SCIENTIFIC CONTEXT

Recent advances in Artificial Intelligence (AI), particularly Large Language Models (LLMs), have enabled the development of increasingly autonomous systems capable of reasoning, planning, using external tools, and interacting with their environments. These systems, commonly referred to as *agentic AI*, go beyond traditional AI models by autonomously selecting and executing actions to achieve specific objectives. However, their growing autonomy raises important questions about reliability and correctness. Current approaches largely rely on empirical evaluation, testing, and runtime safeguards, which cannot generally provide formal guarantees about the behaviours these systems may produce. This limitation becomes particularly critical when AI agents interact with other autonomous entities or operate in environments where incorrect decisions may have significant consequences.

In parallel, research on Multi-Agent Systems (MAS) has developed formal methods for reasoning about the behaviour of autonomous and interacting agents. In particular, *strategic logics*, such as Alternating-time Temporal Logic (ATL), provide a rigorous foundation for expressing and verifying what individual agents or coalitions of agents can achieve through their strategies. These techniques make it possible to reason about cooperation, competition, and the ability of agents to enforce desired properties despite the actions of other agents. However, existing verification frameworks typically assume finite, well-defined models with known actions and transition relations. Such assumptions are difficult to maintain in modern agentic AI systems, where actions may be generated dynamically, environments may evolve, and information about the system may be incomplete.

The **Sound or Silent (SoS)** project aims to bridge this gap by investigating a new approach to *sound-by-construction agentic AI*. The central idea is to integrate strategic reasoning and formal verification directly into the decision-making process of autonomous agents. Rather than verifying an agent only after it has been developed or monitoring its behaviour after execution, SoS investigates whether an agent can produce actions that are formally certified to satisfy a given specification. Under this approach, candidate actions are executed only when they can be established to belong to a strategy guaranteeing the required properties. When such a guarantee cannot be established, the agent abstains from acting.

Achieving this objective requires extending strategic verification techniques to open, evolving, and partially specified environments, as well as investigating suitable notions of correctness under incomplete information. The project will also explore an architecture combining an untrusted generative component, responsible for proposing candidate actions or strategies, with a trusted formal verification kernel, responsible for checking their correctness. Particular attention will be devoted to the expressiveness of the underlying formal models, the computational cost of certification, and the trade-off between formal guarantees and the ability of agents to perform useful tasks.

The research will combine theoretical developments with the implementation and experimental evaluation of prototype systems, considering scenarios involving autonomous planning, tool-using AI agents, and multi-agent coordination. The overall objective is to establish the foundations of agentic AI systems whose autonomy is constrained by formally verifiable strategic guarantees.

SKILLS

Required skills, experience, and knowledge:

- Strong knowledge of formal methods, logics, model checking, or strategy synthesis.
- Experience with multi-agent systems, strategic logics, partial observability, runtime verification, or related verification tools.
- Experience with LLM-based agents, agentic tool use, and/or verifiable generative AI workflows.
- Experience in research software development and algorithm implementation.
- Ability to formulate research questions, design experiments, and analyse results.
- Fluency in written and spoken English.

Other abilities and skills:

- Ability to work in a team, dialogue and writing skills
- Interpersonal and pedagogical skills
- Ability to synthesize