Glass-Box Compliance: Ontological Reasoning for Regulatory Risk Classification

This project explores how **ontological modeling and logical reasoning** can be used to support regulatory risk classification for AI systems, using Article 6 and Annex III of the EU AI Act as a motivating domain. Rather than treating compliance as a probabilistic labeling problem, the project reframes it as a **traceable reasoning task grounded in realist ontology**.

The central question is not “Can we predict whether a system is high-risk?” but rather “Can we make explicit *why* a system would qualify as high-risk, and under what assumptions?”

Motivation and Problem Context

Regulatory frameworks are written in natural language, while AI systems are described through a mix of technical documentation, policy language, and vendor claims. This creates a persistent gap between legal interpretation and technical system description.

Many existing approaches attempt to bridge this gap using keyword search, similarity scoring, or embedding-based classification. While these methods are useful for discovery, they struggle to provide **deterministic guarantees or clear justification**, which are essential in regulatory and audit contexts.

The motivation for this project is to explore whether a semantic pipeline grounded in ontology and logic can help structure this problem in a way that is inspectable, explainable, and reusable across domains.

Conceptual Approach

The project implements a **glass-box compliance pipeline** that separates three concerns:

1. Interpretation of unstructured text
2. Ontological representation of system capabilities
3. Logical inference over those representations

This separation is deliberate. It avoids collapsing interpretation, representation, and reasoning into a single opaque step, and instead makes each stage explicit and reviewable.

The result is not an automated compliance engine, but a framework for **making assumptions visible and reasoning steps explicit**.

Role of Language Models

Large Language Models are used only at the boundary between text and structure. Their role is limited to **candidate extraction** from unstructured documents, such as identifying language that suggests the presence of a particular capability (for example, biometric identification).

Crucially, these outputs are treated as *proposals*, not conclusions. Whether a proposed capability is accepted and asserted in the ontology remains a human decision. The ontology and reasoner operate only on explicitly asserted facts.

This design reflects a clear epistemic boundary: language models assist with interpretation, while **symbolic reasoning governs validation**.

Ontological Grounding

Once a capability is accepted, it is modeled using a **BFO-aligned ontology**.

AI systems are treated as real-world entities, and system capabilities are modeled as **dispositions**. This choice is intentional. A system may pose regulatory risk because of what it is capable of doing, even if that capability is not currently exercised.

By modeling capabilities as dispositions rather than behaviors, the ontology captures **latent risk**, not just observed activity. Regulatory provisions are modeled as information content entities that are explicitly *about* these capability universals, preserving a clean separation between reality-side entities and representation-side artifacts.

Reasoning and Inference

Risk classification is derived using OWL restrictions and a description logic reasoner. The reasoner does not decide policy outcomes; it verifies whether a conclusion is **logically entailed** by the asserted premises.

For example, if:

- a system bears a biometric identification capability, and
- Annex III provisions are about that capability,

then a high-risk determination follows as a logical consequence of the model.

Because the inference is logical rather than statistical, it is reproducible and inspectable. If the outcome is disputed, the disagreement can be traced to specific premises rather than hidden model behavior.

Outputs and Value

The primary output of the system is not a label, but a **justification structure**. The model makes explicit:

- which system capability mattered,
- which regulatory content applied,
- and how the inference was derived.

This enables auditability and supports human review. Stakeholders can challenge assumptions, refine definitions, or revise mappings without obscuring the reasoning process itself.

Broader Significance

This project is not intended to automate legal judgment or replace regulatory decision-making. Its contribution lies in demonstrating how **ontological realism and logical reasoning** can be used to structure complex regulatory analysis in a way that is transparent and defensible.

More broadly, the approach suggests a path forward for integrating AI systems into governance contexts without relying on black-box classification. By combining linguistic interpretation with formal ontology and reasoning, compliance analysis can move from document handling toward **computable, inspectable reasoning processes**.