

RITE 2026

Romanian Internet Technologies Event (RITE)

The New AI Landscape: From Intelligent Models to Trustworthy Autonomous Systems

Prof. Adina Magda Florea

Director of the AI Center of Excellence at POLITEHNICA Bucharest

OCTOBER 27th 2026

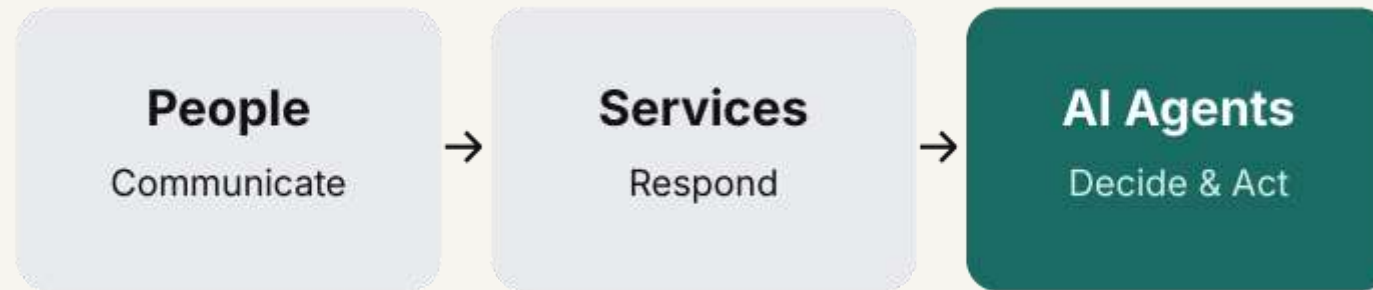
Internet of Information vs. Internet of Intelligent Action

THE TECHNOLOGICAL SHIFT

The Internet is evolving from connecting people, information and services to connecting intelligent systems that make decisions and act

THE CHALLENGE

The next frontier of AI is not just greater intelligence, but to make autonomous intelligence trustworthy

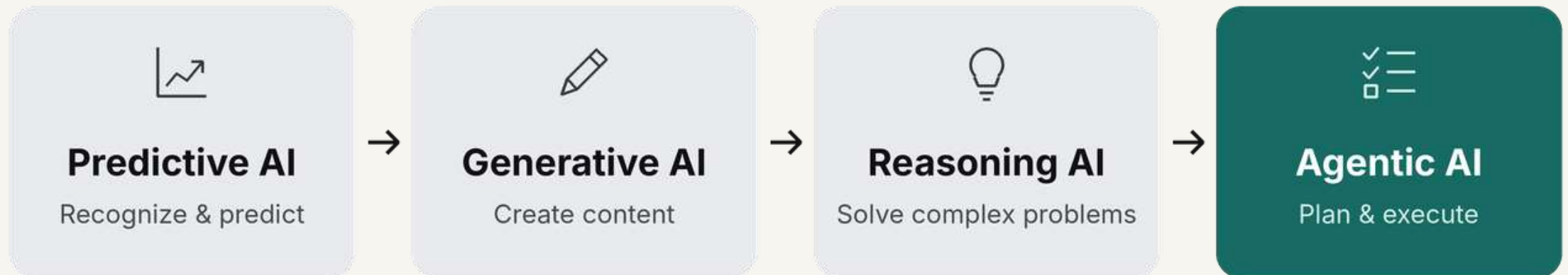


THE STRATEGIC OPPORTUNITY

The future of AI will be shaped not only by those who develop more powerful models, but by those who can build reliable intelligent systems and the ecosystems that sustain them

The AI Paradigm Shift: From Prediction to Autonomous Action

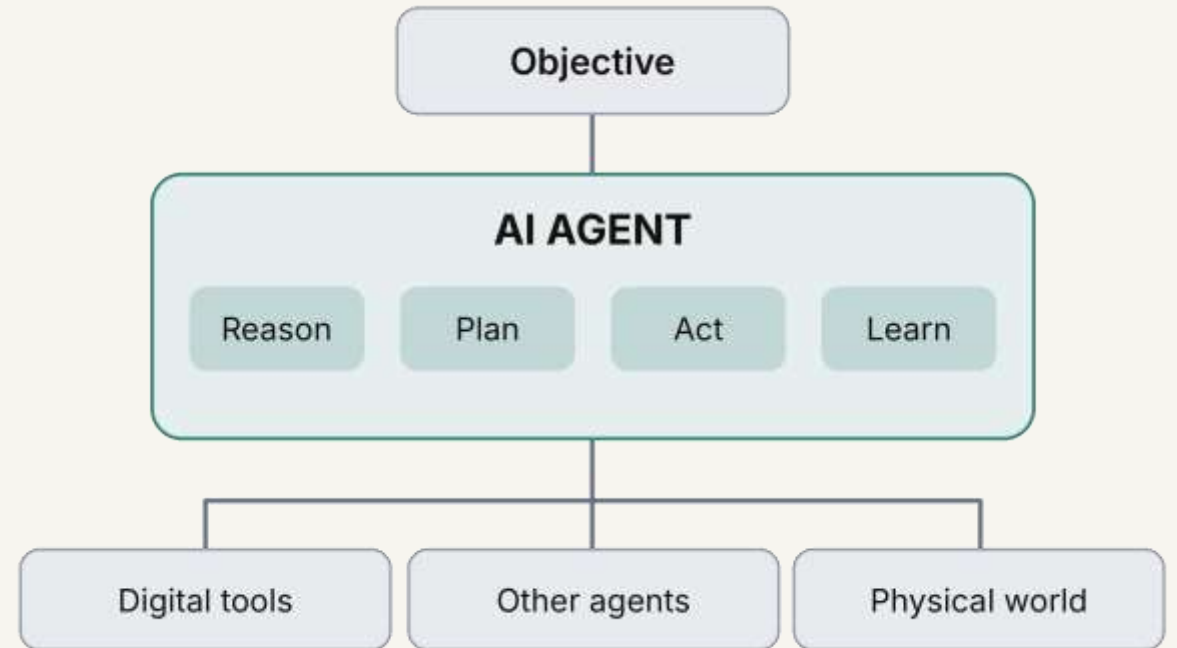
- From recognizing patterns to generating knowledge
- From generating responses to planning and executing tasks
- From isolated models to intelligent systems interacting with their environment



AI Agents are Now Autonomous Actors

Assist → Recommend → Execute with approval → Execute within constraints

- We are moving from telling computers how to perform every operation to specifying objectives and supervising how intelligent systems achieve them
- Autonomy must be bounded according to risk, context, and human control



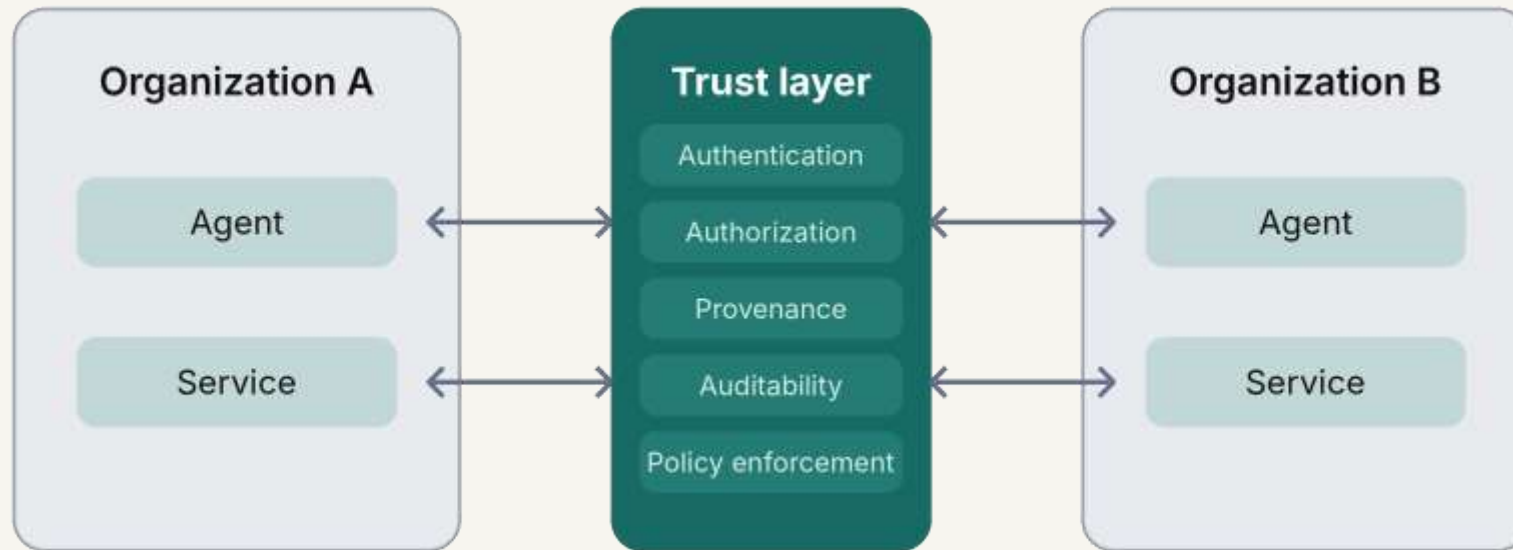
AUTONOMOUS COLLABORATION CREATES NEW FORMS OF DISTRIBUTED INTELLIGENCE

The Internet of AI Agents

A NEW LAYER OF DIGITAL INTERACTION

Agents collaborating across organizations and services

- Who is responsible when autonomous agents exchange information, negotiate decisions, and trigger actions across organizational boundaries?
- Issues beyond traditional networking: authentication, authorization, provenance, auditability, policy enforcement, and system-level robustness



TRUST MUST EXTEND FROM NETWORKS AND USERS TO INTELLIGENT AGENTS

Trustworthy Agents

Trustworthy agents require trustworthy infrastructure, and trustworthy infrastructure increasingly requires understanding agent behavior



Identity

Who is the agent? Who authorized it?



Security

What can it access and execute?



Resilience

Can the system withstand failure or attack?



Accountability

Can actions be traced, verified, and controlled?

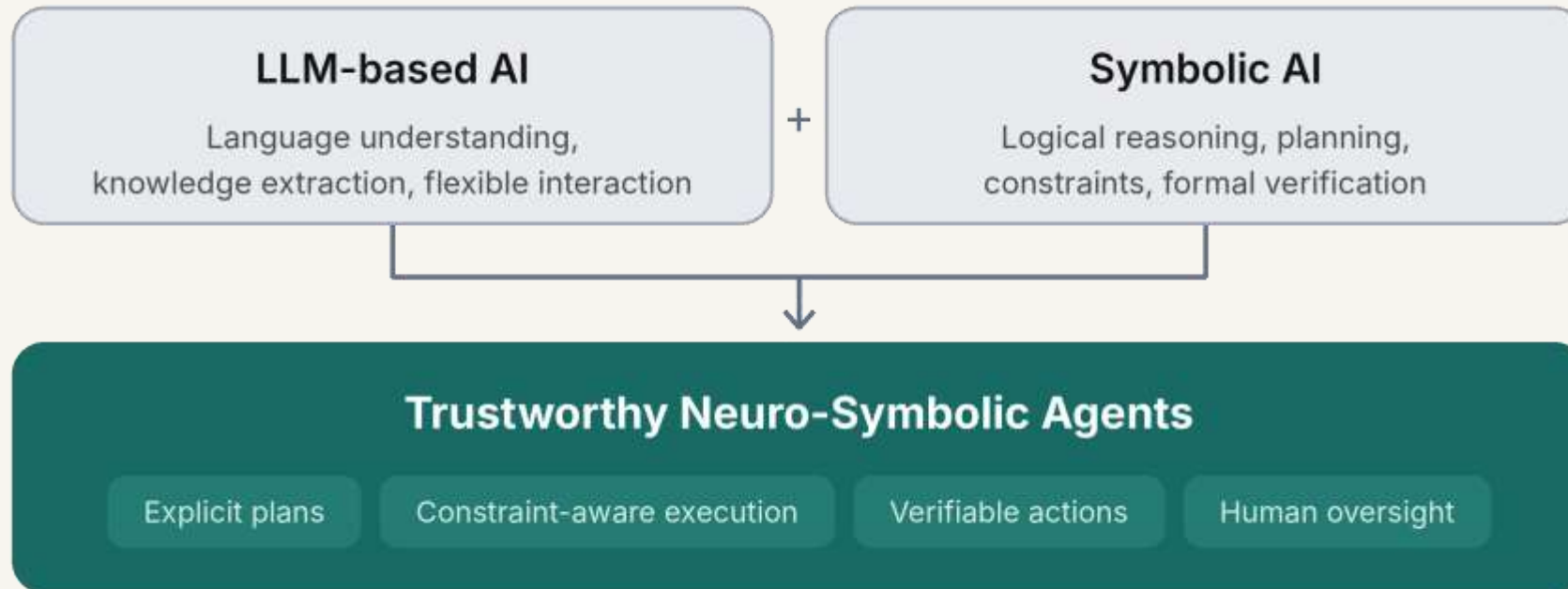
**WHEN AI AGENTS BECOME ACTORS, TRUST MUST EXTEND BEYOND THE
USER AND THE NETWORK**

Two dimensions of trustworthiness

	Behavioral trustworthiness	Institutional trustworthiness
Main question	Is the agent's behavior correct, safe, and compliant with specified constraints?	Is the agent authorized to take this action, under the relevant policies and delegated authority?
Research topics	Neuro-symbolic reasoning, planning, constraints, verification	Identity, delegated authority, runtime policy enforcement, traceability
Challenge	Reliable decision-making and execution	Accountable autonomous action

Neuro-Symbolic Models for Trustworthy Agents

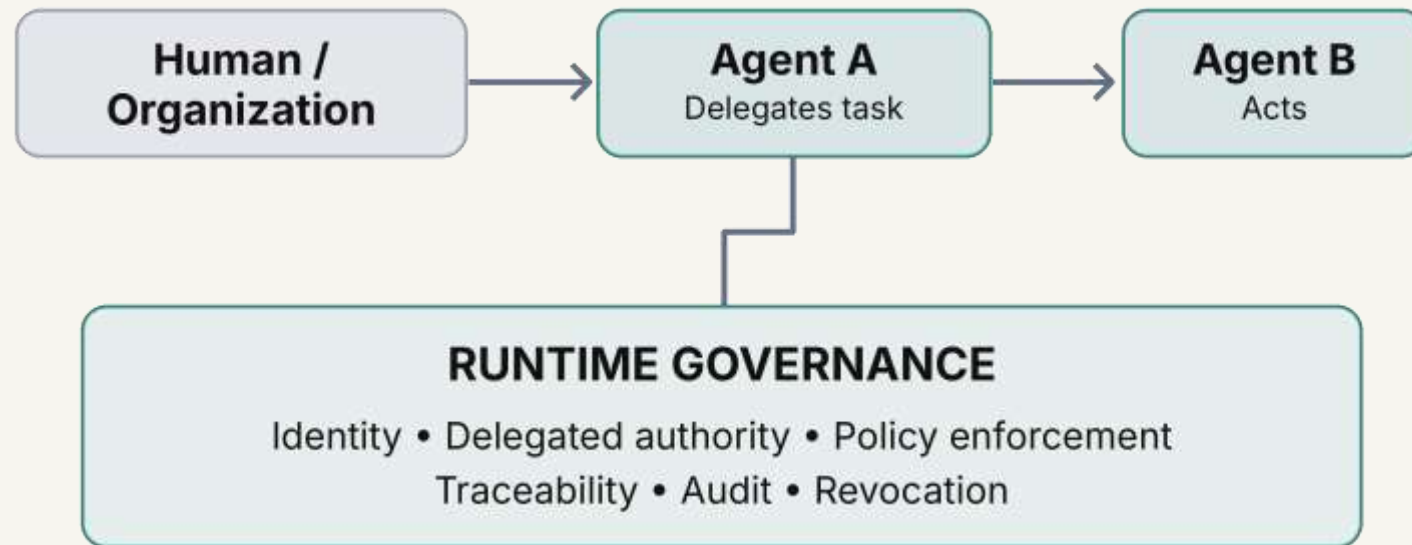
- How to move beyond agents that rely primarily on large language models?
- Combine the flexibility of **generative AI** with the reliability and controllability of **symbolic AI**
- Integrate heterogeneous reasoning mechanisms into a coherent agent architecture
- Expected result: Make the agents' behavior more predictable, inspectable, and controllable



**TRUSTWORTHY AI AGENTS REQUIRE MORE THAN FOUNDATION MODELS
THEY REQUIRE EXPLICIT REASONING, CONSTRAINTS, VERIFICATION, AND CONTROL**

Governable Systems of AI Agents

- Autonomous agents begin collaborating across systems and organizations
- Who authorizes autonomous action of agents?
- Develop a runtime governance layer; governability must be an architectural property
- Expected result: Enable collaboration and delegation among autonomous agents without losing accountability and human control



WHO?

Authorize the action

UNDER WHAT CONDITIONS?

Constraints and policies

WITH WHAT RIGHTS?

Scope of delegated authority



Cofinanțat de
Uniunea Europeană



Programul Creștere Inteligentă, Digitalizare și Instrumente
Financiare 2021-2027 (PoCIDIF), co-finanțat prin Fondul European
de Dezvoltare Regională (FEDR)

HUB Român de Inteligență Artificială

Cod SMIS 2021- 351416

Romanian Hub for Artificial Intelligence: Building Romania's Capacity for Trustworthy and Autonomous AI

7

Universities

8

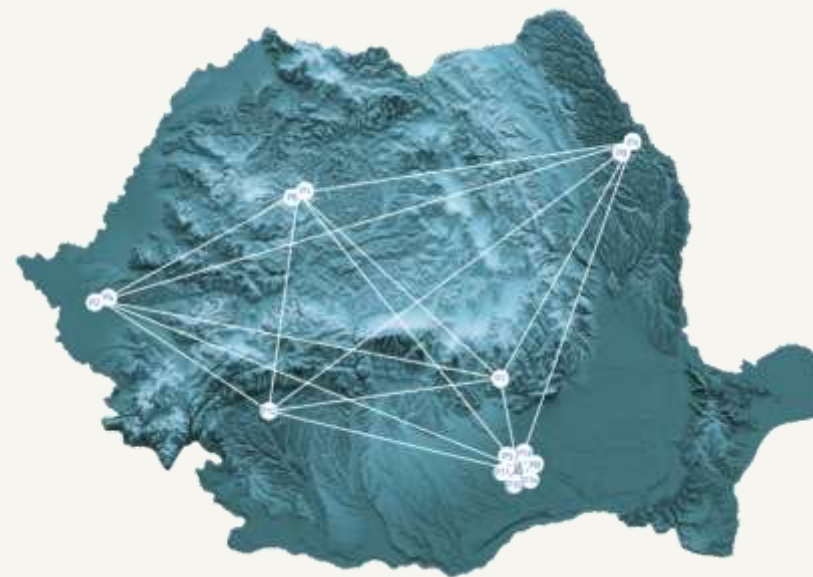
Companies

8

Research directions

HRIA is a national ecosystem connecting excellence in AI
research with the needs of industry and society

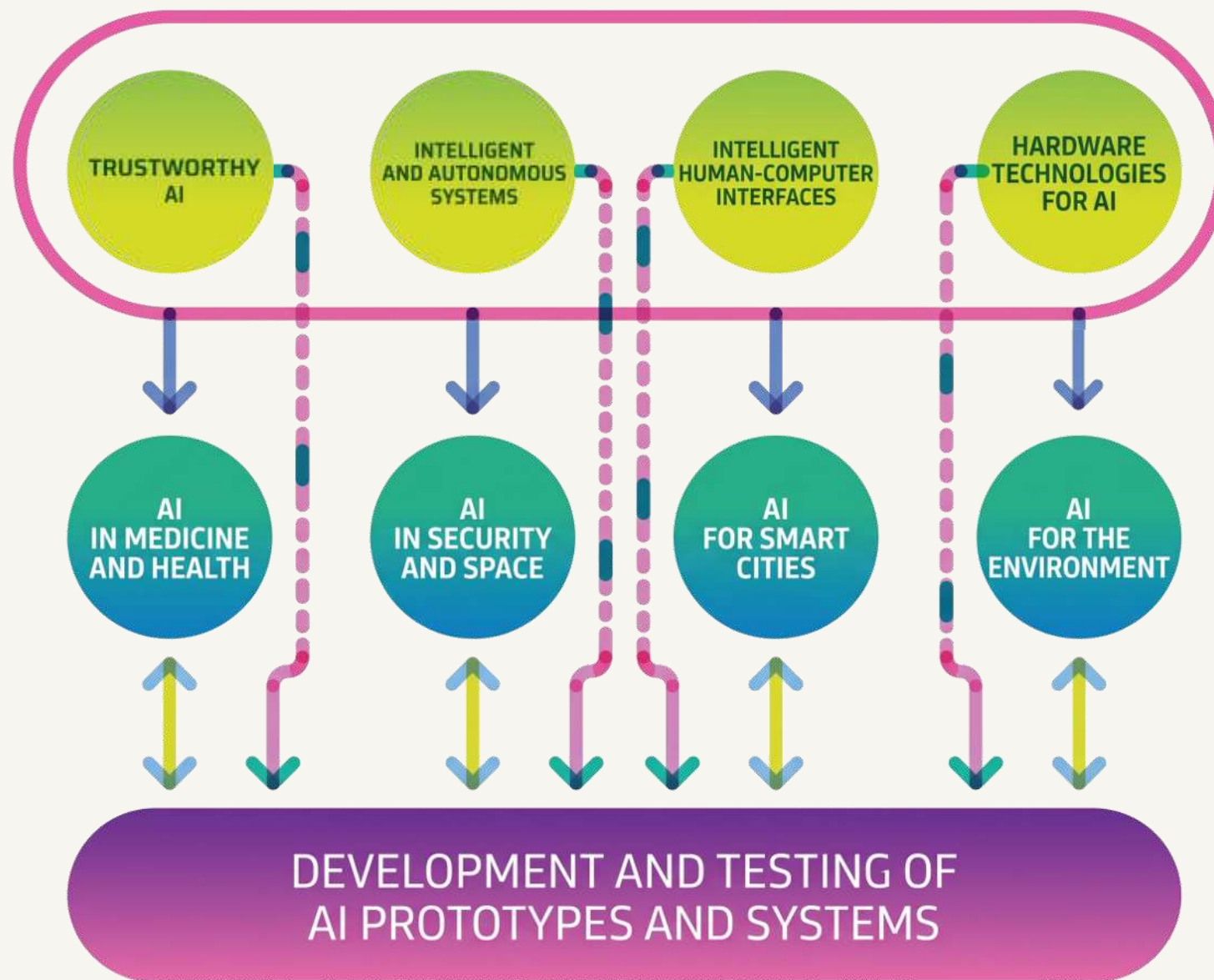
<https://hria.upb.ro/>





HUB Român de Inteligență Artificială

Cod SMIS 2021- 351416



Perspectives for a New Generation of Autonomous Systems

- **Agents' autonomy must be built on trust**
- **Agents must be trustworthy by design** — Integrate verification, runtime monitoring, safety constraints, and human oversight into autonomous execution
- **Beyond LLM-centric agents** — Combine foundation models with symbolic reasoning, planning, constraint solving, and verification
- **Governable autonomy** — Establish agent identity, explicit delegated authority, accountability, and traceable chains of action
- **Beyond traditional XAI** — Develop role-aware, context-sensitive explanations of agent decisions, plans, and actions
- **Necessity-driven AI architectures** — Use foundation models where they add value, considering reliability, cost, efficiency, and sustainability

Thank you
adina.florea@upb.ro