UMEÅ UNIVERSITET

FORMAL METHODS FOR VERIFICATION IN HUMAN-AGENT INTERACTION

Andreas Brännström

Akademisk avhandling

som med vederbörligt tillstånd av Rektor vid Umeå universitet för avläggande av filosofie doktorsexamen framläggs till offentligt försvar i HUM.D.220, humanisthuset, måndagen den 5 maj, kl. 13:15.

Avhandlingen kommer att försvaras på engelska.

Fakultetsopponent: Professor, Stefania Costantini,

University of L'Aquila, Italy.

Department of Computing Science

Organization
Umeå University
Department of Computing Science

Document type
Doctoral thesis

Date of publication
14 April 2025

Author
Andreas Brännström

Title
Formal Methods for Verification in Human-Agent Interaction

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.

Keywords
formal verification, human-agent interaction, non-monotonic reasoning, theory of mind

Language
English

ISBN
print: 978-91-8070-682-7
PDF: 978-91-8070-683-4

ISSN
0348-0542

Number of pages
262