Formal Verification of AI Reasoning in Economic and Financial Environments: Applications to Insurance Markets

Abstract

This PhD project investigates how formal methods and logics for strategic reasoning can support the specification, verification, and robustness analysis of Artificial Intelligence (AI) systems operating in economic and financial environments. These environments are characterized by uncertainty, asymmetric information, heterogeneous preferences, and strategic interactions among agents with different objectives and informational states. Such characteristics pose significant challenges for AI-based decision-making, particularly when reliability and the ability to establish guarantees about system behavior are essential. The project pursues three complementary objectives: (i) investigating how epistemic, doxastic, and strategic logics can represent economic agents with heterogeneous preferences and incomplete information; (ii) adapting formal verification techniques to analyze strategic properties under uncertainty; and (iii) investigating how formal specifications can support the development and evaluation of robust AI systems, including those based on Generative AI (GenAI). Building on research on preference heterogeneity and multidimensional private information in insurance markets, the project will develop formal models of economic interaction and investigate how their properties can be verified under explicitly defined assumptions. The expected contribution is a methodological framework connecting strategic reasoning, formal verification, and robust AI, supported by theoretical results and computational case studies.

1. Introduction and Motivation

Formal verification of Multi-Agent Systems (MAS) provides systematic methods for analyzing interacting agents and determining whether a system satisfies properties expressed in a formal logic [1, 2]. Techniques such as model checking [3] enable the analysis of possible system behaviors, while strategic logics allow properties concerning the capabilities and objectives of individual agents and coalitions to be expressed formally.

Economic and financial environments provide a relevant setting for these methods because decisions depend not only on objective conditions but also on heterogeneous preferences, private information, and beliefs about other agents. Research on insurance markets illustrates these challenges. Cutler, Finkelstein, and McGarry [4] investigate the role of preference heterogeneity in explaining insurance choices, while Finkelstein and McGarry [5] examine multiple dimensions of private information, including individual risk and preferences for insurance. Together, these studies motivate models that distinguish observable behavior from the underlying characteristics and informational states of economic agents.

These challenges become particularly relevant when AI systems are used to analyze or support decisions in such environments. Models that inadequately represent agents' preferences or information may produce unreliable conclusions when the underlying assumptions change. Moreover, although formal verification can establish guarantees about precisely specified models, transferring these guarantees to AI implementations remains an open methodological challenge.

This project investigates how formal methods can address these issues by connecting the representation of strategic interaction with the verification and robustness analysis of AI-based systems. The central research problem is how to specify and verify relevant properties of strategic reasoning when agents have heterogeneous preferences, incomplete information, and dynamically evolving beliefs, and how these properties can inform the evaluation of AI systems operating under such conditions.

2. Research Directions

2.1. Formal Modeling of Economic and Strategic Interaction

The first direction concerns the extension or combination of epistemic, doxastic, and strategic formalisms to represent economic agents with heterogeneous preferences, private information, and dynamic belief changes. Existing approaches to strategic reasoning [1, 2] provide a foundation for investigating how these dimensions influence agents' decisions and the resulting system behavior.

The project will initially consider insurance markets as a motivating case study, examining how differences in risk, preferences, and information can be represented in a formal multi-agent model. Epistemic and doxastic logics will provide tools for representing informational states, while strategic logics such as Alternating-time Temporal Logic (ATL) [1] will support the specification of properties concerning agent capabilities and system evolution. The objective is to identify representations that capture relevant economic mechanisms while remaining amenable to formal analysis.

2.2. Formal Verification and Robustness

The second direction investigates how formal verification techniques can establish properties of multi-agent systems operating under uncertainty and asymmetric information. Model checking [3] and runtime verification [6] will be considered according to the nature of the properties and the computational structure of the models.

Robustness will be investigated in terms of whether specified properties remain valid under explicitly defined variations in agent preferences, informational states, or environmental conditions. The project will distinguish guarantees established for abstract formal models from the empirical robustness of their computational implementations. This distinction is essential when evaluating AI systems whose behavior may not be fully captured by a symbolic model.

2.3. Formal Specifications and GenAI-Based Systems

The third direction concerns the relationship between formal specifications and AI-based reasoning or synthesis. The project will investigate whether formal representations of strategic reasoning can improve the correctness and reliability of algorithms generated by GenAI systems, compared with equivalent descriptions in natural language or pseudocode.

Controlled experiments and ablation studies will be used to assess the contribution of different components of formal specifications. Generated implementations will be evaluated independently against predefined properties using appropriate verification techniques. A further theoretical question concerns how the computational structure of formal reasoning procedures can be represented within neural architectures, building on existing results on the computational capabilities of decoder-only Transformers with Chain-of-Thought reasoning [7]. The project will investigate which properties can be formally established and under what conditions such guarantees can be related to an AI implementation.

3. Research Questions

The project will be guided by the following research questions:

- **RQ1 - Formal modeling:** How can epistemic, doxastic, and strategic logics represent heterogeneous preferences, private information, and dynamic belief changes in economic and financial interactions?
- **RQ2 - Verification:** How can model checking and runtime verification be used to specify and verify strategic properties of multi-agent systems under uncertainty?
- **RQ3 - Robustness:** Which properties remain valid under variations in agent characteristics, informational states, and environmental conditions, and how can these conditions be formally characterized?
- **RQ4 - GenAI reliability:** Can formal specifications improve the correctness of GenAI systems, and which components contribute most to their reliability?
- **RQ5 - Neural computation:** Under what conditions can formal guarantees established at the symbolic level be related to the behavior of neural or generative implementations?

4. Methodology and Expected Contributions

The research will combine formal modeling, logical analysis, verification techniques, and computational evaluation to investigate the robustness of AI-based systems in economic environments characterized by uncertainty, heterogeneous preferences, and asymmetric information.

First, formal models of economic interaction will be developed to represent agents, their preferences, information, beliefs, and strategic choices. Relevant epistemic, strategic, and temporal properties will be expressed using appropriate logical frameworks and analyzed through model-checking techniques. Where relevant, runtime verification will be considered to monitor whether specified properties hold during system execution.

Second, robustness will be studied by examining how the validity of these properties depends on variations in model assumptions, information structures, agents' beliefs, and strategic choices. This analysis will identify the conditions under which desired properties are preserved or violated and establish a formal basis for evaluating robustness.

Third, the proposed methods will be applied to the evaluation of GenAI systems performing economic or strategic reasoning. Their outputs and decisions will be assessed against formally specified properties under controlled variations of problem descriptions, available information, and interaction conditions. This will help investigate whether GenAI systems preserve intended strategic behavior across different scenarios, and whether formal specifications and verification techniques can identify systematic failures or limitations in their robustness. An insurance-market model will provide an initial case study, with possible extensions to other economic and financial settings.

References

[1] Alur, R., Henzinger, T. A., & Kupferman, O. (2002). Alternating-time temporal logic. *Journal of the ACM, 49*(5), 672–713.

[2] Mogavero, F., Murano, A., Perelli, G., & Vardi, M. Y. (2014). Reasoning about strategies: On the model-checking problem. *ACM Transactions on Computational Logic, 15*(4), 1–47.

[3] Clarke, E. M., Grumberg, O., & Peled, D. A. (1999). *Model checking*. MIT Press.

[4] Cutler, D. M., Finkelstein, A., & McGarry, K. (2008). Preference heterogeneity and insurance markets: Explaining a puzzle of insurance. *American Economic Review, 98*(2), 157–162.

[5] Finkelstein, A., & McGarry, K. (2006). Multiple dimensions of private information: Evidence from the long-term care insurance market. *American Economic Review, 96*(4), 938–958.

[6] Bartocci, E., Falcone, Y., Francalanza, A., & Reger, G. (2018). Introduction to runtime verification. In *Lectures on runtime verification: Introductory and advanced topics* (Lecture Notes in Computer Science, Vol. 10457, pp. 1–33). Springer.

[7] Merrill, W., & Sabharwal, A. (2024). The expressive power of transformers with chain of thought. In *Proceedings of the International Conference on Learning Representations (ICLR)*.