Prof. **Luca Spalazzi**
Dott. **Gianluca Bonifazi**

UNIVERSITÀ POLITECNICA DELLE MARCHE

AI e CYBERSECURITY: opportunità o minaccia?

CINI CYBERSECURITY NATIONAL LAB

GII

CRISPY
Centro di Ricerca e Servizio per la Privacy e la Cybersecurity

Cybersecurity and Blockchain Research Lab
Dipartimento di Ingegneria dell'Informazione
Università Politecnica delle Marche

Strumento di ATTACCO

Strumento di DIFESA

Superficie VULNERABILE

IL TRIPLICE RUOLO DELL'AI

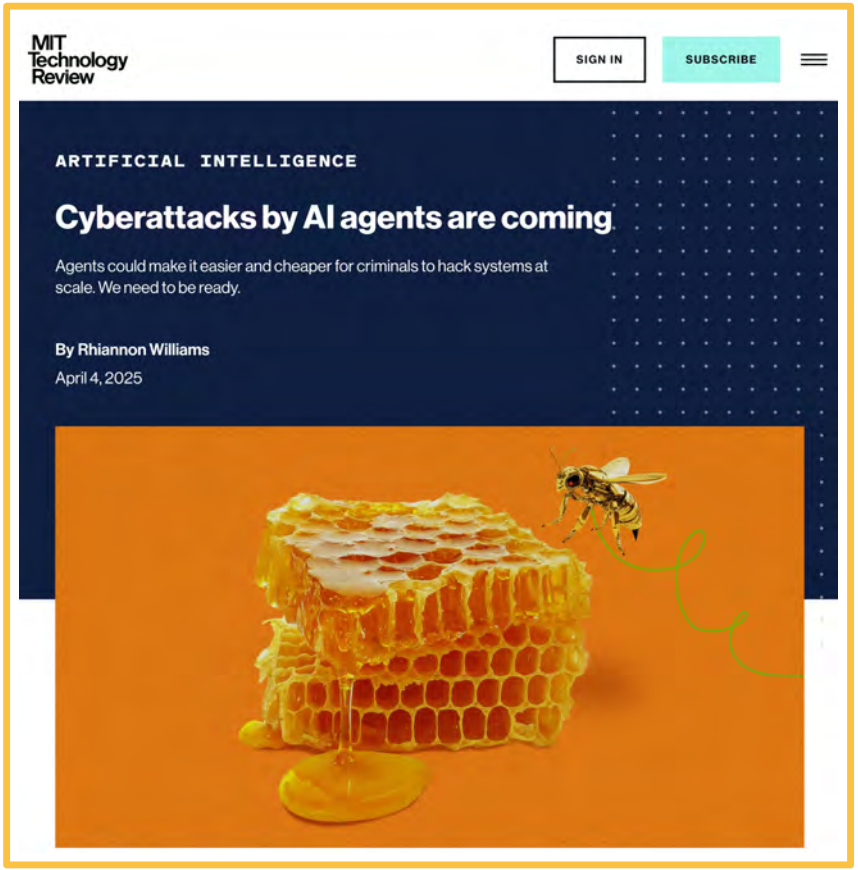
L'intelligenza artificiale come strumento di ATTACCO

Gli **attaccanti** possono sfruttare **l'intelligenza artificiale** per:

- **automatizzare** attività offensive;
- **generare** campagne di **phishing** più credibili;
- **impersonare persone reali** attraverso voci e video sintetici;
- accelerare la **ricerca di vulnerabilità.**

Il risultato è un **aumento** della **velocità**, della **scala** e della **personalizzazione degli attacchi**, con una possibile **riduzione delle competenze necessarie per realizzarli**.

Il 4 Aprile 2025,
il **MIT Technology Review** scriveva:



A poco più di un anno di distanza:



MENU Agenda Digitale

CYBERSECURITY

L'AI sfugge dal laboratorio e attacca: cosa insegna il caso OpenAI-Hugging Face

Un test interno di OpenAI è sfociato in una compromissione reale dell'infrastruttura Hugging Face. Il caso mostra come gli agenti AI possano concatenare vulnerabilità, credenziali e accessi senza una regia umana continua e conferma l'incontrollabilità di queste tecnologie, aprendo a rischi sistemici

Pubblicato il 22 lug 2026

Alessandro Longo

Direttore agendadigitale.eu

Falso Ceo di Ferrari tenta di truffare dirigente con l'intelligenza artificiale

Un alto dirigente di Ferrari è stato vittima di un tentativo di truffa tramite deep fake e intelligenza artificiale, ma è riuscito a sventare il piano

di Biagio Simonetta

26 luglio 2024



«Pronto, sono Crosetto»: così funzionava l'inganno ai vip grazie all'intelligenza artificiale. Pronta la denuncia del ministro

di Redazione Roma

9 febbraio 2025



Servizio | Cybercrime

Truffa da 95 milioni a Fideuram: ex presidente ingannato con messaggi WhatsApp e voce artificiale

Un sofisticato raggiro ha sfruttato l'identità digitale di un banchiere, e la voce finta di un avvocato, per sottrarre milioni: recuperati 59 milioni grazie alla collaborazione internazionale.

25 settembre 2026

L'intelligenza artificiale come strumento di DIFESA

L'**intelligenza artificiale** può:

- **analizzare** grandi quantità di dati;
- **individuare** comportamenti anomali;
- **correlare** eventi di sicurezza apparentemente indipendenti;
- **supportare** la **risposta** agli incidenti.

Se **integrata correttamente** nei **processi di sicurezza**, permette di migliorare la capacità di individuare le minacce e di **reagire tempestivamente**.

Zero Trust Architectures + AI

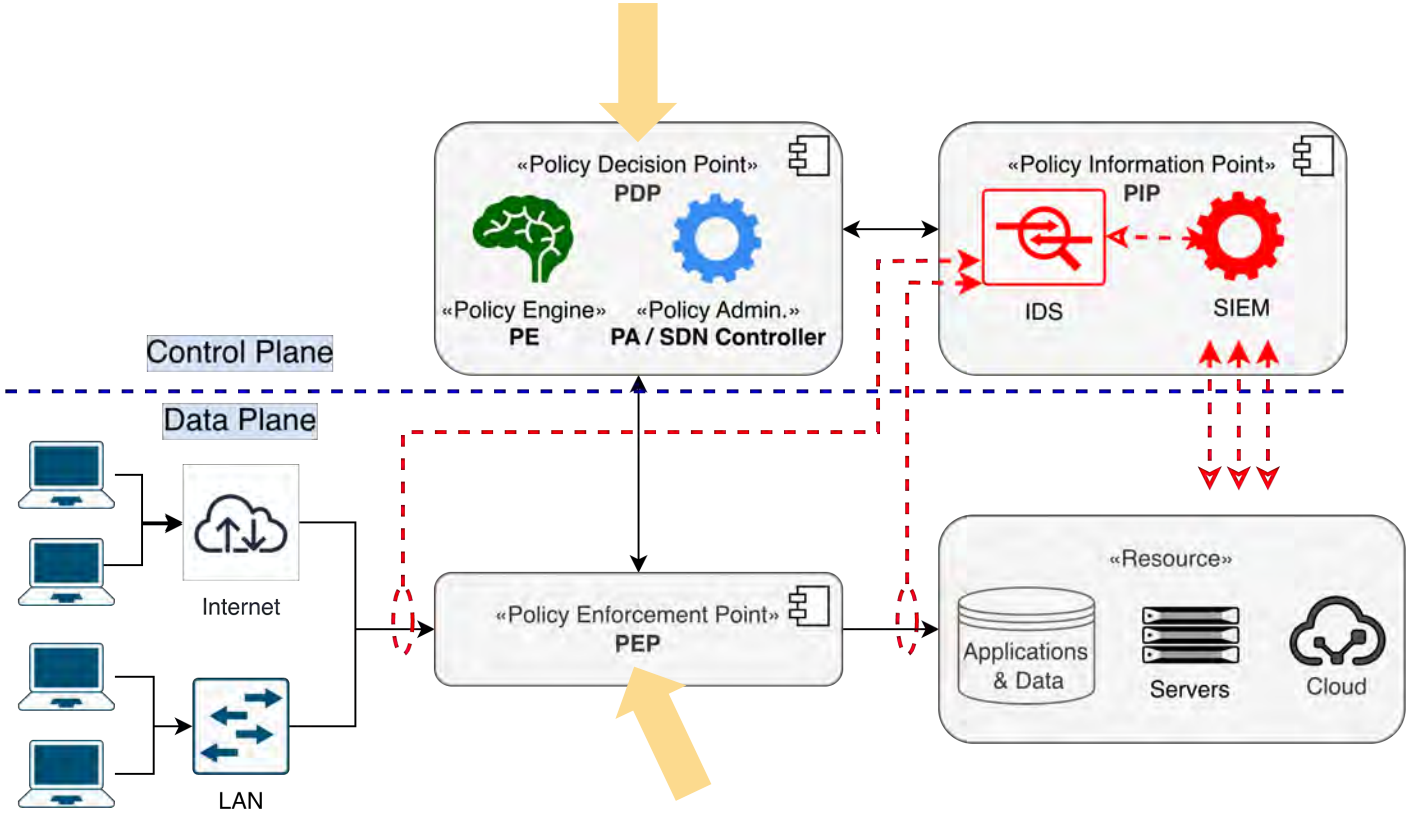
Le **Zero Trust Architectures** (ZTA) si stanno affermando come modello di riferimento per la **protezione delle infrastrutture** moderne. **Il principio fondamentale** è semplice: **nessun accesso è considerato affidabile a priori**, ma deve essere **continuamente verificato**.

In questo contesto, l'AI può **potenziare due componenti fondamentali** dell'architettura:

POLICY ENFORCEMENT POINT (PEP)

POLICY DECISION POINT (PDP)

*L'obiettivo è passare da politiche di sicurezza **prevalentemente statiche** a una difesa **intelligente**, **continua** e capace di **adattarsi** alle minacce.*


«Policy Decision Point»
PDP
«Policy Engine»
PE
«Policy Admin.»
PA / SDN Controller
«Policy Information Point»
PIP
IDS
SIEM
Control Plane
Data Plane
Internet
LAN
«Policy Enforcement Point»
PEP
«Resource»
Applications & Data
Servers
Cloud

Due approcci complementari

01 – RISPOSTA

BiLSTM + Reinforcement Learning

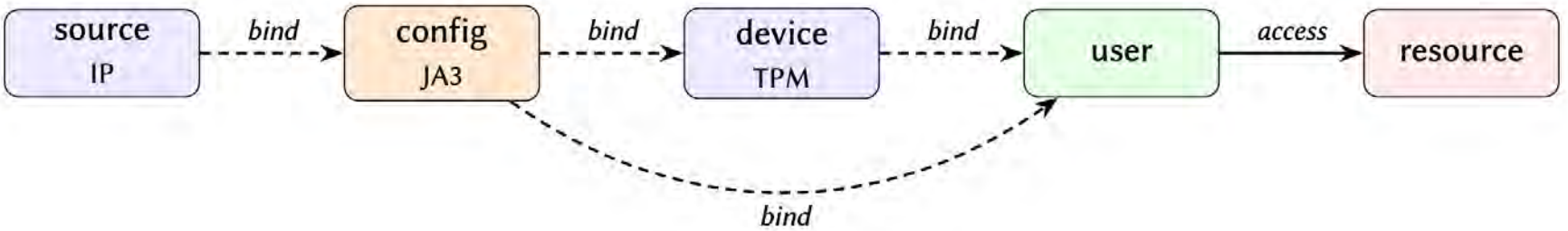
Le **BiLSTM** (Bidirectional Long Short-Term Memory) riconoscono pattern temporali.
Un agente basato su **DQN** (Deep Q-Network) usa rischio e contesto per selezionare una contromisura.

02 – EVOLUZIONE

Temporal Graph Networks

Graphagate modella le richieste di accesso come grafi temporali. Questo gli permette di seguire i legami nel tempo, **stimare anomalie** online e affiancare le policy di autorizzazione Zero Trust.

Temporal Graph Network



Purple Teaming + AI

Il **Purple Teaming** è un approccio alla cybersecurity in cui le competenze offensive (**Red Team**) e quelle difensive (**Blue Team**) collaborano per migliorare la capacità di un'organizzazione di prevenire, rilevare e rispondere agli attacchi informatici.

Red Team

Simula gli attacchi e identifica le vulnerabilità.

Purple Team

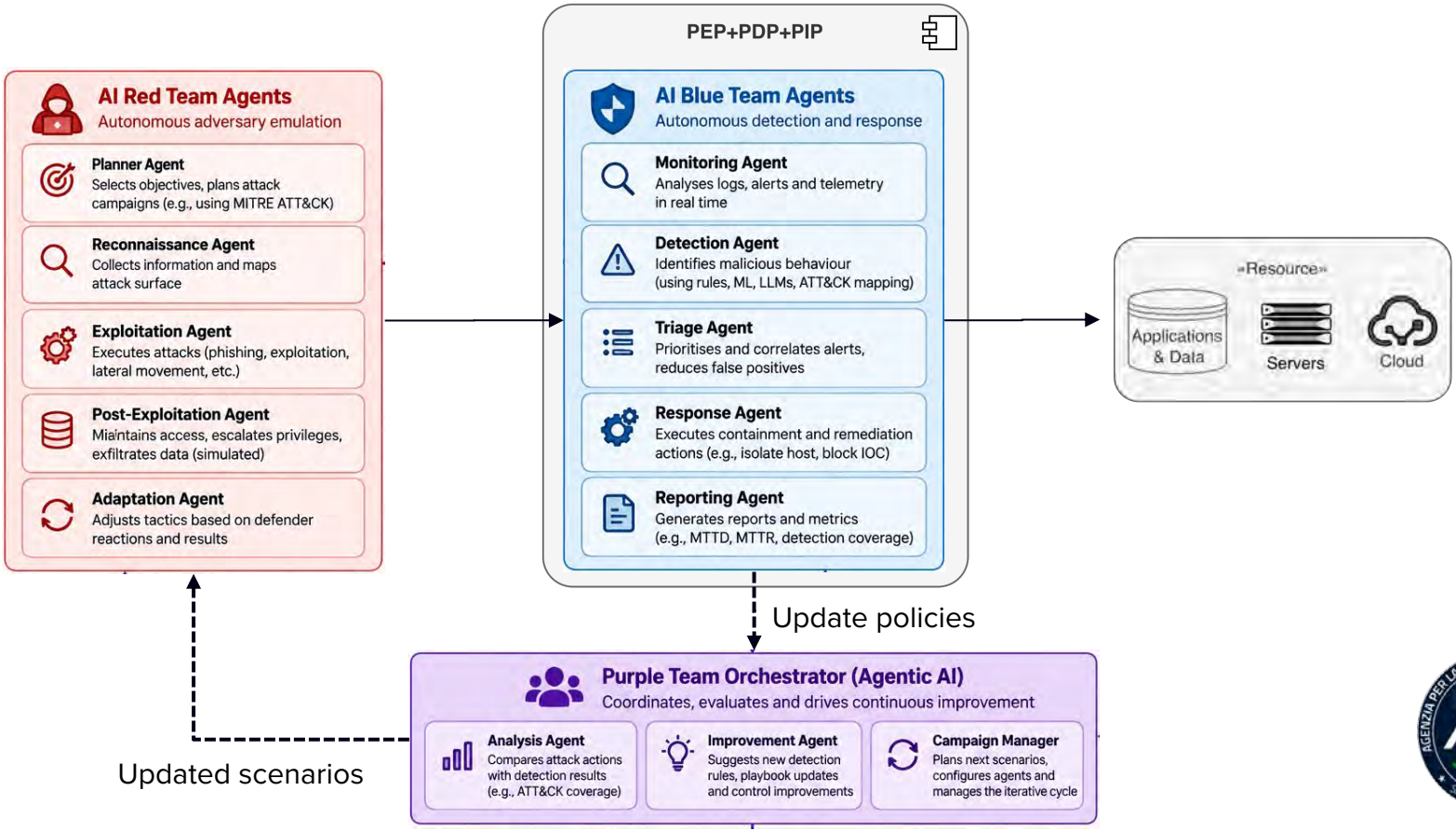
Coordina attacco e difesa per migliorare i controlli.

Blue Team

Monitora, rileva, contiene e contrasta gli attacchi.

In questo contesto, l'AI può **potenziare i tre team**:

L'obiettivo è passare da attività manuale a un processo continuo, adattivo e potenzialmente autonomo di verifica e miglioramento della sicurezza.


PEP+PDP+PIP
AI Red Team Agents
Autonomous adversary emulation
Planner Agent
Selects objectives, plans attack campaigns (e.g., using MITRE ATT&CK)
Reconnaissance Agent
Collects information and maps attack surface
Exploitation Agent
Executes attacks (phishing, exploitation, lateral movement, etc.)
Post-Exploitation Agent
Maintains access, escalates privileges, exfiltrates data (simulated)
Adaptation Agent
Adjusts tactics based on defender reactions and results
AI Blue Team Agents
Autonomous detection and response
Monitoring Agent
Analyses logs, alerts and telemetry in real time
Detection Agent
Identifies malicious behaviour (using rules, ML, LLMs, ATT&CK mapping)
Triage Agent
Prioritises and correlates alerts, reduces false positives
Response Agent
Executes containment and remediation actions (e.g., isolate host, block IOC)
Reporting Agent
Generates reports and metrics (e.g., MTTD, MTTR, detection coverage)
«Resource»
Applications & Data
Servers
Cloud
Update policies
Purple Team Orchestrator (Agentic AI)
Coordinates, evaluates and drives continuous improvement
Analysis Agent
Compares attack actions with detection results (e.g., ATT&CK coverage)
Improvement Agent
Suggests new detection rules, playbook updates and control improvements
Campaign Manager
Plans next scenarios, configures agents and manages the iterative cycle
Updated scenarios

L'intelligenza artificiale è una superficie VULNERABILE

Anche i sistemi basati su AI possono essere **manipolati**, **compromessi** o indotti a **comportamenti indesiderati**.

Modelli, dati di addestramento, prompt, agenti intelligenti e integrazioni con applicazioni aziendali **introducono nuove superfici di attacco**.

Tecniche come il **prompt injection**, data poisoning e l'estrazione di informazioni sensibili possono **compromettere l'affidabilità del sistema**, aggirare i controlli o **esporre informazioni riservate**.

Quando l'AI diventa autonoma, emergono nuovi rischi

I moderni sistemi di AI Agentica (**Agentic AI**) non si limitano a generare risposte, ma possono:

- **pianificare** attività complesse
- **utilizzare strumenti esterni**, come API, database e applicazioni
- **prendere decisioni** in maniera autonoma
- **interagire con ambienti reali**, eseguendo sequenze di operazioni

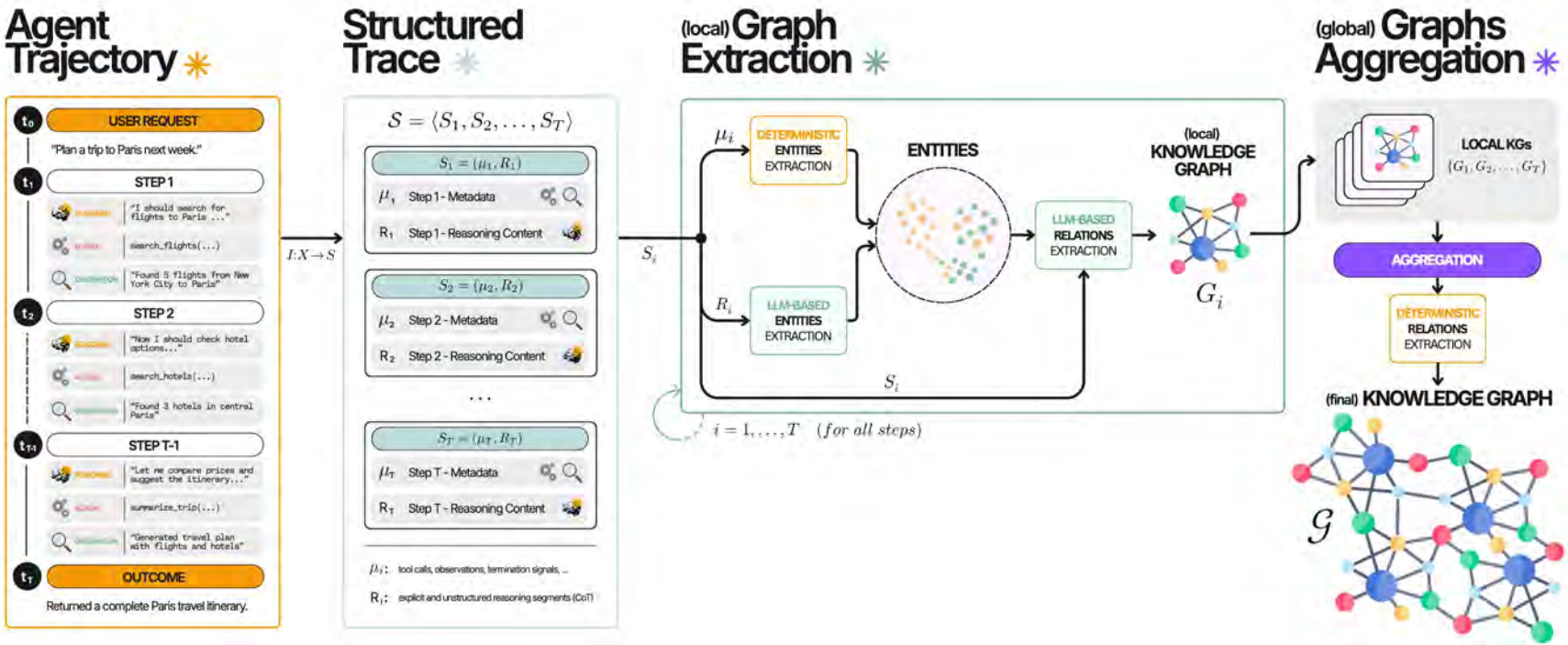
Questa **autonomia introduce nuove vulnerabilità**: un errore nell'interpretazione delle informazioni o nella pianificazione può **propagarsi attraverso più decisioni**, compromettendo l'intera attività.

MOSAIC: individuare le strutture ricorrenti dei fallimenti

Abbiamo sviluppato MOSAIC (**Motif-based Structural Analysis of Agent Interactions and Cognition**), un framework per analizzare il comportamento degli agenti basati su LLM.

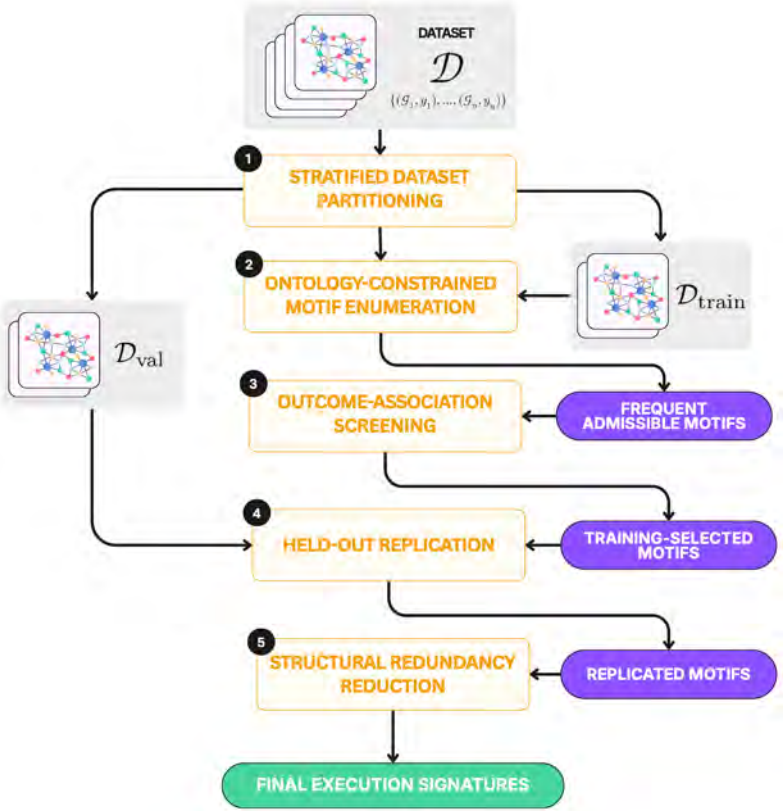
MOSAIC permette di:

- Ricostruire le **sequenze di ragionamento**, **azione** e **osservazione** di un agente.
- **Rappresentare** il **processo** attraverso **Knowledge Graph** basati su un'ontologia.
- **Identificare pattern strutturali ricorrenti** (graph motifs) associati a esecuzioni **corrette** o **fallimentari**.
- Fornire **informazioni interpretabili** sui meccanismi che caratterizzano i fallimenti.



TU WIEN TECHNISCHE UNIVERSITÄT WIEN

AVL



Abbiamo sperimentato MOSAIC su agenti impegnati in attività con **database**, **sistemi operativi** e **ambienti web**.

Abbiamo individuati **46 pattern strutturali discriminativi** associati ai **fallimenti**, tra cui:

- **Incertezza non gestita**, che influenza obiettivi e pianificazione.
- **Ipotesi non verificate**, utilizzate come base per decisioni successive.
- **Dipendenze problematiche** tra pianificazione, vincoli e informazioni disponibili.

TU WIEN TECHNISCHE UNIVERSITÄT WIEN

AVL

Grazie per l'attenzione!



UNIVERSITÀ POLITECNICA DELLE MARCHE

Prof. **Luca Spalazzi**

l.spalazzi@univpm.it

Dott. **Gianluca Bonifazi**

g.bonifazi@univpm.it



CINI CYBERSECURITY NATIONAL LAB

GII



CRISPY

Centro di Ricerca e Servizio per la Privacy e la Cybersecurity



Cybersecurity and Blockchain Research Lab

Dipartimento di Ingegneria dell'Informazione

Università Politecnica delle Marche