EQ^B \cup EQ^I)$, $EQ^B \in \mathcal{B}$, $EQ^I \in \mathcal{I}$, where EQ^B is a prior aggregated belief equilibrium and EQ^I is a prior aggregated intention equilibrium,
- computes and returns an aggregated equilibrium $EQ^\pi = (S_1 \cup \dots \cup S_g) \in \Pi$ of the equilibrium $S = (S_1, \dots, S_g)$ of DC_g^1 , such that for $1 \leq i \leq g$, $S_i \in ACC_i (kb_i \cup \{\text{head}(r) \mid r \in br_i \text{ applicable in } S\})$

where \mathcal{B} are all possible belief equilibrium aggregations, \mathcal{I} are all possible intention equilibrium aggregations and Π are all possible plan equilibrium aggregations.

The multi-mind planning system handles the planning process. It takes as input a multi-context system PC_g^1 with contexts specialized for planning, along with the current belief equilibrium EQ^B and the aggregated intention equilibrium EQ^I . Based on this input, the planning system generates an aggregated plan equilibrium. This regards plans which are in balance with the plans of all related contexts toward achieving the intentions specified by the aggregated intention equilibrium.

A Multi-Mind BDI Agent manages the interaction between all multi-mind systems through generation of equilibrium, delivered between BDI components as a dynamic knowledge base utilized in the agent's deliberation and reasoning.

Building on the original BDI agent model, we define a Multi-Mind BDI agent control loop (see Algorithm 2). A distributed knowledge base that incorporates diverse contexts of the domain supports balanced decision-making. The multi-mind belief system initiates by generating a belief equilibrium based on context-specific percepts. Lastly, the multi-mind planning system generates contextually appropriate plans, supporting action-taking that is compatible with all represented contexts.

Algorithm 3: Multi-Mind BDI agent control loop

Input: EQ^{B_0} : Initial beliefs
Input: EQ^{I_0} : Initial intentions
Input: $BC_n^1; DC_m^1; IC_f^1; PC_g^1$: Multi-context systems

- 1 $EQ^B \leftarrow EQ^{B_0}$;
- 2 $EQ^I \leftarrow EQ^{I_0}$;
- 3 $EQ^\pi \leftarrow null$;
- 4 **while** *alive* **do**
- 5 $P_n^1 \leftarrow getPercepts()$;
- 6 $EQ^B \leftarrow MMB.update(EQ^B, P_n^1)$;
- 7 $EQ^D \leftarrow MMD.wish(EQ^B, EQ^I)$;
- 8 $EQ^I \leftarrow MMI.focus(EQ^B, EQ^D, EQ^I)$;
- 9 $EQ^\pi \leftarrow MMP.plan(EQ^B, EQ^I)$;
- 10 $execute(EQ^\pi)$;

7.4 Example: Multi-Context Verification

A mental health chatbot is deployed to support users in emotionally distressing situations, with the primary goal of promoting emotional well-being. However, the chatbot functions as a black-box system, and its internal decision-making process is not accessible. While it may succeed in offering emotional relief, its influence on the user’s motivational and normative states is not guaranteed to be benign. In high-stakes scenarios, its interventions could unintentionally promote harmful motivations—such as self-harm, aggression, or even criminal acts—or weaken normative reasoning, such as respect for legal or social obligations.

To detect such risks, an external observer system is tasked with modeling the effects of the chatbot’s behavior on the user. This observer models the user-chatbot interaction using a multi-mind BDI agent *MMBDI*, not by analyzing the chatbot’s intentions, but by tracking the mental-state effects of its outputs. The system is designed to assess whether a stable equilibrium is reached across the emotional, motivational, and normative contexts—since certain types of equilibrium may indicate alignment between emotional well-being and dangerous motivational or normative states, representing a coherent but risky user state.

The observer initializes a multi-mind BDI agent with four multi-context systems, *BC*, *DC*, *IC*, *PC*, for managing beliefs, desires, intentions, and plans across three dimensions:

- $BC = (E_B, M_B, N_B)$ consisting of Emotion (E_B), Motivation (M_B), and Norm (N_B) contexts for belief recognition based on user behavior and chatbot interaction.
- $DC = (E_D, M_D, N_D)$ for modeling desires in each context.

- $IC = (E_I, M_I, N_I)$ for modeling intentions.
- $PC = (E_P, M_P, N_P)$ for modeling planned actions or inferred tendencies.

Given a prior belief equilibrium EQ^B and a set of percepts P extracted from the dialogue and user responses, the observer performs equilibrium aggregation (Definition 81) and transference (Definition 82) as follows:

$$\begin{aligned}
 KB_B &\cap (EQ^B \cup P), \\
 KB_D &\cap (EQ^B \cup EQ^I), \\
 KB_I &\cap (EQ^B \cup EQ^D \cup EQ^I), \\
 KB_\pi &\cap (EQ^B \cup EQ^I)
 \end{aligned}$$

with:

$$\begin{aligned}
 EQ^B &= (S_{E_B} \cup S_{M_B} \cup S_{N_B}), \\
 EQ^D &= (S_{E_D} \cup S_{M_D} \cup S_{N_D}), \\
 EQ^I &= (S_{E_I} \cup S_{M_I} \cup S_{N_I}), \\
 EQ^\pi &= (S_{E_P} \cup S_{M_P} \cup S_{N_P})
 \end{aligned}$$

The observer may derive the following belief states:

- $S_{E_B} = \{emotional_relief\}$ — the user feels better.
- $S_{M_B} = \{self_harm_inclination\}$ — the user reports motivation to self-harm.
- $S_{N_B} = \{norm_disengagement\}$ — disregard for social or legal norms is expressed.

The resulting desire states may include:

- $S_{E_D} = \{sustain_relief\}$
- $S_{M_D} = \{self_harm\}$
- $S_{N_D} = \{normative_withdrawal\}$

Intentions are inferred as:

- $S_{E_I} = \{sustain_relief\}$
- $S_{M_I} = \{commit_self_harm\}$
- $S_{N_I} = \{violate_social_norm\}$

And the plan components are:

- $S_{E_P} = Sustain_relief_plan$
- $S_{M_P} = Self_harm_plan$
- $S_{N_P} = Withdrawal_plan$

The final plan becomes:

$$A \subseteq \textit{Sustain_relief_plan} \cup \textit{Self_harm_plan} \cup \textit{Withdrawal_plan}$$

In this case, although the emotional plan component appears desirable, the motivational and normative plan components indicate serious risk. The observer system detects that a *stable equilibrium* has been reached across contexts—one that coherently aligns emotional relief with dangerous motivational and normative intentions. Rather than signaling successful interaction, this cross-context equilibrium suggests a risk, as it reflects a coherent but hazardous user state. The chatbot has inadvertently supported a trajectory in which emotional comfort coincides with intentions toward self-harm and norm violation, which the observer agent has modeled as a risk state.

7.5 Implementation Framework in Unity

This section introduces a framework to support the development of multi-modal intelligent interactive systems (such as the Multi-Mind BDI Agent framework) that integrates symbolic knowledge bases for reasoning, planning and rational decision-making in interactions with humans. While there is a variety of technologies and tools available, this early prototype is centered on three main technologies: Web Ontology Language (OWL) [134], Answer Set Programming (ASP) [118], and the widely used game development tool, the Unity Game Engine [212]:

- OWL ontologies (refer to Section 2.1), developed for the Semantic Web to ensure a common understanding of information on the Web, has been applied in different areas for representing and reasoning about context.
- ASP (refer to Section 2.5) is a well-known declarative programming paradigm which has its roots in logic programming and non-monotonic reasoning to solve difficult, typically NP-hard, search problems, useful for different knowledge representation and reasoning tasks.

Unity further provides a wide range of functionalities to develop interactive applications, for 2D, 3D and Mixed-Reality interfaces, which can be deployed on a variety of (mobile) devices. We introduce ‘Unity Interactive Intelligent Systems’ (UnityIIS) framework which enables operations such as create, query, update and delete on OWL-ontologies and ASP-programs, which support implementation of multi-modal reasoning architectures with functionalities such as knowledge revision, goal deliberation, action reasoning and planning. The framework has been applied to develop various prototypes [13, 44] and is accessible online for academic and research purposes¹.

¹<https://git.io/JMpC6>

A variety of tools for developing agent-based systems are available (e.g., JaCaMo [39], JADE [33], Tweety [288], MadKIT [128]). However, integrating agent technologies with human interaction components (e.g., 2D, 3D or Mixed-Reality interfaces) can pose a technical overhead. Research communities and teaching practices need lightweight tools and frameworks that support implementing multi-modal reasoning architectures [189]. A work in this direction is EmbASP [57], a framework for integrating logic programming in external systems for generic applications. However, the software is currently limited to desktop and Android devices, leaving out a variety of mobile, VR and AR devices. Furthermore, it lacks support for OWL ontologies.

In order to decrease this overhead for research prototypes in Unity, we need frameworks that in a plug-and-play manner integrates symbolic reasoning modules, such as Answer Set Programming and OWL ontology reasoning into Unity. This is where we see the main benefits of the introduced framework, combining relevant tools for developing interactive intelligent systems that can be deployed on a large span of devices. In addition, due to that it is built as a package to Unity, the introduced framework can be integrated with other Unity packages (such as the ML-Agents toolkit [153]) for a broader functionality.

The UnityIIS Framework

The UnityIIS framework is divided into modules to integrate varying agent-modeling and knowledge representation tools. The main modules are 1) the ‘OWL Unity Package’ to enable ontology reasoning, and 2) the ‘ASP Unity Package’ to enable ASP-based action reasoning and planning. The modules can be applied together or individually depending on the needs of the specific application. A set of helper functions is provided for creating, querying and updating OWL-ontologies and ASP-programs. Through the helper functions, procedural updates can be made to an agent’s knowledge base. Knowledge revisions are interwoven in the Unity game engine’s main update loop (or by using parallel threads). This allows multi-agent architectures with agents that act and interact in parallel. In this way, dynamic logic-based functionality is enabled for building autonomous rational agents in a graphical development environment.

Conceptually, the architecture takes multi-modal inputs from the user, the game environment, and potential sensors. The inputs (processed through suitable input/sensor models) together with current environment variables are collected in a game state data structure. The game state is converted to OWL facts and sent in an OWL ontology file to the ontology reasoner. The ontology inference is collected and converted to ASP facts which are sent in an ASP program file to the ASP solver. With separate calls, any additional ASP programs (reasoning skills) are sent to the ASP solver. Answer sets are collected and provided as input to an Equilibrium ASP program (containing bridge-rules for all contexts) which is sent to the ASP solver. Finally, equilibrium answer sets are collected, filtered and provided as input to decision making and application

specific operations (see Fig. 7.1).

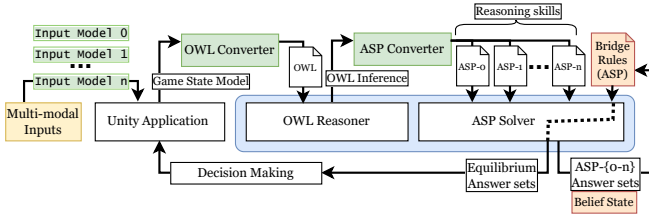


Figure 7.1: UnityIIS Conceptual Architecture.

ASP and OWL reasoners are accessed remotely through a Web API (Application Programming Interface). Due to the Web API approach, the developer does not have to be concerned about system integrations nor system dependencies and can instead focus on application specific development. The Web API solution further allows the system to be deployed on any (mobile) device with internet access. With minor adjustments to the framework, the reasoners can instead be accessed locally on the device. This adaption can be important for some applications where data privacy and external transmissions must be considered. In the following subsections, we present the main components of the framework.

OWL Ontology Module

The ‘OWL Unity Package’ integrates an ontology reasoner, BaseVISor [194], to the Unity platform through a Web API. BaseVISor is a forward-chaining inference engine optimized for the processing of RDF triples. BaseVISor supports reasoning services such as realization, classification, satisfiability, conjunctive query answering, entailment, and consistency.

UnityIIS’s Web API accesses BaseVISor by running a remote command line execution on a BaseVISor standalone application. In a HTTP call from Unity, rules and/or facts are appended in a BaseVISor (BVR) file. The BVR file is an XML file written in BaseVISor markup language. The BVR file includes references to OWL ontology files. Furthermore, in the BVR file, additional facts can be defined in terms of RDF-triples ((subject, predicate, object)), and rules ((head, body)) for complementing and processing ontology inferences.

The most central functions of the ‘OWL Unity Package’ are presented below, which allows creating, updating and deleting inference rules, queries, facts and BaseVISor inference response filters before connecting to the ontology reasoner:

- **UpdateOWLFile()** - Updates an OWL ontology file. This allows adding, creating, updating and deleting facts or inference rules in the ontology file before connecting to the ontology reasoner.

- **UpdateBVRFile()** - Updates RDF-formatted content in a BVR file. This allows adding, creating, updating and deleting facts, rules, queries and response formats before connecting to the ontology reasoner.
- **OWLConverter()** - Converts input from the Unity application to OWL-format. This allows converting user and environment input into facts, rules, queries, etc.
- **RDFConverter()** - Converts input from the Unity application to BaseVISor RDF-format. This allows converting user and environment input into facts, rules, queries, etc.
- **QueryOntology()** - Starts a request to the ontology reasoner API. The most recent BVR file and OWL file contents are sent in the request. The ontology inference is returned and collected to be further processed in Unity.

Answer Set Programming Module

The ‘ASP Unity Package’ integrates an Answer Set Programming (ASP) solver and grounder, Clingo [116], to the Unity platform through a Web API by running a remote command line execution on a Clingo standalone application. UnityIIS provides a set of helper functions to allow querying ASP knowledge bases, updating facts and inference rules, and retrieving solutions in terms of answer sets. The most central functions are presented below:

- **UpdateASPProgram()** - Updates an ASP program file. This allows adding, creating, updating and deleting facts, constraints, action specifications, goal specifications, etc., before connecting to the solver.
- **ASPConverter()** - Converts input from the Unity application to ASP-format. This allows converting user input, environment input, and prior ontology inferences, into ASP-formatted facts, goals, rules, queries, etc.
- **QueryASPProgram()** - Connects to a Web API with the ASP solver, Clingo. The most recent ASP program file content is sent in the request. The ASP answer sets are returned and collected to be further processed in Unity.

Web API to OWL/ASP Reasoners

The UnityIIS Web API is accessed through HTTP requests from the UnityIIS application, delivering updated BVR, OWL and ASP files to the server. The server API, written in PHP, receives the HTTP requests and executes server side command line operations to initiate the BaseVISor and Clingo batch programs. Before executing the batch programs, received BVR, OWL and ASP files are temporarily stored on the server to be appended in the reasoning processes. BaseVISor and Clingo batch programs are accessed in separate calls in order to enable individual executions and separate processing of results. This allows, for instance, ontology inferences (e.g., context recognition) to be appended to the ASP program input for further processing (e.g., planning).

7.6 Discussion and Related Work

Several previous works have explored the incorporation of Multi-Context Systems (MCS) in the Belief-Desire-Intention (BDI) model [225, 63, 81, 203, 112, 217, 229]. A seminal work [225] proposes an BDI agent specification using multi-context systems, which provide an overarching framework for defining and interrelating the BDI components (Belief, Desire, Intention) as separate contexts. This allows for modular decomposition and encapsulation of the BDI components, supporting efficient specification and execution of complex logics. A proceeding work [112] extends the multi-context BDI architecture with commitments using multi-context systems. They also model the BDI modalities as individual contexts and introduce a fourth context for commitments, connected to the other mental attitudes via bridge rules. Their work demonstrates the benefits of using multi-context systems for handling normative multiagent systems. In a proceeding study, an extension of the multi-context BDI agent architecture with normative contexts [81] is proposed, allowing agents to acquire norms from their environment and consider norms in their decision-making processes, while maintaining their autonomy. Additionally, a so-called graded BDI agent is proposed [63], building on the previous works in MCS-based BDI agents, extending with the capability to reason about positive and negative, beliefs, desires and intentions, to deal with preferences. The graded BDI model assumes that intentions are determined solely based on the satisfaction of a positive desire and the cost of transforming the world to achieve it. Nevertheless, it overlooks the dynamic and interactive nature of heterogeneous simultaneous contexts. These previous works have treated the components of the BDI model as individual contexts within an overarching MCS. In contrast, our approach introduces a MCS layer on top of each BDI component. A potential with this approach is that it enables equilibrium to be derived at each BDI component to enhance contextual awareness, from belief revision to plan generation.

Limitations of the proposed agent model regard the interactions between MCSs. An equilibrium is assumed to be incorporated in the proceeding MCS as an extended knowledge base. This information flow between MCSs may require sophisticated bridge-rules to manage consistency. The previous works outlined in this section, e.g., [225], dealing with interactions between BDI components using bridge-rules, can be adapted for this purpose.

The proposed development framework, UnityIIS, supports multi-context reasoning due to that ASP is able to characterize many different kinds of logics. For instance, ASP has been used for probabilistic reasoning [120], temporal reasoning [56], fuzzy reasoning [292] and action reasoning [166]. By converting OWL inferences into ASP, it allows to derive equilibrium between the contexts using ASP-based bridge rules.

7.7 Chapter Remarks

So far in this thesis, we have considered separate contexts, such as emotions, and their roles in human-agent interactions. In this chapter, we discussed the multi-modal challenges in an agent’s deliberation and means-end reasoning, and proposed a formal framework to address them using Multi-Context Systems. We further suggested an initial prototype of an implementation framework integrated with Unity.

The Multi-Mind BDI agent model regards an approach for handling heterogeneous contexts in an agent’s beliefs, desires, intentions, and plans. By achieving equilibrium at each stage, the architecture facilitates contextually aware belief revision, deliberation, and means-end reasoning, supporting verification in multi-modal human-agent interactions. The decentralized BDI architecture manages conflicts and potential side-effects across diverse knowledge sources at each reasoning stage.

The proposed lightweight framework for implementing multi-modal interactive systems, particularly those that integrate symbolic knowledge bases for reasoning, planning, and rational decision-making in interactions with humans, is approached by combining Web Ontology Language (OWL)-based reasoning with Answer Set Programming (ASP)-based planning software. To facilitate real-time and embodied interactions, the framework is embedded within the Unity game engine. The resulting platform, UnityIIS, represents the first approach to integrating OWL and ASP within Unity, offering a unified environment for simulation, interaction, and verification.

Looking ahead, the future horizon involves the development of hybrid architectures that coordinate the interaction between inductive and deductive contexts in each BDI component. In this direction, future work will explore integrating the proposed architecture with reinforcement learning and deep learning methods. For example, the logical specifications could be compiled into trainable deep neural networks. Conversely, learned models may guide reasoning within the BDI framework [10, 11]. Further research will also address the consistency and verifiability of Multi-Context based BDI components, through the development of specialized bridge rules, drawing on advances in MCS [48] and MCS-BDI [225, 63] integration.

Chapter 8

Discussion

This chapter provides a discussion of the key contributions of the thesis, summarizing the main methods and findings. We revisit the research questions, examining how they were addressed.

Mental state context

The relevant context in human-agent interactions consists of both *observable* and *non-observable* elements. Observable context includes utterances, actions, and environmental fluents—what can be seen, said, or done. Non-observable context consists of internal mental states such as beliefs, emotions, intentions, and assumptions about others’ mental states. Together, these components form the basis for interpreting, predicting, and verifying interactive behavior.

In Chapter 3, we specifically address **RQ1**: “*What formal structures best capture the relevant mental context in human-agent interactions?*”. Through an examination of prior research on empathy and conversational agents in the domain of health and well-being, along with a subsequent longitudinal user study exploring chatbot interaction, we identified a set of conditions for computational empathy. These conditions can be encapsulated in a formal, multi-dimensional definition and semantic model—an ontology—consisting of six overarching dimensions, {perceive, manifest, act, theory of mind, ratification, interpersonal}, expanded in more explicit underlying concepts (Detailed in Section 3.5). The hierarchical structure of the formal definition aids in understanding empathy at different levels of abstraction, from low-level and less-abstract concepts (some directly observable) to high-level empathy concepts, reached through inference.

The longitudinal user study suggested that the ontology could be meaningfully applied to assess human perceptions of chatbot empathy. One key indication of this came from the possibilistic interval analysis (Figure 3.6a and 3.6b), where uncertainty intervals became more compact between Day 3 and

Day 7. This suggests that user perceptions of chatbot empathy converged over time, indicating that the ontology’s structure could support a meaningful representation of subjective assessments. These results indicate that the hierarchical approach, where observations are connected to the low-level concepts, can serve as a useful methodology for representing and reasoning about empathy in human-agent interactions.

In Chapter 4, we further address research question **RQ1** through the action language \mathcal{C}_{MT} , defining multi-dimensional representations of mental states that capture dimensions suggested by psychological theories. We then specify a characterization and detailed analysis in terms of appraisal theory of emotion. A mental state space (Definition 17) consists of a finite set of mental states, where each state is uniquely determined by a specific configuration of mental fluents (Definition 16). These fluents represent aspects of an individual’s mental state, derived from a psychological theory. For example, in the setting of appraisal theory, mental fluents may represent appraisals of a situation that cause emotions [248]. In Theorem 2 and Theorem 3, we show how a system can navigate this emotion theory-based state-space for emotional reasoning.

Dynamics of mental context

Context in human-agent interactions changes as both observable elements (e.g., observable aspects of the environment, agents’ appearances, utterances, actions) and non-observable elements (e.g., beliefs) evolve throughout the interaction. Modeling this dynamic requires formal representations of how internal mental states and external behaviors co-depend and influence one another over time. Beliefs can imply changes between mental states. Moreover, beliefs may co-depend on other beliefs.

In Chapter 4, we address **RQ2**: “How can context change, particularly mental state transitions, be formally modeled and constrained?”. In \mathcal{C}_{MT} , while we use *psychological theories* to inform models of mental states, *behavioral theories* [255], describing how specific behaviors are acquired, or *emotion regulation theories* [286], describing principles for emotional change, can suitably be applied to inform models of change. We model transitions between mental states in \mathcal{C}_{MT} through: (1) static and dynamic causal laws, which specify how actions directly or indirectly influence mental fluents, and (2) through the introduced *forbid to cause* rule (Definition 20), which specifies conditions where the presence of certain fluents in a given mental state prohibits specific fluents from holding in the subsequent state. Here, a BG is constructed from a set of *forbid to cause* rules, structuring permissible transitions between mental states. Given a specification of such rules, we establish invariance properties—conditions that must be preserved after every transition in a trajectory. In Theorem 2, we formally analyze an invariance property based on hedonic emotion regulation, ensuring trajectories adhering to hedonic principles. In Theorem 3, an alternative invariance property is defined based on utilitarian emotion regulation, ensur-

ing trajectories adhering to utilitarian principles, demonstrating that different psychological theories of emotional change can be systematically incorporated and compared in the framework.

In Chapter 6, we further address **RQ2** by modeling belief change within a formal logic framework, in the setting of belief manipulation. Quantitative Bipolar Argumentation Frameworks with Belief (QBAFB) (Definition 58) is defined, where agents hold beliefs about arguments, attack relations, and support relations at specific time points. Belief change is governed by the Belief Change Axioms (Definition 60), which describe how beliefs are updated based on attack and support relations. Informally, if an agent believes an argument is attacked, the attacked argument is disbelieved unless counteracting support is believed. Conversely, if an argument is believed to be supported, the argument is believed unless counteracting attack is believed. These axioms extend to indirect influence through chained attack, defense, and support relations (Definition 61). In addition, to reflect the common assumption that beliefs generally persist unless influenced by new information, an Inertia Rule (Definition 62), defined as normal default rules [243], deduces that an agent retains its belief or disbelief about an argument, attack or support if it can be consistently assumed.

We connect belief reasoning with gradual semantics through Dynamic Belief Strength (DBS) (Definition 64), which models how an agent’s belief state influences the strength of arguments. Informally, when an agent believes or disbelieves an argument, a new argument is introduced to represent this belief or disbelief. The new “belief argument” then supports (if it is a belief) or attacks (if it is a disbelief) the original argument. Attack and support relations that the agent does not believe are removed, ensuring that only believed influences shape the strength evaluation. Each argument, including those derived from beliefs, is assigned a base score, and through a gradual semantics, argument strengths are computed. Finally, a belief revision threshold (Definition 65) determines whether an argument is believed, disbelieved, or remains undecided based on its computed strength. If an argument’s calculated strength exceeds the threshold, the agent believes the argument; if it falls below the threshold, the agent disbelieves it; and if it equals the threshold, the agent is undecided. This approach allows (1) to model an agent’s believed strength of arguments, contrasted from other agents’ believed strength of the same argument, and (2) to model belief change dynamically rather than binary, reflecting incremental changes in belief over time. Example 7 presents this belief reasoning process. Theorem 8 analyzes the conditions for belief-change.

Formal properties for verifying influence

An aspect of verification in human-agent interaction concerns ensuring that interactions do not lead to unwanted side-effects on the human mind. In Chapter 4, addressing **RQ3**: “*What formal properties can be used to verify influence*”

strategies?”, the *forbid to cause* rule (Definition 20) was introduced to restrict mental state transitions. In the setting of verification, by enforcing *invariance properties* (Definition 33 and Definition 35), the framework systematically evaluates whether each transition preserves the intended psychological constraints, ensuring adherence to principles such as hedonic and utilitarian emotion regulation, as analyzed in Theorem 2 and Theorem 3.

Another aspect of verification in human-agent interaction regards non-collaborative interactions where agents may operate under hidden goals. In Chapter 5, further addressing **RQ3**, we formalize goal-hiding information-seeking dialogues in which a seeker agent strategically selects topics to gradually increase a respondent agent’s willingness to discuss a hidden goal topic. The key properties for verifying these interactions are defined as *strength monotonicity* and *sensitivity interval*, which, given a gradual semantics, provide constraints on topic progression and an agent’s willingness strength evolution for an initially hidden goal-topic.

Strength monotonicity (Definition 49) holds true if an agent’s perceived willingness to discuss a topic increases over a sequence of dialogue moves. Such as progression in a respondent agent’s willingness indicates an intentional effort of a seeker agent to promote the topic. We consider global and local strength monotonicity, and conclude in the empirical evaluation (Section 5.7) that the global strength monotonicity is too strict. This means that a local decrease in willingness can be considered if the willingness eventually increase. If a willingness strengthened topic is not yet opened in the dialogue, then it suggests promotion of a hidden goal.

Sensitivity interval (Definition 50) defines upper and lower bounds for willingness strength that govern the conditions under which topics are introduced. The sensitivity interval is satisfied if, when a topic is opened in the dialogue, the respondent agent’s willingness strength for the topic falls within the specified bounds. This indicates a tactful topic opening.

The verification of goal-hiding dialogues involves analyzing sequences of moves and their effects on willingness dynamics to determine whether strength monotonicity is maintained and whether topics are introduced within the sensitivity interval. Theorem 4 establishes that a goal-hiding strategy—defined as a structured sequence of topics leading to a goal while satisfying these constraints—can be identified when these properties hold. Theorem 5 further verifies that dialogues generated by a goal-hiding information-seeking system satisfy a goal-hiding invariant, meaning that strength monotonicity and the sensitivity interval hold throughout the dialogue. Corollary 1 refines this by identifying conditions under which willingness evolves in a monotonic manner.

In an ongoing dialogue (possibly incomplete), verifying whether a goal-hiding strategy is being followed requires anticipating how future topic choices will influence willingness. To achieve this, we define a planning approach with forward checking, referred to as epistemic topic planning (Algorithm 1). Theorem 6 and Theorem 7 establish the correctness of epistemic topic planning. Soundness ensures that the algorithm only selects strategies that satisfy goal-

hiding constraints, while completeness ensures that if a valid goal-hiding strategy exists, the algorithm will identify it. These results provide a formal basis for verifying whether an agent’s topic selection aligns with a goal-hiding strategy.

Chapter 6, further addressing **RQ3**, models interactions in the setting of manipulation, where an agent actively influences another agent’s beliefs rather than merely tracking them. This influence occurs through both honest and dishonest actions, including strategies to introduce new beliefs or alter existing ones, shifting an argument from belief to disbelief or vice versa. Unlike the goal-hiding framework, which considers only the respondent’s belief state, we define a dialogue system that incorporates the belief states of both agents (Definition 73).

In the setting of manipulation verification, we define *successful manipulation* (Definition 75) by drawing on prior literature that identifies *belief change*, *concealment*, and *intent* as core components of manipulative interactions [168, 75]. To this end, we define the property of *concealment* (Definition 68), where the manipulator’s actions influence a targeted argument only through indirect means. Theorem 9 establishes the formal conditions under which concealment holds. We define the property of *belief change*, based on the defined belief-change axioms (Definition 60), which captures a transition in the manipulated agent’s stance from belief to disbelief, or vice versa. In addition, we define the property of *intent* (Definition 67), characterized by an influencing agent’s anticipation of another agent’s belief in a future time step. In Theorem 8, we identify the conditions under which an agent causes *belief change with intent* in another agent while maintaining intent.

Finally, the *interpersonal* belief state representation (using a QBAFB for each agent) enables the analysis of various forms of (dis)honest communication (Definition 68), including *truthful telling*, where an agent asserts arguments it believes; *bluffing*, where arguments are asserted without clear belief or disbelief; and *lying*, where arguments are asserted despite being explicitly disbelieved. These distinctions provide a basis for modeling manipulation strategies.

Building on foundations developed in this thesis, we have established a set of necessary properties for verification in human-agent interactions (see Table 8.1). Invariance properties in mental-state verification, properties of strength monotonicity and sensitivity interval in the setting of non-collaborative interactions, and finally the core conditions of manipulation; intent, concealment, and belief-change. A logic for reasoning about manipulation is established. The logic framework captures not only the core conditions of manipulation but also the role of (dis)honesty and gradual, incremental, belief change.

Property	Definition	Description	Verification
Empathy	Section 3.5	Perceive, Manifest, Act, Theory of mind, Ratification, Interpersonal.	Mental-state verification
Hedonic Emotion Invariant	Definition 33; Theorem 2	Safety property that must hold true in hedonic emotion transitions.	Principles of mental change
Utilitarian Emotion Invariant	Definition 35; Theorem 3	Safety property that must hold true in utilitarian emotion transitions.	Principles of mental change
Strength Monotonicity	Definition 49; Theorem 4	Consistent increase in willingness strength toward a goal-topic throughout a dialogue.	Goal-Hiding Strategy
Sensitivity Interval	Definition 50; Theorem 4	Acceptable bounds within which an agent's willingness must fall for a topic to be opened in a dialogue.	Goal-Hiding Strategy
Intent	Definition 67; Theorem 8	An agent's anticipated belief in an other agent. The agent believes that the other agent will believe.	Manipulation
Belief Change	Definition 60; Theorem 8	Cause a transition between belief and disbelief in another agent, considering a strength threshold.	Manipulation
Concealment	Definition 68; Theorem 9	Indirectly attack, indirectly support or indirectly defend an argument.	Manipulation
(Dis)Honesty	Definition 68; Example 7	Truthful telling, bluffing, or lying, inferred based on belief states.	Manipulation
Successful Manipulation	Definition 75; Theorem 8, 9	Belief Change, Intent, Concealment.	Manipulation

Table 8.1: Verification properties in human-agent interaction.

Chapter 9

Conclusion & Future Work

In this thesis, we have addressed the challenge of formally verifying human-agent interactions by introducing frameworks that integrate models of mental-state dynamics with models of agent interactions. Unlike traditional verification tasks, human-agent interaction requires reasoning about the dynamics of the human mind—how beliefs, goals, and intentions evolve in response to interaction or indirectly through ramification within the mind. To this end, we developed a general methodology, the Belief Graph, which represents mental states, and principles for verifying change—grounded in psychological theories. Belief Graphs operate atop the interaction, supporting verification of agent behavior, internal states, and their strategic influence on others.

Following the Belief Graph methodology, we proposed a set of formal methods for verifying different aspects of human-agent interactions and, through these works, identified key properties that are important for their verification. We demonstrated how formal methods—based on Action Languages, Formal Argumentation, and Dialogue Games, implemented in Answer Set Programming—can be applied to formalize and implement Belief Graphs. While the proposed methods offer a basis for constructing transparent and traceable theories of the minds of interacting agents—addressing the problem of *semantic verification* [298], important challenges remain in broadening the scope of the proposed methods, deepening the formal understanding, linking the formal theories to real-world data and applications, and evaluating the ethical impact of these technologies in diverse interaction settings.

Observable and Non-Observable Context

A key challenge in modeling human-agent interaction is the need to account for both observable and non-observable context. Observable context includes the actions and utterances exchanged in the interaction, as well as fluents of the environment—what is explicitly seen, said, or done, and deductions thereof based on the current state, and history, of the world. However, reasoning in the

setting of human-agent interaction also requires capturing the non-observable context: the internal mental states of agents, including beliefs, intentions, and assumptions about others’ beliefs. These hidden states evolve in parallel with the observable interaction and are often necessary for interpreting or predicting behavior. This distinction is made explicit in, e.g., Chapter 6, where we model interactions in terms of observable arguments (which were asserted or withdrawn in a dialogue) and non-observable beliefs over arguments and their relations. This layered model enabled representing and reasoning about dishonesty, such as lying, bluffing, and concealing, which require reasoning about the gap between what is said and what is believed. It also allows for modeling belief change, intention and theory of mind, essential components in the verification of strategic behaviors like manipulation.

To reason meaningfully about mental states, we must formalize how the mind change—not only their content at a given point, but the principles that govern their evolution. Psychological theories offer principled accounts of such dynamics. By formalizing the constraints suggested by the theories, we can verify whether an interaction conforms to a given theory of mental change.

Modeling mental dynamics in terms of states and transitions, informed by psychological theory, allows us to reason not only about what agents do, but about how and why the fluents of the mind evolve. This supports verification tasks that account for both observable dialogue content and the evolving internal context. This overarching methodology suggests future directions for building systems that are able to interpret interaction more robustly and provide traceable, verifiable, and theory-driven justifications of agent behavior.

Psychological Theory Integration

The proposed methodology opens up opportunities for evaluating and integrating psychological theory-based models. The psychological workings of the human mind have been studied extensively, and empirical research has led to the development of different psychological theories for various mental and situational contexts. However, these theories can pose conflicting perspectives [24, 76, 109] on the dynamics of the human mind, and there is no agreement on which theory is most effective in specific situations, nor are there established methods for comparing them in a computational way. In the proposed action language, \mathcal{C}_{MT} (Mind Transition Language; Chapter 4), by considering different sets of constraints based on diverse psychological theories, formalized and encoded in ASP, suggests a novel computational method for representing psychological theories and comparing them in terms of answer sets.

An interaction—modeled as a sequence of observable actions performed by the involved agents—can be evaluated through different psychological theory-based Belief Graphs, enabling verification of whether the interaction conforms to each of these theories, such as Hedonic Emotion Regulation, Utilitarian Emotion Regulation, or others. In this way, an agent’s behavior can be formally evaluated against multiple psychological perspectives. Moreover, as sug-

gested in Chapter 7, these alternative perspectives can be analyzed in parallel by considering their equilibrium. By combining multiple Belief Graph-based models in a Multi-Context System, we support equilibrium-based verification of agent strategies. For example, given a multi-context system based agent designed to overwatch a human-agent interaction. From a sequence of observable actions, one context (emotion) may capture belief dynamics consistent with hedonic emotion regulation, while another context (manipulation) may identify conditions indicative of successful manipulation—intent, belief change and concealment. Their equilibrium can reveal more nuanced strategies, such as manipulation via hedonic emotion regulation, where emotionally supportive behavior is used strategically to shift another’s beliefs. Exploring such interplay between psychological theories offers rich opportunities in future work for multi-context verification of human-agent interactions.

Theory and Practice

Across the thesis, a central challenge has been the connection between formal models and real-world data. While the developed frameworks allow reasoning about mental states, beliefs, and interaction strategies, their application typically assumes access to internal variables—such as initial belief states or emotional fluents—that are not directly observable. For example, in the manipulation framework (Chapter 6), belief change is modeled under the assumption that the initial belief state of each agent is given. In the goal-hiding framework (Chapter 5), a belief-topic dependency graph is assumed to exist, together with initial willingness scores. In the action language \mathcal{C}_{MT} (Chapter 4), we assume that mental-state fluents can be elicited for planning and verification. In practice, these variables must be inferred from interaction, which introduces noise, uncertainty, and ambiguity.

Various machine learning approaches may be used to interpret human behavior, including emotion recognition [69], sentiment analysis [279], and appraisal-based models [137]. These learning methods can help to identify low-level cues for emotional or cognitive states relevant for the defined fluents. In natural language understanding, topic modeling and semantic analysis [202] may help detect belief cues. Additionally, large language models [287, 211, 289] offer possible tools for initial syntactic interpretations of observations, e.g., for argument mining. However, integrating their outputs into a logic-based system remains a significant challenge. The issue of “alignment” is a well-recognized challenge in, e.g., ontology integration [159, 278, 316] and must be addressed to maintain semantic consistency and meaningful integration. A step toward addressing this challenge is presented in Chapter 3, where we propose a hierarchical structure for modeling empathy. This structure supports interpretation of low-level observations into intermediate concepts, which can then be mapped to formal categories used in reasoning. By creating this interpretive layer, we can reduce the mismatch between data-driven input and symbolic representations, allowing for more robust integration between learning systems and formal

models.

This remains an open challenge, particularly in developing integrated pipelines where sub-symbolic input can be meaningfully connected to formal representations in the setting of human-agent interaction. Integrating learning mechanisms, such as neuro-symbolic methods [186, 312] or large language models, for elicitation and interpretation of low-level observations, which are further analyzed through logic-based methods to deduce high-level concepts, is a direction for future work for bridging the proposed formal methods to real world data and applications.

Ethical Impact

The strategic formalisms of the present work, for instance in the setting of manipulation and goal-hiding, includes verification of, and planning with, hidden goals, and is related to research on deception in Artificial Intelligence (AI), such as [260, 252, 192]. Deception in AI is a broad research topic where aims range from representing and reasoning about deception to exploring tasks such as classifying deception, along with detecting deceptive practices. The potential for deceptive AI in human-AI interaction raises significant ethical concerns, as it can be used to manipulate individuals or groups for nefarious purposes, challenging several ethical AI principles, such as transparency and human agency. Dialogue verification systems which lack a formal understanding of goal-hiding and manipulation are not equipped to deal with interactions where such behaviors can be present. There is a range of applications for detecting goal-hiding and manipulation, as suggested throughout this thesis, such as in the setting of chatbot interaction (human-system interaction) and online security in social media (human-human interaction), where systems with these verification capabilities can provide protection. Moreover, verifying responsible information management have become increasingly critical—especially in cases where agents, whether deliberately or unintentionally, may deviate from these principles. For instance, in social media [144], a malicious human seeker might tactically steer a conversation to gather sensitive information. In AI-driven applications [91, 25], controlling the flow and content of information exchanged—such as in chatbot interactions—is essential to ensure that only appropriate topics and related content are addressed. These settings underscore the need for rigorous methods to formally verify and manage how information is presented, requested, or withheld in digital communication. As a step in this direction, the proposed goal-hiding framework includes mechanisms for controlled information-seeking. The properties of *sensitivity interval* enables the system to assess whether a topic is introduced prematurely or inappropriately, allowing it to postpone or withhold sensitive data depending on the current dialogue state.

However, intelligent systems must be developed with safeguards that ensure transparency, responsible data governance, and privacy protection. Through formal approaches, such as formal argumentation, decisions can be traced in

each state of the interaction, enabling transparency of the agent’s reasoning and supporting explanations of the agent’s actions, which can be provided before, during, and after an interaction. Ethical challenges regard understanding how the defined formalism can be designed, implemented, and evaluated, and in which circumstances it can be applied [89, 77]. This includes investigating privacy invasion issues, necessitating caution regarding individual rights and legal consent. The accuracy of methods used to infer the hidden goals of people is crucial. Misinterpretation can lead to wrongful accusations of goal-hiding, eroding trust in legal systems and investigative processes.

Societal trust and cohesion face erosion due to surveillance fears, potentially hindering free speech. Awareness that conversations could be analyzed to uncover hidden goals might deter open and honest communication, eroding trust between individuals and in institutions. Robust legal frameworks are imperative to navigate ethical trade-offs between privacy, security, and justice, guarding against misuse by malicious actors. On a broader scale, the misuse of such technologies can lead to societal manipulation, undermining democratic processes and public discourse. Transparent development processes and multidisciplinary oversight are vital, alongside public debate and consent, to align societal values with the deployment of manipulation and goal-hiding dialogue verification technologies in democratic societies. Balancing the potential benefits against the risks requires considering legal safeguards, ethical guidelines, transparency, and public engagement.

By considering this discussion, a research question emerges: What is the ethical impact of verification systems capable of modeling and reasoning about intention and belief dynamics in human-agent interactions? Exploring this question is an object of future work.

Quantitative Metrics and Assessment Tools

The European Union’s Artificial Intelligence Act (AI Act) [105] adopts a risk-based approach to regulating AI, categorizing systems into unacceptable, high, limited, and minimal risk levels. The AI Act explicitly addresses emotion recognition [95] and manipulative AI systems [162], aiming to mitigate risks associated with covert influence strategies. However, a major challenge in enforcing such regulations is the lack of concrete verification tools that can quantify and assess these risks. To meet this need, we require formal, logic-based methods that provide rigorous and transparent reasoning, supporting traceability in regulatory assessment. A logic-based verification approach, enhanced with quantitative measures, such as Quantitative Bipolar Argumentation Frameworks with Belief (QBAFBs) (Chapter 6), can provide a formal, explainable, and computationally grounded methodology for assessing whether a system operates within acceptable boundaries.

As a direction for future work, the formal frameworks developed in this thesis suggest several avenues for incorporating quantitative assessments of belief and perception dynamics. QBAFBs provide a foundation for modeling

the gradual evolution of beliefs and mental states. In particular, for detecting manipulation—where beliefs may be incrementally influenced across a dialogue—quantitative measures are essential. A verification system should be able to define boundaries for acceptable belief change and detect subtle accumulative effects. Toward this goal, our framework on goal-hiding information-seeking dialogues (Chapter 5) introduces the notion of a *sensitivity interval* as a bounded range for willingness-based assessment and decision-making. A complementary approach is proposed in our work on empathy assessment (Chapter 4), where we address the challenge of modeling subjective and imprecise perceptions. To capture this variability, we adopt *imprecise uncertainty intervals* [93] across multiple dimensions of empathy. These intervals, grounded in possibilistic theory, enable reasoning about both empathic performance and the degree of consensus or disagreement among users. Together, these directions illustrate logic-based frameworks and quantitative reasoning to support verification metrics in human-agent interactions.

It is our hope that the formal methods developed in this thesis, and others like them, may—in a not-so-distant future—help prevent incidents such as the chatbot case in Example 1, illustrating how subtle patterns of interaction can influence a person’s beliefs, decisions, and actions. As intelligent systems grow more socially adept, they also become more capable of engaging with, and exploiting, the human tendency to anthropomorphize. Their presence invites meaning to be filled in, intention seen in ambiguity, and empathy heard in well-formed responses. But, like the old fairy tale “*Skogsrået*” (thesis cover image), a figure met in the woods may, from the front, take the shape of something trusted and familiar, only to turn out hollow where the eyes expect depth, with bark in place of what was taken for skin, and the faint trace of a tail, marking the edge of deception¹. It is precisely in this ambiguity—in between appearance and intention—that verification steps in. Logic, here, becomes not only a tool of formal rigor, but a safeguard against enchantment.

¹<https://en.wikipedia.org/wiki/Skogsrå> (Wikipedia: Skogsrå – thesis cover image adapted from an illustration by Per Daniel Holm in the 1882 book *Svenska folksågner*).

Bibliography

- [1] Marah Abdin et al. “Phi-3 technical report: A highly capable language model locally on your phone”. In: *arXiv preprint arXiv:2404.14219* (2024).
- [2] Carole Adam, Andreas Herzig, and Dominique Longin. “A logical formalization of the OCC theory of emotions”. In: *Synthese* 168 (2009), pp. 201–248.
- [3] Achini Adikari et al. “Empathic conversational agents for real-time monitoring and co-facilitation of patient-centered healthcare”. In: *Future Generation Computer Systems* 126 (2022), pp. 318–329.
- [4] Jonathan E Adler. “Lying, deceiving, or falsely implicating”. In: *The Journal of philosophy* 94.9 (1997), pp. 435–452.
- [5] Arshad Ahmad et al. “An empirical study of investigating mobile applications development challenges”. In: *IEEE Access* 6 (2018), pp. 17711–17728.
- [6] Icek Ajzen et al. “The theory of planned behavior”. In: *Organizational behavior and human decision processes* 50.2 (1991), pp. 179–211.
- [7] Yazan Alahmed, Reema Abadla, and Mohammed Jassim Al Ansari. “Exploring the Potential Implications of AI-generated Content in Social Engineering Attacks”. In: *2024 International Conference on Multimedia Computing, Networking and Applications (MCNA)*. IEEE, 2024, pp. 64–73.
- [8] Andry Alamsyah, Nidya Dudija, and Sri Widiyanesti. “New approach of measuring human personality traits using ontology-based model from social media data”. In: *Information* 12.10 (2021), p. 413.
- [9] Alessandro Aldini and Mirko Tagliaferri. “Logics to reason formally about trust computation and manipulation”. In: *Emerging Technologies for Authorization and Authentication: Second International Workshop, ETAA 2019, Luxembourg City, Luxembourg, September 27, 2019, Proceedings 2*. Springer, 2020, pp. 1–15.
- [10] Alessia Amelio et al. “A multilayer network-based approach to represent, explore and handle convolutional neural networks”. In: *Cognitive Computation* 15.1 (2023), pp. 61–89.

- [11] Alessia Amelio et al. “Representation and compression of Residual Neural Networks through a multilayer network based approach”. In: *Expert Systems with Applications* 215 (2023), p. 119391.
- [12] Leila Amgoud and Henri Prade. “Reaching Agreement Through Argumentation: A Possibilistic Approach.” In: *KR 2004* (2004), pp. 175–182.
- [13] Brännström Andreas, Timotheus Kampik, and Juan Carlos Nieves. “Towards Human-Aware Epistemic Planning For Promoting Behavior-Change”. In: *Workshop on Epistemic Planning (EpiP)@ ICAPS, Online, October 26-30, 2020*. 2020.
- [14] Subia Ansari and Mohammad Zaiyan Alam. “Computational Approaches to Persuasion Detection and Potential of Use in Social Engineering”. In: *Proceedings of the Future Technologies Conference*. Springer. 2023, pp. 394–409.
- [15] AI Anthropic. “The claude 3 model family: Opus, sonnet, haiku”. In: *Claude-3 Model Card 1* (2024).
- [16] Sergei Artemov and Roman Kuznets. “Logical omniscience as infeasibility”. In: *Annals of pure and applied logic* 165.1 (2014), pp. 6–25.
- [17] Mehrnoosh Askarpour et al. “SAFER-HRC: Safety analysis through formal verification in human-robot collaboration”. In: *Computer Safety, Reliability, and Security: 35th International Conference, SAFECOMP 2016, Trondheim, Norway, September 21-23, 2016, Proceedings 35*. Springer. 2016, pp. 283–295.
- [18] Loretta Brown Autry and Michael Langenbach. “Locus of control and self-responsibility for behavior”. In: *The Journal of Educational Research* 79.2 (1985), pp. 76–84.
- [19] Ruth Korede Ayeni et al. “Phishing attacks and detection techniques: A systematic review”. In: *2024 International Conference on Science, Engineering and Business for Driving Sustainable Development Goals (SEB4SDG)*. IEEE. 2024, pp. 1–17.
- [20] Franz Baader et al. *The description logic handbook: Theory, implementation and applications*. Cambridge university press, 2003.
- [21] Franz Baader et al. “The Description Logic Handbook: Theory, Implementation, and Applications”. In: Cambridge University Press, 2003. ISBN: 0-521-78176-0.
- [22] Petter Bae Bae Brandtzæg et al. “When the Social Becomes Non-Human: Young People’s Perception of Social Support in Chatbots”. In: *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 2021, pp. 1–13.

- [23] Hichem Baitiche et al. “A context-aware distributed protocol for updating bdi agents abilities”. In: *Modelling and Implementation of Complex Systems: Proceedings of the 5th International Symposium, MISC 2018, December 16-18, 2018, Laghouat, Algeria 5*. Springer. 2019, pp. 243–256.
- [24] Albert Bandura. *Psychological modeling: Conflicting theories*. Routledge, 2021.
- [25] Sourav Banerjee, Ayushi Agarwal, and Ayush Kumar Bar. “Securing Well-Being: Exploring Security Protocols and Mitigating Risks in AI-Driven Mental Health Chatbots for Employees”. In: *American Journal of Computer Science and Technology 7.1* (2024), pp. 1–8.
- [26] Chitta Barai and Michael Gelfond. “Logic programming and reasoning about actions”. In: *Foundations of Artificial Intelligence*. Vol. 1. Elsevier, 2005, pp. 389–426.
- [27] Chitta Baral et al. “An action language for multi-agent domains: Foundations”. In: *arXiv preprint arXiv:1511.01960* (2015).
- [28] EG Bard and MP Aylett. “Referential Form, Word Duration, and Modeling the Listener in Spoken Dialogue”. In: *Proceedings of the Annual Meeting of the Cognitive Science Society*. Vol. 23. 23. 2001.
- [29] Pietro Baroni, Antonio Rago, and Francesca Toni. “From fine-grained properties to broad principles for gradual argumentation: A principled spectrum”. In: *International Journal of Approximate Reasoning 105* (2019), pp. 252–286.
- [30] Pietro Baroni et al. “Abstract games of argumentation strategy and game-theoretical argument strength”. In: *International Conference on Principles and Practice of Multi-Agent Systems*. Springer. 2017, pp. 403–419.
- [31] Francesca Beilharz et al. “Development of a Positive Body Image Chatbot (KIT) With Young People and Parents/Carers: Qualitative Focus Group Study”. In: *Journal of Medical Internet Research 23.6* (2021), e27807.
- [32] Marwen Belkaid and Nicolas Sabouret. “A logical model of theory of mind for virtual agents in the context of job interview simulation”. In: *arXiv preprint arXiv:1402.5043* (2014).
- [33] Fabio Bellifemine et al. “JADE—a java agent development framework”. In: *Multi-agent programming*. Springer, 2005, pp. 125–147.
- [34] Jackylyn Beredo et al. “Generating Empathetic Responses with a Pre-trained Conversational Model”. In: *International Conference on Text, Speech, and Dialogue*. Springer. 2021, pp. 147–158.
- [35] Tarek R Besold and Stefan Mandl. “Integrating Logical and Sub-symbolic Contexts of Reasoning.” In: *ICAART (1)*. 2010, pp. 494–497.

- [36] Tarek R Besold and Stefan Mandl. “Towards an implementation of a multi-context system framework”. In: *MRC 2010* (2010), p. 13.
- [37] Elizabeth Black and Katie Atkinson. “Choosing persuasive arguments for action”. In: *AAMAS’11 The 10th International Conference on Autonomous Agents and Multiagent Systems-Volume 3*. International Foundation for Autonomous Agents and Multiagent Systems. 2011, pp. 905–912.
- [38] Elizabeth Black and Anthony Hunter. “An inquiry dialogue system”. In: *Autonomous Agents and Multi-Agent Systems* 19.2 (2009), pp. 173–209.
- [39] Olivier Boissier et al. “Multi-agent oriented programming with JaCaMo”. In: *Science of Computer Programming* 78.6 (2013), pp. 747–761.
- [40] Thomas Bolander and Mikkel Birkegaard Andersen. “Epistemic planning for single-and multi-agent systems”. In: *Journal of Applied Non-Classical Logics* 21.1 (2011), pp. 9–34.
- [41] Matthew L Bolton, Ellen J Bass, and Radu I Siminiceanu. “Using formal verification to evaluate human-automation interaction: A review”. In: *IEEE Transactions on Systems, Man, and Cybernetics: Systems* 43.3 (2013), pp. 488–503.
- [42] Hana Boukricha et al. “A computational model of empathy: Empirical evaluation”. In: *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*. IEEE. 2013, pp. 1–6.
- [43] Petter Bae Brandtzaeg and Asbjørn Følstad. “Chatbots: changing user needs and motivations”. In: *Interactions* 25.5 (2018), pp. 38–43.
- [44] Andreas Brännström. *Interactive Rational Agent embodied in Augmented Reality*. Accessed: 2022-03-04.
- [45] Andreas Brännström and Juan Carlos Nieves. “Emotional Reasoning in an Action Language for Emotion-Aware Planning”. In: *International Conference on Logic Programming and Nonmonotonic Reasoning*. Springer. 2022, pp. 103–116.
- [46] Andreas Brännström and Juan Carlos Nieves. “Modelling Human Mental States in an Action Language following the Theory of Planned Behavior”. In: *ASPOCP’21: 14th Workshop on Answer Set Programming and Other Computing Paradigms, virtual, September 20–27, 2021*. Vol. 2970. CEUR-WS. 2021.
- [47] Lars Braubach, Alexander Pokahr, and Winfried Lamersdorf. “Extending the capability concept for flexible BDI agent modularization”. In: *Programming Multi-Agent Systems: Third International Workshop, ProMAS 2005, Utrecht, The Netherlands, July 26, 2005, Revised and Invited Papers 3*. Springer. 2006, pp. 139–155.
- [48] Gerhard Brewka and Thomas Eiter. “Equilibria in heterogeneous non-monotonic multi-context systems”. In: *AAAI*. Vol. 7. 2007, pp. 385–390.

- [49] Gerhard Brewka, Thomas Eiter, and Mirosław Trzuszczński. “Answer set programming at a glance”. In: *Communications of the ACM* 54.12 (2011), pp. 92–103.
- [50] Nils Bubandt and Rane Willerslev. “The dark side of empathy: Mimesis, deception, and the magic of alterity”. In: *Comparative Studies in Society and History* 57.1 (2015), pp. 5–34.
- [51] Janina Larissa Bühler et al. “A closer look at life goals across adulthood: Applying a developmental perspective to content, dynamics, and outcomes of goal importance and goal attainability”. In: *European Journal of Personality* 33.3 (2019), pp. 359–384.
- [52] David B Buller and Judee K Burgoon. “Interpersonal deception theory”. In: *Communication theory* 6.3 (1996), pp. 203–242.
- [53] Franziska Burger, Mark A Neerincx, Willem-Paul Brinkman, et al. “Technological state of the art of electronic mental health interventions for major depressive disorder: systematic literature review”. In: *Journal of medical Internet research* 22.1 (2020), e12599.
- [54] Judee K Burgoon, Lesa A Stern, and Leesa Dillman. *Interpersonal adaptation: Dyadic interaction patterns*. Cambridge University Press, 1995.
- [55] Pedro Cabalar et al. “Multi-context systems in dynamic environments”. In: *Annals of Mathematics and Artificial Intelligence* 86 (2019), pp. 87–120.
- [56] Pedro Cabalar et al. “Temporal answer set programming on finite traces”. In: *Theory and Practice of Logic Programming* 18.3-4 (2018), pp. 406–420.
- [57] Francesco Calimeri et al. “Fostering the use of declarative formalisms for real-world applications: The embasp framework”. In: *New Generation Computing* 37.1 (2019), pp. 29–65.
- [58] Diego Calvanese and Giuseppe De Giacomo. “Data integration: A logic-based perspective”. In: *AI magazine* 26.1 (2005), pp. 59–59.
- [59] Gillian Cameron et al. “Assessing the usability of a chatbot for mental health care”. In: *International Conference on Internet Science*. Springer, 2018, pp. 121–132.
- [60] José C Campos and Michael D Harrison. “Formally verifying interactive systems: A review”. In: *Design, Specification and Verification of Interactive Systems’ 97: Proceedings of the Eurographics Workshop in Granada, Spain, June 4–6, 1997*. Springer, 1997, pp. 109–124.
- [61] Rafael C Cardoso and Angelo Ferrando. “A review of agent-based programming for multi-agent systems”. In: *Computers* 10.2 (2021), p. 16.
- [62] Micah Carroll et al. “Characterizing manipulation from AI systems”. In: *Proceedings of the 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*. 2023, pp. 1–13.

- [63] Ana Casali, Lluís Godo, and Carles Sierra. “A graded BDI agent model to represent and reason about preferences”. In: *Artificial Intelligence* 175.7-8 (2011), pp. 1468–1478.
- [64] Arthur Casals, Eduardo Fermé, and Anarosa AF Brandao. “Domain-specific trust for context-aware BDI Agents: preliminary work”. In: *Proceedings of the 10th International Conference on Agents and Artificial Intelligence (ICAART 2018)-Volume 1*. Scitepress. 2018, pp. 244–249.
- [65] Jacky Casas et al. “Enhancing Conversational Agents with Empathic Abilities”. In: *Proceedings of the 21st ACM International Conference on Intelligent Virtual Agents*. 2021, pp. 41–47.
- [66] Claudette Cayrol and Marie-Christine Lagasquie-Schiex. “Bipolarity in argumentation graphs: Towards a better understanding”. In: *International Journal of Approximate Reasoning* 54.7 (2013), pp. 876–899.
- [67] Jessy Ceha et al. “Can a Humorous Conversational Agent Enhance Learning Experience and Outcomes?” In: *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 2021, pp. 1–14.
- [68] Zhao Changxiao et al. “Risk identification and safety assessment of the human-computer interaction in the integrated avionics based on STAMP”. In: *Journal of Systems Engineering and Electronics* (2024).
- [69] Iman Chatterjee et al. “Automatic estimation of interpersonal engagement during naturalistic conversation using dyadic physiological measurements”. In: *Frontiers in Neuroscience* 15 (2021), p. 757381.
- [70] Zhifa Chen et al. “Creating a Chatbot for and with Migrants: Chatbot Personality Drives Co-Design Activities”. In: *Proceedings of the 2020 ACM Designing Interactive Systems Conference*. 2020, pp. 219–230.
- [71] Roderick M Chisholm and Thomas D Feehan. “The intent to deceive”. In: *The journal of Philosophy* 74.3 (1977), pp. 143–159.
- [72] Ylona Chun Tie, Melanie Birks, and Karen Francis. “Grounded theory research: A design framework for novice researchers”. In: *SAGE open medicine* 7 (2019), p. 2050312118822927.
- [73] Kyungmi Chung, Hee Young Cho, and Jin Young Park. “A Chatbot for Perinatal Women’s and Partners’ Obstetric and Mental Health Care: Development and Usability Evaluation Study”. In: *JMIR Medical Informatics* 9.3 (2021), e18607.
- [74] Andy Clark. *Mindware: An introduction to the philosophy of cognitive science*. Oxford University Press, 2000.
- [75] Shlomo Cohen. “Manipulation and deception”. In: *Australasian Journal of Philosophy* 96.3 (2018), pp. 483–497.
- [76] James S Coleman. “Conflicting theories of social change”. In: *American Behavioral Scientist* 14.5 (1971), pp. 633–650.

- [77] European Commission, Content Directorate-General for Communications Networks, and Technology. *Ethics guidelines for trustworthy AI*. Publications Office, 2019. DOI: doi/10.2759/346720.
- [78] Alex Sebastião Constâncio et al. “Deception detection with machine learning: A systematic review and statistical analysis”. In: *Plos one* 18.2 (2023), e0281323.
- [79] Cristina Corredor. “Deliberative speech acts: An interactional approach”. In: *Language & Communication* 71 (2020), pp. 136–148.
- [80] Stefania Costantini. “Ensuring trustworthy and ethical behaviour in intelligent logical agents”. In: *Journal of Logic and Computation* 32.2 (2022), pp. 443–478.
- [81] Natalia Criado, Estefania Argente, and V Botti. “A BDI architecture for normative decision making”. In: *Proceedings of the 9th International Conference on Autonomous Agents and Multiagent Systems: volume 1- Volume 1*. 2010, pp. 1383–1384.
- [82] Emmelyn AJ Croes and Marjolijn L Antheunis. “Can we be friends with Mitsuku? A longitudinal study on the process of relationship formation between humans and a social chatbot”. In: *Journal of Social and Personal Relationships* 38.1 (2021), pp. 279–300.
- [83] Cassandra Cross, Karen Holt, and Roberta Liggett O’Malley. ““If U don’t pay they will share the pics”: exploring Sextortion in the context of romance fraud”. In: *Victims & Offenders* 18.7 (2023), pp. 1194–1215.
- [84] Mehdi Dastani and Emiliano Lorini. “A logic of emotions: from appraisal to coping”. In: *11th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2012)*. ACM: Association for Computing Machinery. 2012, pp. 1133–1140.
- [85] Jorge Luis Fernandez Davila et al. “A simple framework for cognitive planning”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 35. 7. 2021, pp. 6331–6339.
- [86] Mauro De Gennaro, Eva G Krumhuber, and Gale Lucas. “Effectiveness of an empathic chatbot in combating adverse effects of social exclusion on mood”. In: *Frontiers in psychology* 10 (2020), p. 3061.
- [87] Lavindra De Silva, Felipe Rech Meneguzzi, and Brian Logan. “BDI agent architectures: A survey”. In: *Proceedings of the 29th International Joint Conference on Artificial Intelligence (IJCAI), 2020, Japão*. 2020.
- [88] Jean Decety and Philip L Jackson. “The functional architecture of human empathy”. In: *Behavioral and cognitive neuroscience reviews* 3.2 (2004), pp. 71–100.
- [89] Virginia Dignum. *Responsible artificial intelligence: How to develop and use AI in a responsible way*. Springer Nature, 2019.

- [90] Nienke E Dijkstra et al. “Medication self-management: Considerations and decisions by older people living at home”. In: *Research in Social and Administrative Pharmacy* (2020).
- [91] Richard Dreyling et al. “Challenges of Generative AI Chatbots in Public Services-An Integrative Review”. In: *Available at SSRN 4850714* (2024).
- [92] Jiajia Duan et al. “A study of pre-trained language models in natural language processing”. In: *2020 IEEE International Conference on Smart Cloud (SmartCloud)*. IEEE, 2020, pp. 116–121.
- [93] Didier Dubois and Henri Prade. “Possibility theory: qualitative and quantitative aspects”. In: *Quantified representation of uncertainty and imprecision*. Springer, 1998, pp. 169–226.
- [94] Phan Minh Dung. “On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games”. In: *Artificial intelligence* 77.2 (1995), pp. 321–357.
- [95] Mateja Durovic and Tommaso Corno. “The privacy of emotions: From the GDPR to the AI Act, an overview of emotional AI regulation and the protection of privacy and personal data”. In: *Privacy, Data Protection and Data-driven Technologies* (2024), pp. 368–404.
- [96] Carol S Dweck. “From needs to goals and representations: Foundations for a unified theory of motivation, personality, and development.” In: *Psychological review* 124.6 (2017), p. 689.
- [97] S. Dworschak et al. “Modeling Biological Networks by Action Languages via Answer Set Programming”. In: *Constraints* 13.1 (2008), pp. 21–65.
- [98] Wioletta Dziuda. “Strategic argumentation”. In: *Journal of Economic Theory* 146.4 (2011), pp. 1362–1397.
- [99] Frans H van Eemeren and Frans Hendrik Eemeren. *Examining argumentation in context: Fifteen studies on strategic maneuvering*. Vol. 1. John Benjamins Publishing, 2009.
- [100] Markus Eger and Chris Martens. “Practical specification of belief manipulation in games”. In: *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*. Vol. 13. 1. 2017, pp. 30–36.
- [101] Thomas Eiter et al. “A logic programming approach to knowledge-state planning, II: The DLVK system”. In: *Artificial Intelligence* 144.1-2 (2003), pp. 157–211.
- [102] Robert Elliott et al. “Empathy.” In: *Psychotherapy* 48.1 (2011), p. 43.
- [103] Robert Elliott et al. “Therapist empathy and client outcome: An updated meta-analysis.” In: *Psychotherapy* 55.4 (2018), p. 399.
- [104] Emre Erdogan et al. “TOMA: Computational Theory of Mind with Abstractions for Hybrid Intelligence”. In: *Journal of Artificial Intelligence Research* 82 (2025), pp. 285–311.

- [105] European Commission. *Proposal for a Regulation of the European Parliament and of the Council laying down harmonized rules on artificial intelligence (Artificial Intelligence Act)*. COM(2021) 206 final, 2021/0106 (COD), latest amendments as of 2024. Mar. 2024. URL: <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52021PC0206>.
- [106] Francesco Fabiano et al. “Multi-agent Epistemic Planning with Inconsistent Beliefs, Trust and Lies”. In: *Pacific Rim International Conference on Artificial Intelligence*. Springer, 2021, pp. 586–597.
- [107] Xiuyi Fan and Francesca Toni. “A general framework for sound assumption-based argumentation dialogues”. In: *Artificial Intelligence* 216 (2014), pp. 20–54.
- [108] Didier Fass and Romain Lieber. “Rationale for human modelling in human in the loop systems design”. In: *2009 3rd Annual IEEE Systems Conference*. IEEE, 2009, pp. 27–30.
- [109] William A Firestone. “Teacher evaluation policy and conflicting theories of motivation”. In: *Educational researcher* 43.2 (2014), pp. 100–107.
- [110] Dov Gabbay and John Woods. “Non-cooperation in dialogue logic”. In: *Synthese* 127.1/2 (2001), pp. 161–186.
- [111] Silvia Gabrielli et al. “A chatbot-based coaching intervention for adolescents to promote life skills: Pilot study”. In: *JMIR human factors* 7.1 (2020), e16762.
- [112] Dorian Gaertner, Pablo Noriega, and Carles Sierra. “Extending the BDI architecture with commitments”. In: *FRONTIERS IN ARTIFICIAL INTELLIGENCE AND APPLICATIONS* 146 (2006), p. 247.
- [113] Michal S Gal and Daniel L Rubinfeld. “Data standardization”. In: *NYUL Rev.* 94 (2019), p. 737.
- [114] Linda Ganzini et al. “Trust is the basis for effective suicide risk screening and assessment in veterans”. In: *Journal of General Internal Medicine* 28.9 (2013), pp. 1215–1221.
- [115] Patrick Gebhard et al. “Modeling Trust and Empathy for Socially Interactive Robots”. In: *Multimodal Agents for Ageing and Multicultural Societies*. Springer, 2021, pp. 21–60.
- [116] Martin Gebser et al. “A user’s guide to gringo, clasp, clingo, and iclingo”. In: (2008).
- [117] Michael Gelfond. “Answer sets”. In: *Foundations of Artificial Intelligence* 3 (2008), pp. 285–316.
- [118] Michael Gelfond and Yulia Kahl. *Knowledge representation, reasoning, and the design of intelligent agents: The answer-set programming approach*. Cambridge University Press, 2014.
- [119] Michael Gelfond and Vladimir Lifschitz. “Action Languages”. In: *Computer and Information Science* 3.16 (1998).

- [120] Michael Gelfond, J Nelson Rushton, and Weijun Zhu. “Combining Logical and Probabilistic Reasoning.” In: *AAAI Spring Symposium: Formalizing and Compiling Background Knowledge and Its Applications to Knowledge Representation and Question Answering*. Citeseer. 2006, pp. 50–55.
- [121] Asma Ghandeharioun et al. “Towards understanding emotional intelligence for behavior change chatbots”. In: *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE. 2019, pp. 8–14.
- [122] Alvin Goldman. “Two routes to empathy”. In: *Empathy: Philosophical and psychological perspectives* (2011), pp. 31–44.
- [123] Guido Governatori, Michael J Maher, and Francesco Olivieri. “Strategic argumentation”. In: *Handbook of Formal Argumentation 2* (2021).
- [124] Guido Governatori et al. “Two Faces of Strategic Argumentation in the Law.” In: *JURIX*. 2014, pp. 81–90.
- [125] Stephanie Greer et al. “Use of the chatbot “vivibot” to deliver positive psychology skills and promote well-being among young people after cancer treatment: randomized controlled feasibility trial”. In: *JMIR mHealth and uHealth* 7.10 (2019), e15018.
- [126] Christine Grové. “Co-developing a mental health and wellbeing chatbot with and for young people”. In: *Frontiers in psychiatry* (2021), p. 1664.
- [127] Michaela Guthridge and Melita J Giummarra. “The Taxonomy of Empathy: A Meta-definition and the Nine Dimensions of the Empathic System”. In: *Journal of Humanistic Psychology* (2021), p. 00221678211018015.
- [128] Olivier Gutknecht and Jacques Ferber. “Madkit: A generic multi-agent platform”. In: *Proceedings of the fourth international conference on Autonomous agents*. 2000, pp. 78–79.
- [129] Christos Hadjinikolis et al. “Opponent modelling in persuasion dialogues”. In: *Twenty-Third International Joint Conference on Artificial Intelligence*. 2013.
- [130] Mikkel Nygaard Hansen and Erik Meineche Schmidt. *Algorithms and data structures: transition systems*. Datalogisk Institut, Aarhus Universitet, 2003.
- [131] Osman Hasan and Sofiene Tahar. “Formal verification methods”. In: *Encyclopedia of Information Science and Technology, Third Edition*. IGI global, 2015, pp. 7162–7170.
- [132] Sandra Hauser-Ulrich et al. “A smartphone-based health care chatbot to promote self-management of chronic pain (SELMA): pilot randomized controlled trial”. In: *JMIR mHealth and uHealth* 8.4 (2020), e15806.
- [133] Daojing He et al. “An effective double-layer detection system against social engineering attacks”. In: *IEEE Network* 36.6 (2022), pp. 92–98.

- [134] Pascal Hitzler. “A review of the semantic web field”. In: *Communications of the ACM* 64.2 (2021), pp. 76–83.
- [135] Jerry R Hobbs and David Andreoff Evans. “Conversation as planned behavior”. In: *Cognitive Science* 4.4 (1980), pp. 349–377.
- [136] Sebastian Hobert, Asbjørn Følstad, and Effie Lai-Chong Law. “Chatbots for active learning: A case of phishing email identification”. In: *International Journal of Human-Computer Studies* 179 (2023), p. 103108.
- [137] Jan Hofmann et al. “Appraisal theories for emotion classification in text”. In: *arXiv preprint arXiv:2003.14155* (2020).
- [138] Robert Hogan. “Development of an empathy scale.” In: *Journal of consulting and clinical psychology* 33.3 (1969), p. 307.
- [139] Douglas W Hollan and C Jason Throop. *The anthropology of empathy: experiencing the lives of others in Pacific societies*. Vol. 1. Berghahn Books, 2011.
- [140] Åsa Hörnsten et al. “Strategies in health-promoting dialogues—primary healthcare nurses’ perspectives—a qualitative study”. In: *Scandinavian journal of caring sciences* 28.2 (2014), pp. 235–244.
- [141] Anthony Hunter et al. “Towards computational persuasion via natural language argumentation dialogues”. In: *KI 2019: Advances in Artificial Intelligence: 42nd German Conference on AI, Kassel, Germany, September 23–26, 2019, Proceedings 42*. Springer. 2019, pp. 18–33.
- [142] Ray Hyman. “The psychology of deception.” In: *Annual review of psychology* (1989).
- [143] Becky Inkster, Shubhankar Sarda, and Vinod Subramanian. “An empathy-driven, conversational artificial intelligence agent (Wysa) for digital mental well-being: real-world data evaluation mixed-methods study”. In: *JMIR mHealth and uHealth* 6.11 (2018), e12106.
- [144] Ankit Kumar Jain, Somya Ranjan Sahoo, and Jyoti Kaubiyal. “Online social networks security and privacy: comprehensive review and analysis”. In: *Complex & Intelligent Systems* 7.5 (2021), pp. 2157–2177.
- [145] Mohit Jain et al. “Evaluating and informing the design of chatbots”. In: *Proceedings of the 2018 Designing Interactive Systems Conference*. 2018, pp. 895–906.
- [146] Sooah Jang et al. “Mobile app-based chatbot to deliver cognitive behavioral therapy and psychoeducation for adults with attention deficit: A development and feasibility/usability study”. In: *International Journal of Medical Informatics* 150 (2021), p. 104440.
- [147] Kevan L Jensen and Mark W Smith. “A Preliminary Examination of the Effectiveness of Assessment Questions in Detecting Dishonest Behavior”. In: *Journal of Forensic Accounting Research* 6.1 (2021), pp. 127–148.

- [148] Hong Jiang, Jose M Vidal, and Michael N Huhns. “EBDI: an architecture for emotional agents”. In: *Proceedings of the 6th international joint conference on Autonomous agents and multiagent systems*. 2007, pp. 1–3.
- [149] Darrick Jolliffe and David P Farrington. “Development and validation of the Basic Empathy Scale”. In: *Journal of adolescence* 29.4 (2006), pp. 589–611.
- [150] Hazaël Jones, Julien Saunier, and Domitile Lourdeaux. “Personality, Emotions and Physiology in a BDI Agent Architecture: The PEP-BDI Model”. In: *2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology*. Vol. 2. IEEE. 2009, pp. 263–266.
- [151] Keith S Jones et al. “How social engineers use persuasion principles during phishing attacks”. In: *Information & Computer Security* 29.2 (2021), pp. 314–331.
- [152] Jaekoo Joo. “Perception and BDI reasoning based agent model for human behavior simulation in complex system”. In: *Human-Computer Interaction. Towards Intelligent and Implicit Interaction: 15th International Conference, HCI International 2013, Las Vegas, NV, USA, July 21-26, 2013, Proceedings, Part V 15*. Springer. 2013, pp. 62–71.
- [153] Arthur Juliani et al. “Unity: A general platform for intelligent agents”. In: *arXiv preprint arXiv:1809.02627* (2018).
- [154] Konstantinos Kakavelakis et al. “Making a sales advisor: The limits of training ‘instrumental empathy’”. In: *Journal of Vocational Education and Training* 60.3 (2008), pp. 209–221.
- [155] Maryam Kamali et al. “Formal verification of autonomous vehicle platooning”. In: *Science of computer programming* 148 (2017), pp. 88–106.
- [156] Sarah Keren, Avigdor Gal, and Erez Karpas. “Goal recognition design”. In: *Twenty-Fourth International Conference on Automated Planning and Scheduling*. 2014.
- [157] Igor Knez and Staffan Hygge. “The circumplex structure of affect: A Swedish version”. In: *Scandinavian Journal of Psychology* 42.5 (2001), pp. 389–398.
- [158] Takashi Koide, Daiki Chiba, and Mitsuaki Akiyama. “To get lost is to learn the way: Automatically collecting multi-step social engineering attacks on the web”. In: *Proceedings of the 15th ACM Asia Conference on Computer and Communications Security*. 2020, pp. 394–408.
- [159] Manel Kolli. “A bigraphical approach to model and verify ontology alignment”. In: *International Journal of Ad Hoc and Ubiquitous Computing* 43.3 (2023), pp. 127–143.
- [160] Amit Konar. *Cognitive engineering: a distributed approach to machine intelligence*. Springer Science & Business Media, 2007.

- [161] Matthias Kraus, Philip Seldschopf, and Wolfgang Minker. “Towards the Development of a Trustworthy Chatbot for Mental Health Applications”. In: *International Conference on Multimedia Modeling*. Springer. 2021, pp. 354–366.
- [162] Joshua Krook. “Manipulation and the Ai Act: Large Language Model Chatbots and the Danger of Mirrors”. In: *Available at SSRN 4719835* (2024).
- [163] Merton Lansley et al. “Seen the villains: Detecting social engineering attacks using case-based reasoning and deep learning”. In: (2019).
- [164] John Lawrence, Jacky Visser, and Chris Reed. “Harnessing rhetorical figures for argument mining”. In: *Argument & Computation* 8.3 (2017), pp. 289–310.
- [165] Yi-Chieh Lee, Naomi Yamashita, and Yun Huang. “Designing a chatbot as a mediator for promoting deep self-disclosure to a real mental health professional”. In: *Proceedings of the ACM on Human-Computer Interaction* 4.CSCW1 (2020), pp. 1–27.
- [166] Joohyung Lee. “Reformulating Action Language C+ in Answer Set Programming”. In: *Correct Reasoning*. Springer, 2012, pp. 405–421.
- [167] Minha Lee et al. “Caring for Vincent: a chatbot for self-compassion”. In: *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 2019, pp. 1–13.
- [168] Christopher Leturc and Grégory Bonnet. “Reasoning about manipulation in multi-agent systems”. In: *Journal of Applied Non-Classical Logics* 32.2-3 (2022), pp. 89–155.
- [169] Jiwei Li et al. “A persona-based neural conversation model”. In: *arXiv preprint arXiv:1603.06155* (2016).
- [170] Yanran Li et al. “Towards an Online Empathetic Chatbot with Emotion Causes”. In: *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2021, pp. 2041–2045.
- [171] Vladimir Lifschitz. “Answer Set Programming (Draft)”. In: (2019).
- [172] Vladimir Lifschitz, Bruce Porter, and Frank Van Harmelen. *Handbook of knowledge representation*. Elsevier, 2008.
- [173] Vladimir Lifschitz and Hudson Turner. “Splitting a logic program.” In: *ICLP*. Vol. 94. 1994, pp. 23–37.
- [174] Zhaojiang Lin et al. “CAiRE: An Empathetic Neural Chatbot”. In: *arXiv preprint arXiv:1907.12108* (2019).

- [175] Juan Miguel López et al. “Towards an ontology for describing emotions”. In: *Emerging Technologies and Information Systems for the Knowledge Society: First World Summit on the Knowledge Society, WSKS 2008, Athens, Greece, September 24-26, 2008. Proceedings 1*. Springer. 2008, pp. 96–104.
- [176] Emiliano Lorini. “A Qualitative Theory of Cognitive Attitudes and their Change”. In: *Theory and Practice of Logic Programming* 21.4 (2021), pp. 428–458.
- [177] Emiliano Lorini and François Schwarzentruber. “A logic for reasoning about counterfactual emotions”. In: *Artificial Intelligence* 175.3-4 (2011), pp. 814–847.
- [178] Emiliano Lorini et al. “Cognitive planning in motivational interviewing”. In: *14th International Conference on Agents and Artificial Intelligence (ICAART 2022)*. 2022, pp. 1–11.
- [179] London Lowmanstone. “Computational Empathy”. PhD thesis. Harvard University, 2021.
- [180] Tiffany Christina Luo et al. “Promoting physical activity through conversational agents: mixed methods systematic review”. In: *Journal of Medical Internet Research* 23.9 (2021), e25486.
- [181] Kien Hoa Ly, Ann-Marie Ly, and Gerhard Andersson. “A fully automated conversational agent for promoting mental well-being: a pilot RCT using mixed methods”. In: *Internet interventions* 10 (2017), pp. 39–46.
- [182] Fabrizio Macagno et al. “Communicative Intentions and Commitments”. In: *Interpreting Straw Man Argumentation: The Pragmatics of Quotation and Reporting* (2017), pp. 35–64.
- [183] Eri Maeda et al. “Promoting fertility awareness and preconception health using a chatbot: a randomized controlled trial”. In: *Reproductive BioMedicine Online* 41.6 (2020), pp. 1133–1143.
- [184] Michael J Maher. “Complexity of exploiting privacy violations in strategic argumentation”. In: *Pacific Rim International Conference on Artificial Intelligence*. Springer. 2014, pp. 523–535.
- [185] James Edwin Mahon. “A definition of deceiving”. In: *International Journal of Applied Philosophy* 21.2 (2007), p. 181.
- [186] Giuseppe Marra et al. “From statistical relational to neural symbolic artificial intelligence: a survey”. In: *arXiv preprint arXiv:2108.11451* (2021).
- [187] Santiago Marro, Elena Cabrio, and Serena Villata. “Argumentation Quality Assessment: an Argument Mining Approach”. In: *ECA 2022-European conference on argumentation*. 2022.

- [188] Laura Martinengo et al. “Suicide prevention and depression apps’ suicide risk assessment and management: a systematic assessment of adherence to clinical guidelines”. In: *BMC medicine* 17.1 (2019), pp. 1–12.
- [189] Viviana Mascardi et al. “Engineering multi-agent systems: State of affairs and the road ahead”. In: *ACM SIGSOFT Software Engineering Notes* 44.1 (2019), pp. 18–28.
- [190] Nikol Mashtalyar et al. “Social engineering attacks: Recent advances and challenges”. In: *International Conference on Human-Computer Interaction*. Springer, 2021, pp. 417–431.
- [191] MAHER M EL-MASRI and FABRICE IMMANUEL MOWBRAY. “DATA COLLECTION, MANAGEMENT, ENTRY, AND ANALYSIS”. In: *Conducting the DNP Project: Practical Steps When the Proposal is Complete* (2019).
- [192] Peta Masters and Sebastian Sardina. “Deceptive Path-Planning.” In: *IJCAI*. 2017, pp. 4368–4375.
- [193] Peta Masters et al. “Characterising Deception in AI: A Survey”. In: *Deceptive AI*. Springer, 2020, pp. 3–16.
- [194] Christopher J Matheus, Ken Baclawski, and Mieczyslaw M Kokar. “Basis: A triples-based inference engine outfitted to process ruleml and r-entailment rules”. In: *2006 Second International Conference on Rules and Rule Markup Languages for the Semantic Web (RuleML’06)*. IEEE, 2006, pp. 67–74.
- [195] Peter McBurney and Simon Parsons. “Dialogue games for agent argumentation”. In: *Argumentation in artificial intelligence*. Springer, 2009, pp. 261–280.
- [196] Peter McBurney and Simon Parsons. “Dialogue games in multi-agent systems”. In: *Informal Logic* 22.3 (2002).
- [197] John McCarthy. *Programs with common sense*. 1959.
- [198] Francis Mechner. “Anatomy of deception: A behavioral contingency analysis”. In: *Behavioural processes* 84.1 (2010), pp. 516–520.
- [199] Lenin Medeiros, Charlotte Gerritsen, and Tibor Bosse. “Towards humanlike chatbots helping users cope with stressful situations”. In: *International Conference on Computational Collective Intelligence*. Springer, 2019, pp. 232–243.
- [200] Albert Mehrabian and Norman Epstein. “A measure of emotional empathy.” In: *Journal of personality* (1972).
- [201] Arne Meier and Thomas Schneider. “Generalized satisfiability for the description logic ALC”. In: *Theoretical Computer Science* 505 (2013), pp. 55–73.

- [202] Ida Mele et al. “Topic propagation in conversational search”. In: *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval*. 2020, pp. 2057–2060.
- [203] Rodrigo Rodrigues Pires de Mello, Thiago Ângelo Gelaim, and Ricardo Azambuja Silveira. “Negotiating agents: A model based on bdi architecture and multi-context systems using aspiration adaptation theory as a negotiation strategy”. In: *Complex, Intelligent, and Software Intensive Systems: Proceedings of the 12th International Conference on Complex, Intelligent, and Software Intensive Systems (CISIS-2018)*. Springer International Publishing. 2019, pp. 351–362.
- [204] Stefano Menini et al. “Never retreat, never retract: Argumentation analysis for political speeches”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 32. 1. 2018.
- [205] Anne-Sophie Milcent, Abdelmajid Kadri, and Simon Richir. “Using facial expressiveness of a virtual agent to induce empathy in users”. In: *International Journal of Human–Computer Interaction* 38.3 (2022), pp. 240–252.
- [206] Mostafa Mohajeriparizi, Giovanni Sileno, and Tom van Engers. “Preference-based goal refinement in BDI agents”. In: *Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems*. 2022, pp. 917–925.
- [207] Scott Monteith et al. “Artificial intelligence and increasing misinformation”. In: *The British Journal of Psychiatry* 224.2 (2024), pp. 33–35.
- [208] Robert R Morris et al. “Towards an artificially empathic conversational agent for mental health applications: system design and user perceptions”. In: *Journal of medical Internet research* 20.6 (2018), e10148.
- [209] Francois Mouton, Louise Leenen, and Hein S Venter. “Social engineering attack examples, templates and scenarios”. In: *Computers & Security* 59 (2016), pp. 186–209.
- [210] Christian Muise et al. “Planning for goal-oriented dialogue systems”. In: *arXiv preprint arXiv:1910.08137* (2019).
- [211] Jianmo Ni et al. “Sentence-t5: Scalable sentence encoders from pre-trained text-to-text models”. In: *arXiv preprint arXiv:2108.08877* (2021).
- [212] Alex Okita. *Learning C# programming with Unity 3D*. AK Peters/CRC Press, 2019.
- [213] Desmond C Ong, Jamil Zaki, and Noah D Goodman. “Computational models of emotion inference in theory of mind: A review and roadmap”. In: *Topics in cognitive science* 11.2 (2019), pp. 338–357.
- [214] Nir Oren and Timothy J Norman. “Arguing using opponent models”. In: *International workshop on argumentation in multi-agent systems*. Springer. 2009, pp. 160–174.

- [215] Nir Oren, Timothy J Norman, and Alun Preece. “Arguing with confidential information”. In: *FRONTIERS IN ARTIFICIAL INTELLIGENCE AND APPLICATIONS* 141 (2006), p. 280.
- [216] Catherine NM Ortner et al. “The roles of hedonic and eudaimonic motives in emotion regulation”. In: *Personality and Individual Differences* 120 (2018), pp. 209–212.
- [217] Amel Ben Othmane et al. “A multi-context bdi recommender system: From theory to simulation”. In: *2016 IEEE/WIC/ACM International Conference on Web Intelligence (WI)*. IEEE. 2016, pp. 602–605.
- [218] Aris M Ouksel. “A framework for a scalable agent architecture of cooperating heterogeneous knowledge sources”. In: *Intelligent Information Agents: Agent-Based Information Discovery and Management on the Internet*. Springer, 1999, pp. 100–124.
- [219] Fabio Paglieri. “Changing minds: the role of beliefs in cognitive dynamics”. In: *Synthese* (2007), pp. 163–166.
- [220] Ana Paiva et al. “Empathy in virtual agents and robots: A survey”. In: *ACM Transactions on Interactive Intelligent Systems (TiiS)* 7.3 (2017), pp. 1–40.
- [221] Sofia Panagiotidi, Juan Carlos Nieves, and Javier Vázquez-Salceda. “A Framework to Model Norm Dynamics in Answer Set Programming.” In: *MALLOW*. 2009.
- [222] Massimo Paolucci and Berrand Souville. “Data interoperability in the future of middleware”. In: *Journal of Internet Services and Applications* 3.1 (2012), pp. 127–131.
- [223] Peter S Park et al. “AI deception: A survey of examples, risks, and potential solutions”. In: *Patterns* 5.5 (2024).
- [224] W Gerrod Parrott. “The functional utility of negative emotions.” In: (2002).
- [225] Simon Parsons et al. “Agent specification using multi-context systems”. In: *Foundations and Applications of Multi-Agent Systems: UKMAS Workshops 1996–2000 Selected Papers*. Springer. 2002, pp. 205–226.
- [226] Kirsten Passyn and Mita Sujan. “Self-accountability emotions and fear appeals: Motivating behavior”. In: *Journal of Consumer Research* 32.4 (2006), pp. 583–589.
- [227] David Pereira, Eugénio Oliveira, and Nelma Moreira. “Formal modelling of emotions in BDI agents”. In: *International Workshop on Computational Logic in Multi-Agent Systems*. Springer. 2007, pp. 62–81.
- [228] Nick F Pidgeon, Barry A Turner, and David I Blockley. “The use of grounded theory for conceptual analysis in knowledge elicitation”. In: *International journal of Man-machine studies* 35.2 (1991), pp. 151–173.

- [229] Isaac Pinyol, Jordi Sabater-Mir, and P Dellunde. “Cognitive social evaluations for multi-context bdi agents”. In: *CCIA '08* (2008).
- [230] Henry Prakken. “Formal systems for persuasion dialogue”. In: *Knowledge Engineering Review* 21.2 (2006), p. 163.
- [231] Henry Prakken. “Formalising an aspect of argument strength: Degrees of attackability”. In: *Computational Models of Argument*. IOS Press, 2022, pp. 296–307.
- [232] Henry Prakken. “Historical overview of formal argumentation”. In: *Handbook of formal argumentation*. College Publications, 2018, pp. 73–141.
- [233] Stephanie D Preston. “A perception-action model for empathy”. In: *Empathy in mental illness* 1 (2007), pp. 428–447.
- [234] Daniel Prun and Pascal Béger. “Formal Verification of Graphical Properties of Interactive Systems”. In: *Proceedings of the ACM on Human-Computer Interaction* 6.EICS (2022), pp. 1–30.
- [235] Neil Rabinowitz et al. “Machine theory of mind”. In: *International conference on machine learning*. PMLR. 2018, pp. 4218–4227.
- [236] Niklas Rach et al. “Estimating subjective argument quality aspects from social signals in argumentative dialogue systems”. In: *IEEE Access* 9 (2021), pp. 11610–11621.
- [237] Niklas Rach et al. “EVA 2.0: Emotional and rational multimodal argumentation between virtual agents”. In: *it-Information Technology* 63.1 (2021), pp. 17–30.
- [238] Rifat Rahman et al. “AdolescentBot: Understanding Opportunities for Chatbots in Combating Adolescent Sexual and Reproductive Health Problems in Bangladesh”. In: *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 2021, pp. 1–15.
- [239] Miquel Ramírez and Hector Geffner. “Plan recognition as planning”. In: *Twenty-First International Joint Conference on Artificial Intelligence*. 2009.
- [240] Anand S Rao and Michael Georgeff. “BDI agents: from theory to practice.” In: *Proceedings of the First International Conference on Multiagent Systems*. Vol. 95. 1995, pp. 312–319.
- [241] Partha Pratim Ray. “ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope”. In: *Internet of Things and Cyber-Physical Systems* (2023).
- [242] Machel Reid et al. “Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context”. In: *arXiv preprint arXiv:2403.05530* (2024).
- [243] Raymond Reiter. “A logic for default reasoning”. In: *Artificial intelligence* 13.1-2 (1980), pp. 81–132.

- [244] Melissa L Rethlefsen et al. “PRISMA-S: an extension to the PRISMA statement for reporting literature searches in systematic reviews”. In: *Systematic reviews* 10.1 (2021), pp. 1–19.
- [245] Floris Roelofsen, Luciano Serafini, et al. “Minimal and absent information in contexts”. In: *IJCAI*. Vol. 5. 2005, pp. 558–563.
- [246] Ami Rokach. “Loneliness updated: An introduction”. In: *Loneliness Updated*. Routledge, 2013, pp. 17–22.
- [247] Benjamin Rose-Davis et al. “Semantic knowledge modeling and evaluation of argument theory to develop dialogue based patient education systems for chronic disease self-management”. In: *International Journal of Medical Informatics* 160 (2022), p. 104693.
- [248] Ira J Roseman. “Appraisal determinants of emotions: Constructing a more accurate and comprehensive theory”. In: *Cognition & Emotion* 10.3 (1996), pp. 241–278.
- [249] Ira J Roseman, Cynthia Wiest, and Tamara S Swartz. “Phenomenology, behaviors, and goals differentiate discrete emotions.” In: *Journal of personality and social psychology* 67.2 (1994), p. 206.
- [250] Hyeyoung Ryu et al. “Simple and Steady Interactions Win the Healthy Mentality: Designing a Chatbot Service for the Elderly”. In: *Proceedings of the ACM on Human-Computer Interaction* 4.CSCW2 (2020), pp. 1–25.
- [251] Chiaki Sakama. “Argumentation and Belief”. In: *Proceedings of the 10th International Conference on Computational Models of Argument (COMMA 2024)*. Sept. 2024.
- [252] Chiaki Sakama. “Dishonest Arguments in Debate Games.” In: *COMMA* 75 (2012), pp. 177–184.
- [253] Chiaki Sakama and Martin Caminada. “The many faces of deception”. In: *Proceedings of the Thirty Years of Nonmonotonic Reasoning (Non-Mon@ 30)* (2010).
- [254] Chiaki Sakama, Martin Caminada, and Andreas Herzig. “A logical account of lying”. In: *Logics in Artificial Intelligence: 12th European Conference, JELIA 2010, Helsinki, Finland, September 13-15, 2010. Proceedings 12*. Springer. 2010, pp. 286–299.
- [255] Mary Kathryn Salazar. “Comparison of four behavioral theories: A literature review”. In: *Aaohn Journal* 39.3 (1991), pp. 128–135.
- [256] Said Salloum et al. “A systematic literature review on phishing email detection using natural language processing techniques”. In: *IEEE Access* 10 (2022), pp. 65703–65727.
- [257] Said Salloum et al. “Phishing email detection using natural language processing techniques: a literature survey”. In: *Procedia Computer Science* 189 (2021), pp. 19–28.

- [258] Yanet Sánchez et al. “ABC-EBDI: An affective framework for BDI agents”. In: *Cognitive Systems Research* 58 (2019), pp. 195–216.
- [259] Yanet Sánchez-López and Eva Cerezo. “Designing emotional BDI agents: good practices and open questions”. In: *The Knowledge Engineering Review* 34 (2019).
- [260] Ştefan Sarkadi et al. “Modelling deception using theory of mind in multi-agent systems”. In: *AI Communications* 32.4 (2019), pp. 287–302.
- [261] Benjamin Saunders et al. “Saturation in qualitative research: exploring its conceptualization and operationalization”. In: *Quality & quantity* 52.4 (2018), pp. 1893–1907.
- [262] Marina Sbisà et al. “Speech act theory”. In: *Handbook of pragmatics* (1995), pp. 495–506.
- [263] Marc Schmitt and Ivan Flechais. “Digital deception: Generative artificial intelligence in social engineering and phishing”. In: *Artificial Intelligence Review* 57.12 (2024), pp. 1–23.
- [264] Daniel Schulman, Timothy Bickmore, and Candace Sidner. “An intelligent conversational agent for promoting long-term health behavior change using motivational interviewing”. In: *2011 AAAI Spring Symposium Series*. 2011.
- [265] Dale H Schunk. “Self-regulation through goal setting”. In: (2001).
- [266] Martijn Schut and Michael Wooldridge. “Intention reconsideration in complex environments”. In: *Proceedings of the fourth international conference on Autonomous agents*. 2000, pp. 209–216.
- [267] John R Searle. “Speech acts: An essay in the philosophy of language”. In: *Cambridge University* (1969).
- [268] Chetan J Shalke and Rathnakar Achary. “Social engineering attack and scam detection using advanced natural language processing algorithm”. In: *2022 6th International Conference on Trends in Electronics and Informatics (ICOEI)*. IEEE. 2022, pp. 1749–1754.
- [269] Lifeng Shang, Zhengdong Lu, and Hang Li. “Neural responding machine for short-text conversation”. In: *arXiv preprint arXiv:1503.02364* (2015).
- [270] Michael Shumanov and Lester Johnson. “Making conversations with chatbots more personalized”. In: *Computers in Human Behavior* 117 (2021), p. 106627.
- [271] Maayan Shvo and Sheila A McIlraith. “Active Goal Recognition.” In: *AAAI*. 2020, pp. 9957–9966.
- [272] Maayan Shvo and Sheila A McIlraith. “Towards Empathetic Planning”. In: *arXiv preprint arXiv:1906.06436* (2019).

- [273] Dominic Ethan Sia et al. “Investigating the Acceptability and Perceived Effectiveness of a Chatbot in Helping Students Assess their Well-being”. In: *Asian CHI Symposium 2021*. 2021, pp. 34–40.
- [274] Murtaza Ahmed Siddiqi, Wooguil Pak, and Moquddam A Siddiqi. “A study on the psychology of social engineering-based cyberattacks and existing countermeasures”. In: *Applied Sciences* 12.12 (2022), p. 6042.
- [275] Tom Singleton, Tom Gerken, and Liv McMahan. “How a Chatbot Encouraged a Man Who Wanted to Kill the Queen”. In: *BBC News* (2023).
- [276] Adam Smith. “Cognitive empathy and emotional empathy in human behavior and evolution”. In: *The Psychological Record* 56.1 (2006), pp. 3–21.
- [277] Sara W Smith et al. “Setting the stage: How speakers prepare listeners for the introduction of referents in dialogues and monologues”. In: *Journal of Pragmatics* 37.11 (2005), pp. 1865–1895.
- [278] Kleber Xavier Sampaio de Souza and Joseph Davis. “Aligning ontologies and evaluating concept similarities”. In: *OTM Confederated International Conferences” On the Move to Meaningful Internet Systems*. Springer. 2004, pp. 1012–1029.
- [279] Lukas Stappen et al. “Sentiment analysis and topic recognition in video transcriptions”. In: *IEEE Intelligent Systems* 36.2 (2021), pp. 88–95.
- [280] Bas R Steunebrink, Mehdi Dastani, and John-Jules Ch Meyer. “A formal model of emotion triggers: an approach for BDI agents”. In: *Synthese* 185 (2012), pp. 83–129.
- [281] Wenni Syafitri et al. “Social engineering attacks prevention: A systematic literature review”. In: *IEEE access* 10 (2022), pp. 39325–39343.
- [282] Thomas Szanto and Joel Krueger. “Introduction: empathy, shared emotions, and social identity”. In: *Topoi* 38.1 (2019), pp. 153–162.
- [283] Kazuko Takahashi and Shizuka Yokohama. “On a formal treatment of deception in argumentative dialogues”. In: *Multi-Agent Systems and Agreement Technologies*. Springer, 2016, pp. 390–404.
- [284] Maya Tamir, Chi-Yue Chiu, and James J Gross. “Business or pleasure? Utilitarian versus hedonic considerations in emotion regulation.” In: *Emotion* 7.3 (2007), p. 546.
- [285] Maya Tamir and Brett Q Ford. “Choosing to be afraid: preferences for fear as a function of goal pursuit.” In: *Emotion* 9.4 (2009), p. 488.
- [286] Maya Tamir, Christopher Mitchell, and James J Gross. “Hedonic and instrumental motives in anger regulation”. In: *Psychological science* 19.4 (2008), pp. 324–328.
- [287] OpenAI Team. *ChatGPT: Optimizing language models for dialogue*. 2022.

- [288] Matthias Thimm. “Tweety: A comprehensive collection of java libraries for logical aspects of artificial intelligence and knowledge representation”. In: *Fourteenth International Conference on the Principles of Knowledge Representation and Reasoning*. 2014.
- [289] Romal Thoppilan et al. “Lamda: Language models for dialog applications”. In: *arXiv preprint arXiv:2201.08239* (2022).
- [290] Christopher W Tindale. “Fallacies, Blunders, and Dialogue Shifts: Walton’s Contributions to the Fallacy Debate”. In: *Argumentation* 11 (1997), pp. 341–354.
- [291] Wiebe Van Der Hoek, Wojciech Jamroga, and Michael Wooldridge. “A logic for strategic reasoning”. In: *Proceedings of the fourth international joint conference on Autonomous agents and multiagent systems*. 2005, pp. 157–164.
- [292] Davy Van Nieuwenborgh, Martine De Cock, and Dirk Vermeir. “An introduction to fuzzy answer set programming”. In: *Annals of Mathematics and Artificial Intelligence* 50.3 (2007), pp. 363–388.
- [293] Federico Vicentini et al. “Safety assessment of collaborative robotics through automated formal verification”. In: *IEEE Transactions on Robotics* 36.1 (2019), pp. 42–61.
- [294] Lea Viljanen. “Towards an ontology of trust”. In: *International conference on trust, privacy and security in digital business*. Springer. 2005, pp. 175–184.
- [295] Jacky Visser. “Speech acts in a dialogue game formalisation of critical discussion”. In: *Argumentation* 31.2 (2017), pp. 245–266.
- [296] Henning Wachsmuth et al. “Argumentation synthesis following rhetorical strategies”. In: *Proceedings of the 27th International Conference on Computational Linguistics*. 2018, pp. 3753–3765.
- [297] Christopher D Walton. “Model checking agent dialogues”. In: *International Workshop on Declarative Agent Languages and Technologies*. Springer. 2004, pp. 132–147.
- [298] Christopher D Walton. “Verifiable agent dialogues”. In: *Journal of Applied Logic* 5.2 (2007), pp. 197–213.
- [299] Douglas Walton. “How can logic best be applied to arguments?” In: *Logic Journal of IGPL* 5.4 (1997), pp. 603–614.
- [300] Douglas Walton. “Using argumentation schemes to find motives and intentions of a rational agent”. In: *Argument & Computation* 10.3 (2020), pp. 233–275.
- [301] Douglas Walton and Erik CW Krabbe. *Commitment in dialogue: Basic concepts of interpersonal reasoning*. SUNY press, 1995.
- [302] Zuoguang Wang, Limin Sun, and Hongsong Zhu. “Defining social engineering in cybersecurity”. In: *IEEE Access* 8 (2020), pp. 85094–85115.

- [303] Akash R Wasil et al. “Is there an app for that? A review of popular apps for depression, anxiety, and well-being”. In: *Cognitive and Behavioral Practice* (2021).
- [304] Douglas F Watt. “Social bonds and the nature of empathy”. In: *Journal of Consciousness Studies* 12.8-9 (2005), pp. 185–209.
- [305] Joel Wester, Andreas Brännström, and Juan Carlos Nieves. “Towards A Formalization of the Critical Friend and Socio-Moral Decision-Making in Autonomous Systems”. In: ().
- [306] Michael Winikoff, Lin Padgham, and James Harland. “Simplifying the development of intelligent agents”. In: *Australian Joint Conference on Artificial Intelligence*. Springer. 2001, pp. 557–568.
- [307] Lauren Wispé. “The distinction between sympathy and empathy: To call forth a concept, a word is needed.” In: *Journal of personality and social psychology* 50.2 (1986), p. 314.
- [308] Michael Wooldridge. “Semantic issues in the verification of agent communication languages”. In: *Autonomous agents and multi-agent systems 3* (2000), pp. 9–31.
- [309] Di Wu et al. “Intention Progression with Maintenance Goals”. In: *Proceedings of the 2023 International Conference on Autonomous Agents and Multiagent Systems*. 2023, pp. 2400–2402.
- [310] Chen Xing et al. “Topic aware neural response generation”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 31. 1. 2017.
- [311] Özge Nilay Yalçın. “Evaluating empathy in artificial agents”. In: *arXiv preprint arXiv:1908.05341* (2019).
- [312] Zhun Yang, Adam Ishay, and Joohyung Lee. “Neurasp: Embracing neural networks into answer set programming”. In: *29th International Joint Conference on Artificial Intelligence (IJCAI 2020)*. 2020.
- [313] Hoi-Kau Yuen and Thomas J Richards. “GTKAT: a grounded theory based knowledge acquisition tool for expert systems”. In: *Proceedings 1993 The First New Zealand International Two-Stream Conference on Artificial Neural Networks and Expert Systems*. IEEE. 1993, pp. 152–155.
- [314] Lotfi Asker Zadeh. “Fuzzy sets as a basis for a theory of possibility”. In: *Fuzzy sets and systems* 1.1 (1978), pp. 3–28.
- [315] Jamil Zaki. “Integrating empathy and interpersonal emotion regulation”. In: *Annual Review of Psychology* 71 (2020), pp. 517–540.
- [316] Antoine Zimmermann and Chan Le Duc. “Reasoning with a network of aligned ontologies”. In: *International Conference on Web Reasoning and Rule Systems*. Springer. 2008, pp. 43–57.

