rijksuniversiteit groningen / faculteit der letteren

Agentic AI and its Roles in Corporate Negotiations: Benevolent Tools or Manipulative Actors?

Natalie Voo Xin Ru

S4228693

Master's thesis

Digital Humanities

Faculty of Arts

University of Groningen

July 1st, 2025

Abstract

This study examines the efficacy of artificial intelligence (AI) as a negotiation agent across three distinct roles: human replacement, supportive assistant, and undermining assistant. A role-playing experiment (N=43) simulated a corporate negotiation, measuring agreement rates, negotiation expertise, and efficiency. Results indicate the replacement AI achieved the highest agreement rate (95.7%), while the undermining condition significantly reduced success. Supportive AI interactions were 13.5% more efficient than undermined ones. Critical vulnerabilities (primarily within the replacement role) were identified including AI's susceptibility to manipulation and poor temporal awareness. Furthermore, participants eliciting advice from the undermining AI assistant consistently failed to detect AI deception. The findings demonstrate AI's potential to augment negotiation but reveal significant ethical risks and performance inconsistencies that challenge its readiness for unsupervised deployment. All of the code, data and materials used in this study are available at https://github.com/natalievxr/Master_Thesis.git.

Keywords: Artificial Intelligence, Negotiation, Large Language Models, Claude 3.7 Sonnet, Undermine, Support, Replacement, AI agents, Agentic AI, Persuasion

1. Introduction

When was the last time you negotiated? Have you ever used negotiation tactics to achieve what you wanted? If you're like most people, you've likely faced situations where you had to: from something perhaps more high stakes such as discussing a salary with your employer, or something a little more low stakes such as bargaining for the price of fruit at a market – negotiation, we can conclude, is thus almost inevitable in our lives (Park et al., 2012; Zhu et al., 2025).

At its core, negotiation is a process where two or more parties with differing aims engage in productive dialogue with the purpose of bridging gaps, settling disputes, or crafting agreements that benefit everyone involved (Zhan et al., 2024; Zhu et al., 2025). It often comes into play when each party has distinct priorities, requiring give-and-take to reach an agreed upon outcome. Every interaction has limits – what each party is willing to offer or accept, and these limits often differ and thus negotiation can be thought of as a careful, calculated dance where all negotiating partners synchronise to (hopefully) find a happy middle ground (Zhu et al., 2025).

Negotiations can range from cooperative to competitive and take place in all kinds of settings – from everyday casual conversations to high-stakes business or even diplomatic discussions, where negotiations can affect millions across the world, with a current example being Trump's tariffs (Sanger, 2025). Given its frequency and wide range of applications, effective negotiation skills are essential for achieving favourable outcomes in various everyday social and professional settings.

In recent years, the Artificial Intelligence (AI) revolution has swept across the globe. Defined by Li (2024), AI is 'the simulation of human intelligence in machines programmed to perform tasks typically requiring human cognition,' and it has boomed significantly – driven by pioneering tech companies such as OpenAI, Microsoft, Amazon Web Services and Nvidia (Fernandez, 2025), which have developed

cutting-edge technologies, leading the industry. Today, it seems almost impossible to escape the influence of AI. The term is everywhere; with businesses, organisations, and institutions all racing to embrace this new technology, tempted by the grapevine whispers of AI promising more efficiency, productivity, and innovative ideas.

More recently, the term ‘AI agents’ has surged in popularity, becoming this year’s latest buzzword (Belcic & Stryker, 2025; Thompson, 2025). Interestingly, its definition varies across industries and stakeholders, adapting to different contexts. So what exactly is it? An AI agent is a software or system capable of performing tasks autonomously or semi-autonomously, leveraging AI techniques such as machine learning (ML), natural language processing (NLP), or decision-making algorithms (Gutowska, 2024). For instance, in healthcare, AI agents might diagnose diseases or suggest treatments, while in customer service, a more familiar application for many; they often take the form of chatbots or virtual assistants handling inquiries. The retail sector has similarly been transformed by AI agents, which power personalised recommendations, optimise logistics, and deploy interactive chatbots, enhancing online platforms' appeal and often securing consumer purchases before in store visits (Deng et al., 2020).

The concept of an AI agent thus remains fluid, evolving alongside advancements in AI technology and varying based on the industry. However, it is generally understood as an intelligent system that executes tasks, with its precise definition shaped by factors like autonomy level, service, and specific use cases.

Research on negotiation has frequently highlighted that human negotiators are susceptible to cognitive biases, emotional influences and constrained knowledge bases (Zhan et al., 2024), which can obscure valuable implicit information during the negotiation process and lead to suboptimal outcomes. Additionally, while expert negotiators exist, most individuals generally lack the specialised skills required for effective negotiation. This widespread skill gap underlines the central beating heart of this thesis: the

role(s) of AI as a negotiation agent, as AI agents are inherently free from these cognitive and affective constraints due to their seemingly unlimited knowledge and the fact that they do not experience emotion the same way that humans do, which suggests their potential superiority in negotiation contexts. But critical questions remain: How effective are these AI agents in real-world negotiations? Are we at a point in time where AI is already better than humans in something as complex as negotiations? Will AI take over the world and all our jobs? The latter is perhaps a valid question, but unfortunately it is not within the scope of this thesis.

This thesis is specifically interested in three key roles of an AI agent and their implications for negotiation outcomes and efficiency. First, we examine AI as a replacement, where it fully assumes the role of human negotiators. Second, we explore AI as an assistant, analysing both its supportive role of providing legitimate guidance to enhance decision-making, and also its undermining role, in which AI may appear as if it is supportive while covertly manipulating outcomes to the user's disadvantage. Additionally, the study considers how contextual factors, including cultural differences, human expertise, and willingness to collaborate, may influence the effectiveness of AI in negotiation scenarios. By analysing these dimensions, this research aims to contribute to a deeper understanding of AI's evolving impact on negotiation dynamics. In this thesis, the following research questions were formulated and aimed to be answered:

1. What are the impacts of different roles AI (undermine, support, and replacement) can play in negotiation processes in terms of outcomes and efficiency?
2. How do contextual factors (cultural differences, human expertise, and collaboration skills) influence the AI agent's effectiveness in negotiations?

These questions were selected to address critical gaps identified in recent studies: while research demonstrate AI's tendencies for prosocial behavior and high-performance decision-making (Mei et al.,

2024), they also reveal vulnerabilities, such as susceptibility to manipulation (Derner & Batistič, 2023) and inconsistent toxicity (Zhuo et al., 2023), especially when prompted to in the guise of role-playing scenarios (Mao et al., 2024). Understanding these dynamics are essential for harnessing AI's benefits in negotiation while mitigating risks, particularly as AI capabilities advance toward Kurzweil's (2014) predicted singularity theory, which posits that AI will eventually eclipse human cognitive abilities through exponential technological advancement. The findings of this thesis hopes to advance the emerging field of AI behavioural science (Bernasconi & Ferilli, 2024), highlighting the need for further comparative studies while offering pragmatic guidance for human-AI collaboration frameworks.

This study consists of a role-playing experiment simulating a corporate negotiation scenario, with participants assigned to one of three conditions: (1) a replacement role where they negotiate directly with an AI agent role-playing as a human senior HR professional, (2) a support condition where two participants negotiate with one another while receiving AI assistance, with one party unknowingly paired with an undermining AI that provides detrimental advice while the other receives genuinely supportive guidance. This thesis analyses negotiation outcomes and efficiency metrics while examining how contextual factors like cultural background, human expertise, and collaborative willingness (measured through post-negotiation five-point Likert scale surveys) influence the effectiveness of AI in these different roles. Cronbach's alpha is employed to evaluate the survey's internal reliability. Statistical analysis is conducted using the Kruskal-Wallis test, with subsequent Dunn's tests for pairwise comparisons, complemented by descriptive statistics. This experimental design allows for systematic comparison of how AI's various roles impact negotiation dynamics while controlling for important human factors that may moderate these effects.

Key findings show that replacement AI achieved the highest agreement rates (95.7%), primarily through hybrid human-AI solutions. The support AI helped maintain favourable outcomes, however the undermining AI significantly reduced negotiation success for participants.

This study reveals AI's vulnerabilities to user manipulation, inconsistent role adherence, and lack of temporal awareness. The support AI and participant conversations were 13.5% more efficient than the undermining AI interactions with participants. Notably, participants often failed to detect AI deception, raising ethical concerns. These findings highlight both AI's potential to enhance negotiations and risks of undermining human agency when misused.