Operationalizing Human Oversight under the EU AI Act: A Five-Dimensional Framework

欧盟 AI 法案下人类监督的可操作化：一个五维框架

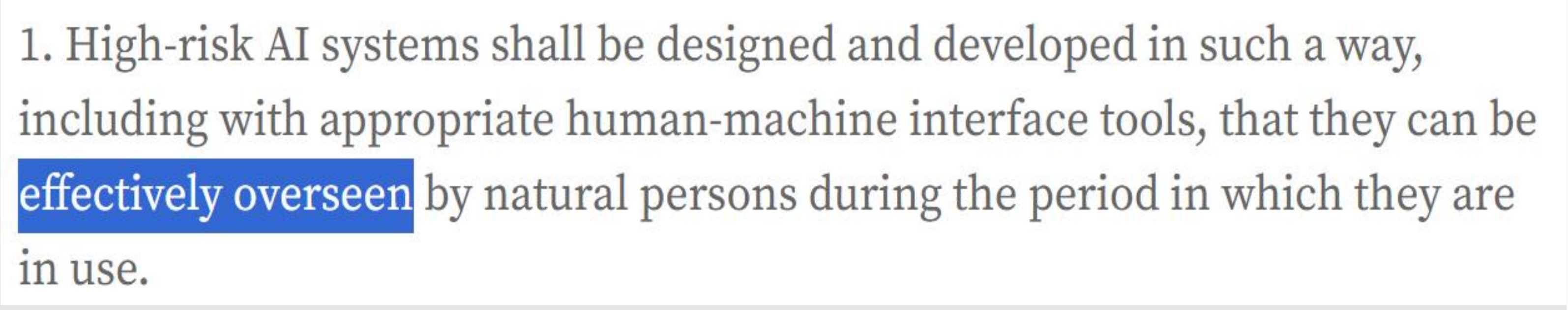
- Author: Xiaotong Sun

- Affiliation: University of Turku

Context & The Gap

1 Introduction: The Legal Mandate

- Article 14 of EU AI Act: Establishes “Human Oversight” as a cornerstone for high-risk AI governance.
- **Goal:** Ensure AI systems are overseen by natural persons to prevent risks to health, safety, and fundamental rights.
- **Key Requirement:** Oversight must be “Effective” (not just symbolic).



1. High-risk AI systems shall be designed and developed in such a way, including with appropriate human-machine interface tools, that they can be effectively overseen by natural persons during the period in which they are in use.

2 Gaps: From Legal Ideal to Practical Illusion

Technical Illusion

- **Functional Opacity**:High-speed, opaque systems trap humans in passive roles.
- **"Strange Loop"**:Overseers cannot explain outputs.

Procedural Illusion

- **Undefined Triggers:** Lack of criteria for when to intervene or override.
- **Disconnected Loops:** Missing feedback channels for recording interventions or reporting system faults upstream.

Cognitive Illusion

- **Automation Bias:** Over-reliance on machine outputs.
- **Cognitive Load:** Fatigue and complexity overwhelm human capacity (MABA-MABA).

Institutional Illusion

- **Responsibility Gap:** Overseers bear formal liability but lack actual authority, resources, or training.
- **Culture Clash**: Profit and efficiency prioritized over robust safety.

Societal Illusion

- **Product-Safety Logic:** Fundamental rights protection is narrowed down to technical product safety checks.
- **Asymmetry**: Public lacks transparency and access to decisions.

Technical Effectiveness

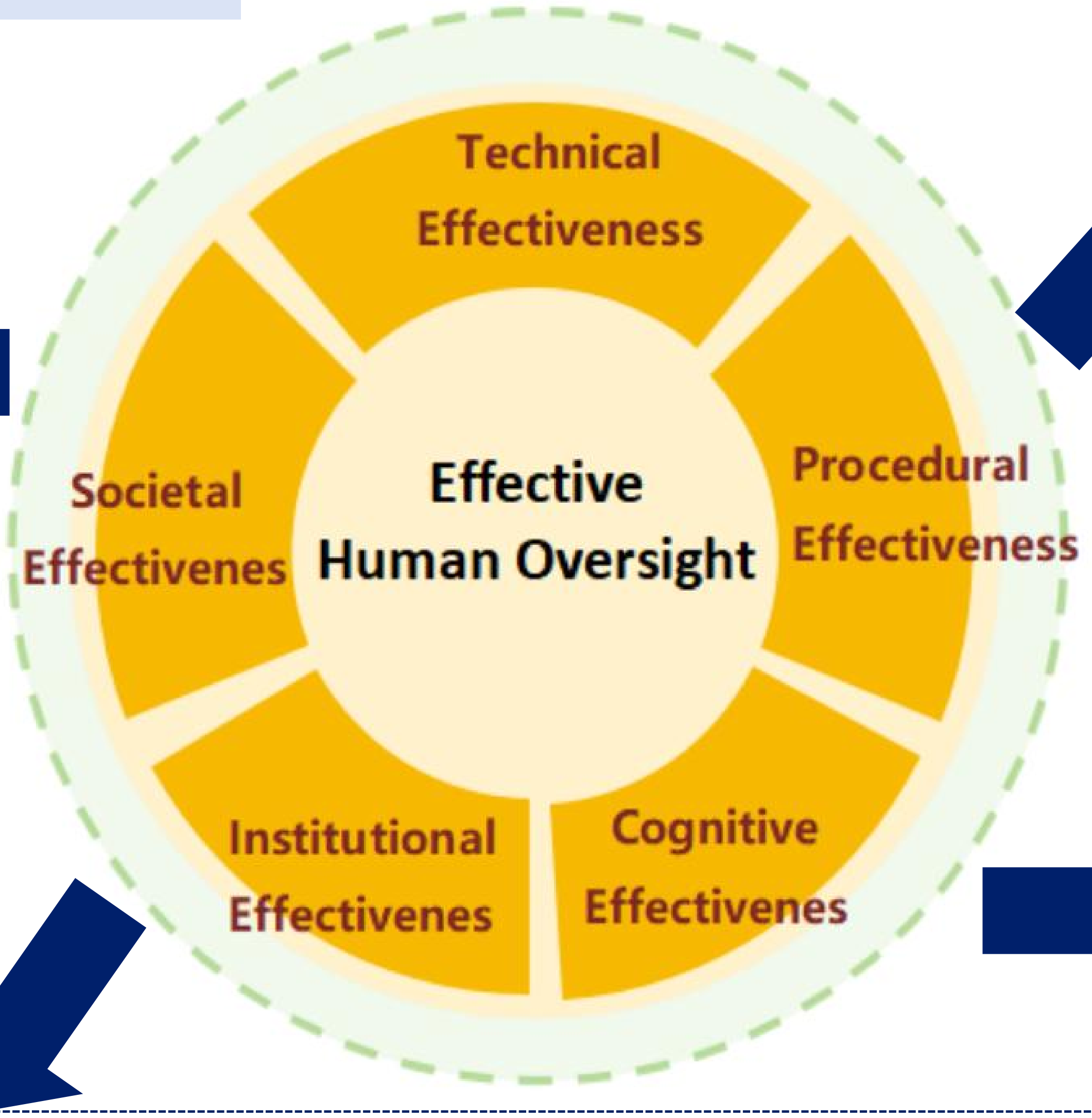
Principle: Effectiveness begins with functions, tools, and interfaces appropriately designed for high-risk AI systems.

- Meaningful Explainability
- Fail-Safe and Degradation Mechanisms
- Informative and Dynamic Interfaces

Procedural Effectiveness

Principle: Just, transparent, and inclusive procedural arrangements turn the mere possibility of oversight into reliable practice.

- Concrete intervention thresholds
- Documentation and Communication
- Training and Competence Maintenance



Societal Effectiveness

Principle: The system must be accountable to the public through transparent processes and redress.

- External Transparency and Public Accountability
- Accessibility of Redress
- Participatory Governance

Cognitive Effectiveness

Principle: Reliable practice is useless if overseers cannot exercise independent and critical judgment.

- AI Literacy
- Cognitive Resilience
- Critical Mindset

Institutional Effectiveness

Principle: Human overseers’ authority must be cultivated by supportive structures, resources, and an oversight-valued culture.

- Due Diligence and Support (Provider)
- Reshape Governance and Authority (Deployer)
- Cultivate a Just Culture (Both)

Funded by the European Union

HAIF

UNIVERSITY OF TURKU