General-purpose AI and the AI Act: Risk Assessment and Societal Safety

The advent of the Artificial Intelligence (AI) Act by the European Commission marks a transformative era in the realm of AI governance. Focused extensively on general-purpose AI systems (GPAIs), the AI Act sets forth comprehensive guidelines aimed at aligning AI technologies with principles of human dignity, rights, and trust. The project endeavors to address these regulatory challenges by developing a robust framework for assessing and managing the risks associated with these systems. Through an agile, multi-layer risk taxonomy and compliance methodologies, the research seeks to reinforce the integrity and ethical alignment of general-purpose AI models.

Project Objectives

The project is activated within the context of the Innovation Lab agreement between Huawei and the University of Bologna. This project is committed to exploring various dimensions of AI risk management, structured around key research challenges and objectives. The central goals include:

1. **Risk Assessment**: Evaluating the intricate risks posed by general-purpose AI systems, particularly when these systems are integrated as components within high-risk domains.
2. **Risk Taxonomy Development**: Crafting an adaptive risk taxonomy that can effectively categorize and manage GPAI-related risks, catering to the nuances between generic and systemic models.
3. **Comprehensive Assessment Methodology**: Employing both quantitative and qualitative measures to develop a robust risk assessment framework, encompassing diverse impact metrics ranging from public health to cybersecurity.
4. **Socio-Technical Mitigation Strategies**: Proposing risk mitigation solutions that conform to the AI Act and resonate with broader legal frameworks like the General Data Protection Regulation (GDPR).

These objectives are crucial to ensuring that AI systems meet compliance benchmarks, thereby safeguarding public interests amidst rapid technological progression.

Research Challenges

The project is structured to tackle several research challenges:

- **Emerging AI Principles**: Understanding and integrating emerging AI principles is pivotal as GPAIs introduce unique complexities that demand nuanced governance approaches.
- **Legal/Governance Framework Analysis**: The project will analyze existing legal frameworks, with a central focus on the EU AI Act, while also considering influencing governance models.
- **AI Lifecycle Mapping**: Identifying and mapping the roles and accountabilities of AI actors throughout the AI lifecycle is essential for comprehensive risk management.

- **Risk Taxonomies/Catalogues**: Developing detailed, layered taxonomies that consider legal, operational, safety, and societal risks, and can adapt to various application scenarios.
- **Risk Management Methods**: Establishing a multi-stage risk management process covering risk identification, assessment, and mitigation.
- **Benchmarking for AI Principles**: Investigating benchmarks to evaluate GPAI alignment with established AI principles, aiming for some automation in the evaluation process.

Implementation Timeline and Tasks
The project is divided into phases, each encompassing distinct tasks and deliverables:

- **Phase 1 (T to T + 2 months)**: Focuses on surveying the GPAI risk management environment and industry practices, resulting in an insight report detailing regulatory requirements and industry risk governance understanding.
- **Phase 2 (T+3 to T + 6 months)**: Entails developing a structured GPAI Risk Taxonomy Model V1.0, integrating insights from international standards and mitigation strategies.
- **Phase 3 (T+7 to T + 9 months)**: Optimizes the Risk Taxonomy Model based on the EU GPAI Code of Practice, producing a comprehensive baseline draft of identified risks.
- **Phase 4 (T+10 to T + 11 months)**: Involves interpreting the EU GPAI Code of Practice to lay down compliance requirements and strategic action suggestions.