MODELLO PER LA PRESENTAZIONE DEL PROGETTO DI RICERCA

a) **Titolo del Progetto: SAFE-AI-Lite: Secure, Accountable, and Fair Enforcement for Autonomous Intelligence (Progetto di Ricerca)**

b) **Proponente (PI): Stefania Costantini**

c) Posizione accademica del proponente: **Professore Ordinario**

d) **Curriculum vitae** del proponente (max 5000 caratteri – circa 2 pagine) con elenco delle pubblicazioni più significative (max 10) nel periodo 21-2025, relative al tema del progetto. Indicatori ASN 2024/26 alla scadenza del bando (solo per i progetti di ricerca) relativamente alla fascia superiore del Settore concorsuale e del Settore scientifico disciplinare di appartenenza.

- *CV Summary*

Stefania Costantini has been a Full Professor of Computer Science at the University of L'Aquila since 2005, with internationally recognised expertise in computational logic, autonomous agents, and trustworthy AI. A pioneer in agent-oriented logic programming and reflective agents, her work has made long-standing contributions to agents, Answer Set Programming and the broader field of computational logic. Her leadership is demonstrated through the coordination of high-impact national and international research projects, an extensive publication record of over 200 papers, and significant roles in prestigious scientific committees. Her influence in the field is highlighted by her inclusion in the Stanford "Top 2% World Scientists" list (2023–2025) and her recognition in 2024 as one of the most influential women and individuals in AI in Italy.

- *Research and Leadership*

  - Research Interests: Computational Logic, Logic Programming, Meta-programming and Meta-reasoning, Agents and Multi-Agent Systems, Automated Reasoning, Knowledge Representation, Neuro-symbolic Hybrid Systems, Ethics in Artificial Intelligence.
  - Project Leadership: Key recent roles include:
    - Principal Investigator (PI): PRIN 2022 TRUSTPACTX, PNRR Cascading Bands EBABLE.
    - Local PI: PRIN 2022 PNNR ADVISOR, FAIR National Project.
    - Vice-Coordinator: COST Action CA17124 "DIGital FORensics: evidence Analysis via intelligent Systems and Practices" (2018–2022).
  - Publications: Author of over 200 scientific publications.
  - Invited Talks: Delivered 7 invited talks at international conferences and workshops in the last 5 years.

- *Scientific and Institutional Service*
  - Executive and Board Roles:
    - Executive Committee Member, Association for Logic Programming (ALP)
    - Former President, Italian Association for Computational Logic (GULP)

    - Former Board Member, Italian Association for Artificial Intelligence (AIxIA)
- Conference Leadership:
    - Co-Chair: ICLP 2023, RuleML+RR 2017, ASP07.
    - Chair: GULP92.
- Editorial Roles:
    - Editor, Special Issues in "Theory and Practice of Logic Programming" and "Journal of Logic and Computation" (Q1 Journals).
- Reviewing and Evaluation:
    - Serves as a reviewer for the main journals and conference in her research field, the EU, Italian Ministry of Universities and Research (MUR), ESF, and numerous national and international universities.
    - In 2024, served as a reviewer for the Italian Chamber of Deputies.
    - PhD Committees: Member of national PhD boards and international scientific committees.

- *Recent Recognition*

- Most Influential Women in AI in Italy (2024): [Link](https://www.repubblica.it/tecnologia/2024/03/08/news/donne_intelligenza_artificiale_italia-422273156/amp/)
- Most Influential Italians in AI (2024): [Link](https://www.repubblica.it/tecnologia/dossier/romecup/2024/03/19/news/intelligenza_artificiale_in_italia_ecco_tutti_nomi_luniversita_e_la_ricerca_le_startup_e_le_grandi_aziende_lart_e_e_la_-422332723/)

Indicatori ASN 2024/26 alla scadenza del bando: according to Scopus, total of 1,472 Citations, h-index 18, journal papers 46; according an IRIS simulation for the positions of “docente di Prima Fascia” and “Commissario” (positive outcome), 23 journal papers in the last 10 years, and 869 citations resulting in h-index 13 in the last 15 years.

Selected publications (2021–2025):

1. S. Costantini, A. Formisano (2025): *A Prototyping Framework for Reduct-Based ELP Solvers: Methodology and Implementation*. Proc. of ECAI, 28th European Conf. on Artif. Intell., Frontiers in AI and Applications, Vol. 413, IOS Press
2. A. Vozna, A. Monaldini, S. Costantini (2025): *A Trust-Aware Architecture for Personalized Digital Health: Integrating Blueprint Personas and Ontology-Based Reasoning*. J. Medical Syst. 49(1)
3. S. Costantini, A. Formisano (2024): *Epistemic Logic Programs: A Study of Some Properties.* Theory Pract. Log. Program. 24(3)
4. L. De Lauretis, F. Persia. S. Costantini, D. D'Auria (2023): *How to leverage intelligent agents and complex event processing to improve patient monitoring*. J. Log. Comput. 33(4)

5. G. De Gasperis, S. Costantini, A. Rafanelli, P Migliarini, I. Letteri, A. Dyoub (2023): *Extension of constraint-procedural logic-generated environments for deep Q-learning agent training and benchmarking*. J. Log. Comput. 33(8)
6. P. Dell'Acqua, S. Costantini (2023): *Empathetic human-agent interaction via emotional behavior trees.* Intelligenza Artificiale 17(1)
7. Dyoub, S. Costantini, F. Lisi (2022): *Learning Domain Ethical Principles from Interactions with Users*. Digit. Soc. 1(3)
8. S. Costantini (2022): *Ensuring trustworthy and ethical behaviour in intelligent logical agents*. J. Log. Comput. 32(2)
9. A. Dyoub, S. Costantini, I. Letteri (2022): *Care Robots Learning Rules of Ethical Behavior Under the Supervision of an Ethical Teacher*. HYDRA/RCRA@LPNMR
10. A. Dyoub, S. Costantini, I. Letteri, F. Lisi (2021): *A Logic-based Multi-agent System for Ethical Monitoring and Evaluation of Dialogues.* ICLP Tech. Comm.

e) Eventuali **componenti del gruppo di ricerca** (solo per i progetti di ricerca di base): Dr. Giovanni De Gasperis.

f) Settore di ricerca ERC di riferimento per la proposta (indicare anche due sotto-settori): **PE6 – Computer Science and Informatics,** sotto-settori PE6_08 Artificial Intelligence, Autonomous Agents e PE6_05 Knowledge Representation and Reasoning

g) **Abstract** (max 1000 caratteri):

The integration of Large Language Models (LLMs) into autonomous agents introduces a critical security gap, as their opaque nature prevents formal guarantees on behaviour. This project bridges that gap by introducing an innovative certification-driven enforcement framework. We will develop a lightweight symbolic supervisor based on logical reflection to monitor and constrain LLM-based components within hybrid agents in real-time, ensuring compliance with security and ethical constraints without sacrificing flexibility. The result will be a formally grounded model and a prototype for trustworthy AI, directly addressing the requirements of the EU AI Act and laying the scientific foundation for a future, broader ERC-scale research agenda. This approach transforms AI safety from reactive mitigation to a proactive and verifiable guarantee.

h) **Descrizione del progetto** (max 8.000 caratteri, compresi eventuali riferimenti bibliografici. È consentito inserire figure nella proposta. Le figure non concorrono alla determinazione del calcolo del numero dei caratteri.)

a. *State of the art:* Autonomous agents, empowered by Large Language Models (LLMs), are revolutionising human-machine interaction. In fact, Agents (or, by common terminology, “AI agents”) can nowadays be built as any combination of symbolic and LLM-based components; they can be autonomous, adaptive, and intelligent, but current systems still lack verifiability, accountability, and fairness. This creates risks in real-world applications. Current research either focuses narrowly on symbolic verification techniques or experimental attempts at monitoring LLM agents, but no framework integrates these approaches systematically. However, this evolution introduces unprecedented risks: LLM models, inherently probabilistic and opaque, can generate unpredictable, non-compliant, or even harmful behaviours. Current countermeasures are

predominantly based on mitigation techniques, such as prompt engineering, ethical fine-tuning, or post-hoc explanations. While useful, these approaches are reactive and do not offer formal guarantees on the safety and reliability of an agent's behaviour during execution. The international scientific community, as highlighted in the debate on AI regulation (see, e.g., the IASEAI'25 and '26 Conferences, URL https://www.iaseai.org/), recognises the urgency of moving from empirical solutions to scientifically founded and verifiable control mechanisms. There is a clear gap between the generative capabilities of LLM-based systems and our ability to ensure their reliability in critical contexts.
Vision: This project addresses this gap by proposing a paradigm shift: from post-hoc analysis to runtime enforcement based on logical principles. In particular, it proposes a hybrid architecture where symbolic reasoning guarantees rigour, while LLMs provide adaptability and interaction, the two components united through self-awareness, self-checking, and proactive enforcement. The project's approach aims to leverage the structured reasoning, interpretability, and reliability of symbolic AI, alongside the flexibility, adaptability, and communicative strengths of LLMs. By thoughtfully combining these complementary paradigms, SAFE-AI intends to pave the way towards frameworks that enable designers to create AI systems capable of robust, ethical, and transparent decision-making, significantly reducing or eliminating risks of unintended behaviours, biases, and emergent threats often associated with autonomous agents

b. *Objectives:* The general objective is to develop a rigorous and lightweight framework to certify and enforce the behaviour of hybrid agents (symbolic-LLM), ensuring security and reliability. The goal is to ensure that AI agents enhance productivity and efficiency without compromising security or ethical standards. Ultimately, research efforts carried on in the SAFE-AI project will contribute to the safe and effective integration of autonomous systems into real-world applications. Over several decades, the PI has carried out preliminary developments and experiments that lay a solid foundation for the SAFE-AI effort. This positions the project and its future developments to transform the field of autonomous agents and contribute to building a secure, fair, and trustworthy AI ecosystem in Europe and globally.

The specific, measurable, and achievable objectives in 12 months are:

1. Define a formal model of a hybrid agent (Months 1-2). Develop an architecture that isolates the LLM component (treated as untrusted) from a trusted symbolic reasoning core, specifying its interfaces and interaction protocols. Deliverable: Technical specification of the architectural model.
2. Develop a certification language (Months 3-5). Create a set of security, ethical, and operational properties (e.g., non-disclosure of sensitive data, adherence to procedural rules) expressible in computational logic and verifiable at runtime. Deliverable: Library of formal properties and their specifications.
3. Design and prototype a Symbolic Runtime Supervisor (Months 6-9). Implement a software component (Symbolic Runtime Supervisor - SRS) that intercepts the actions proposed by the LLM module, verifies them against the certified properties, and applies

enforcement strategies (e.g., blocking, requesting reformulation, fallback to safe procedures). Deliverable: Open-source prototype of the supervisor with documentation.
4. Validate and disseminate the results (Months 10-12). Test the framework on a significant case study (e.g., a digital assistant for managing confidential information), analyse its performance, and prepare the results for scientific publication. Deliverable: Scientific article and final project report.

c. *Methodology:* The project adopts an approach grounded in computational logic and logic programming for agents. The symbolic core of the agent encodes trusted knowledge, goals, and regulatory constraints. The LLM component is treated as a powerful but unreliable "oracle," whose responses are subject to formal verification.

The methodology is articulated as follows:

1. Formal Modelling. The hybrid architecture will be defined using a logic programming language for agents, extending it to integrate calls to external LLMs as special actions. This allows for a clear separation between deliberative reasoning (symbolic) and text generation (LLM).
2. Supervisor Development. The Symbolic Runtime Supervisor (SRS) will be implemented using logical reflection mechanisms. It intercepts the agent's reasoning cycle under certain conditions, analyses the intentions involving the use of LLMs, and validates their output against a set of formal constraints before allowing the corresponding action to be executed.
3. Empirical Validation. The prototype will be tested on concrete scenarios to measure the computational overhead of the supervisor and its effectiveness in preventing security violations. The results will be compared with approaches based solely on prompt engineering.

Communication and Dissemination: SAFE-AI will prioritise publication in high-impact, peer-reviewed Open Access journals, and presentations at leading international conferences. Events will be organised within the University and also for companies and for the general public, to raise awareness, addressing ethical implications.

d. Work Plan: The 12-month project is structured in four sequential phases, each with clear milestones and deliverables, ensuring feasibility with the available resources.

Month(s)	**Main Activity**	**Concrete Milestones and Deliverables**
1-2	Phase 1: Theoretical Foundations - Definition of the formal model of the hybrid agent.	M2: Publication of the technical specification on an open repository (e.g., GitHub).
3-5	Phase 2: Development of the Certification Language - Creation of the property library.	M5: Release of a library of formal constraints for security and ethics.

6-9	Phase 3: Prototyping - Implementation of the Symbolic Runtime Supervisor (SRS).	M9: Release of the alpha version of the SRS prototype with source code and documentation.
10-12	Phase 4: Validation and Dissemination - Testing and writing of the results.	M12: Submission of an article to an A* conference and release of the final report.

This work plan is realistic for a 12-month, €15,000 project, focusing on producing a solid theoretical contribution and a working prototype that will serve as a springboard for the broader ERC proposal.

i) **Elementi di originalità e innovazione** della proposta e impatto in termini di **rilevanza dell'avanzamento nella ricerca di base** per la comunità scientifica di riferimento (max 3000 caratteri):

The originality of the project lies in its radically new approach to trust in AI: instead of trying to make LLM models intrinsically "safe" (a problem that is difficult, maybe impossible, to address in general), we propose to make them verifiably safe in their use. We treat LLM components as "untrusted oracles" whose generative power is harnessed and controlled by a formal reflective logical supervisor. This shifts the problem of security from the model to the architecture, a domain where formal verification is possible.
The innovation manifests on three levels:

1. Architectural Paradigm. Unlike current approaches that merge reasoning and generation, our architecture imposes a hierarchical separation of powers. The symbolic core retains final authority, ensuring that the agent always operates within a pre-certified security perimeter. This concept of "runtime supervision" based on logic for hybrid agents is unprecedented.
2. Proactive Enforcement. The project introduces a proactive and dynamic enforcement mechanism. Instead of merely detecting violations after the fact, our Symbolic Supervisor intervenes before an unsafe action is executed, correcting the agent's trajectory in real-time. This preventive approach is a significant step forward from current reactive mitigation strategies.
3. Feasibility and Scalability. We propose a lightweight and modular solution, designed to be integrated into existing agent systems with contained computational overhead. This not only ensures the feasibility of the annual project but also lays the foundation for a scalable framework capable of managing complex multi-agent systems, as will be outlined in a future ERC proposal.

The scientific impact is twofold. In the short term, we provide the community with a novel formal model and novel techniques tool for building safer hybrid agents. In the long term, this work opens a new line of research on trustworthy AI, aligned with European strategic priorities, and builds the necessary scientific foundations to successfully address the challenges of large-scale projects. By establishing a new formal bridge between symbolic reasoning and sub-symbolic models, this project contributes directly to fundamental

computer science, advancing our understanding of hybrid intelligent systems. By developing rigorous deployment frameworks based on self-conscious, reflective, self-improving agents and by inventing new technologies to control existing agentic systems, AI agents can be steered to serve humanity's best interests.

j) Impatto del progetto in riferimento alle tematiche di genere (facoltativo, max 3000 caratteri)

Piano di spesa

Voce di spesa	*Importo (Euro)*
Borse di ricerca (art.2 del Regolamento per il conferimento di borse di ricerca attualmente in vigore)	
Rinnovo assegni di ricerca	
Materiali di consumo	1000
Attrezzature, strumentazioni, software	3.000
Missioni	8.000
Acquisto prodotti ritenuti necessari per la realizzazione del progetto (es. materiale librario, licenze per l'accesso a banche dati, ecc.)	
Pubblicazioni, organizzazione di convegni e workshop	3.000

Stefania Costantini
09.01.2026
16:51:31
GMT+01:00