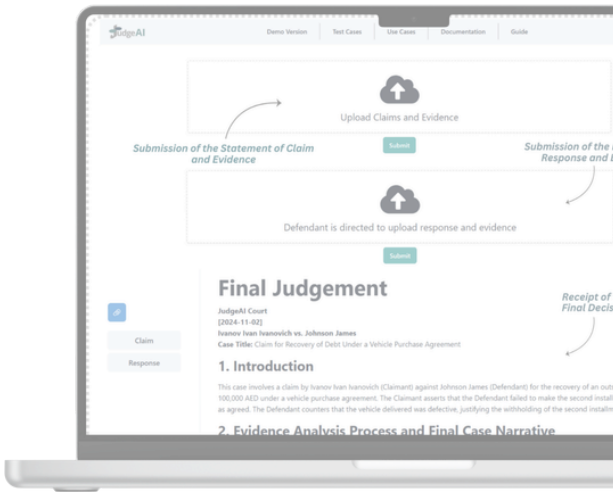
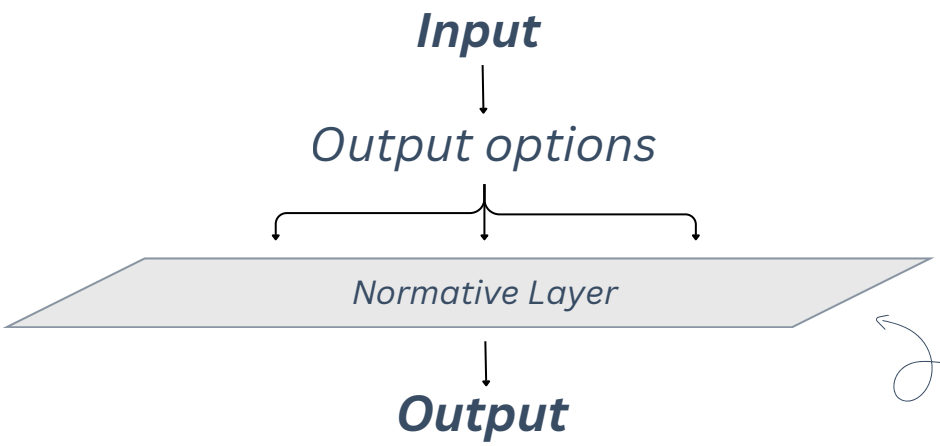
Project Overview

JUDGEAI



2026

JudgeAI focuses on developing a domain-agnostic normative decision layer for AI systems that enforces a unified logic across automated dispute resolution, lawmaking, and autonomous agent decisions.



[context]

Current AI systems perform effectively in the assistant setting, where normative judgment remains with the human. AI development is moving toward agents, systems that independently choose actions and produce outcomes with direct consequences for real people. This transition creates a new requirement inside AI itself, the presence of a normative layer that determines which outputs and actions are permissible or required in a given situation.

Current approaches in alignment, including Constitutional AI and Superalignment, are designed to align model responses with predefined rules and constraints. An AI agent solves a different problem. It must choose between alternative actions that lead to different consequences for different actors. This requires a formal criterion for selecting between alternatives based on their consequences, which makes the task inherently normative and requires a separate layer.

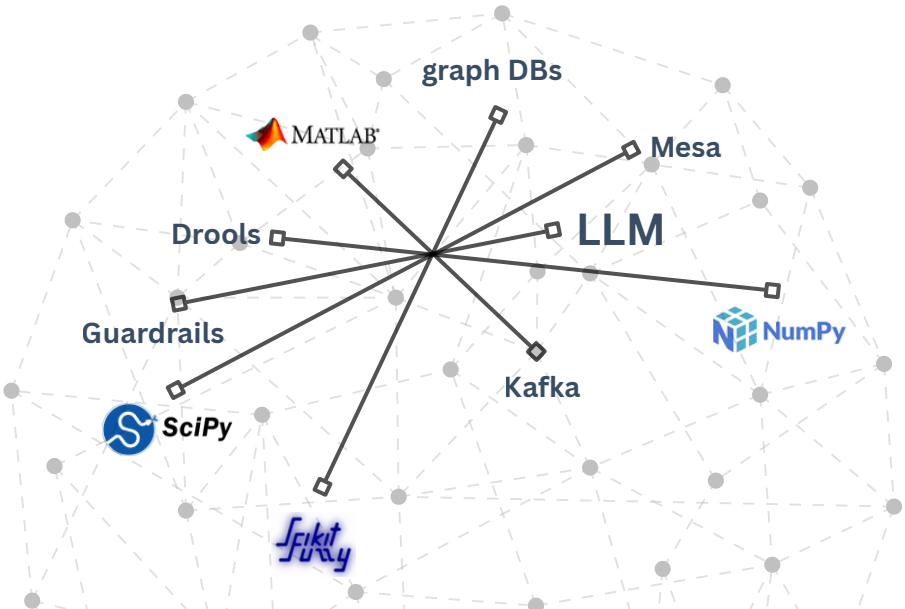
JudgeAI focuses on this problem. The project develops a domain agnostic normative decision layer for AI systems by constructing a formal legal model of the world from empirical premises such as the structure of situations, the behavior of actors, and the causal consequences of alternative decisions. On this basis, JudgeAI forms normative judgment for AI systems as an analogue of morality that is fully traceable and falsifiable. This layer provides the underlying normative foundation that precedes any specific legal system and enables a unified approach across automated dispute resolution, lawmaking, and autonomous agent decision making.

[approach]

To address the identified problem, JudgeAI implements a formalized system of normative reasoning designed for the autonomous formation of admissible legal decisions. The system organizes the transformation of empirical reality into normative conclusions through a strictly defined sequence of logically connected stages, each of which supports verification and reproducible execution.

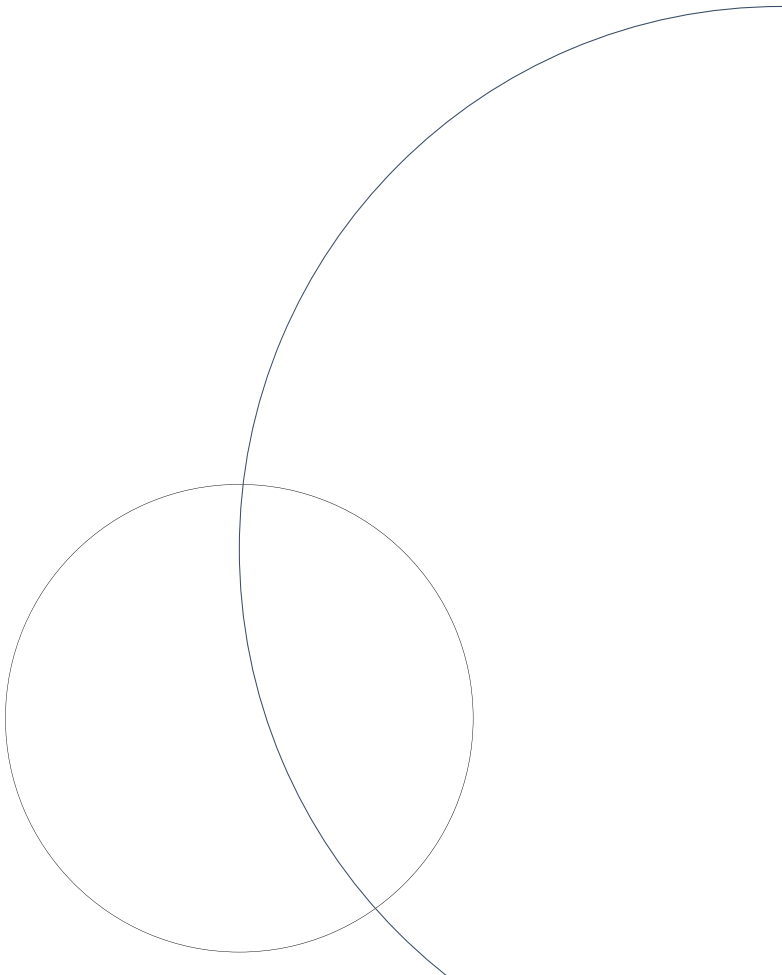
Within this approach, formalized normative logic defines how legal relevance of facts is established, how stakeholder interests are structured and correlated, how causal relationships are constructed, and how potential regulatory consequences are modeled. Normative conclusions are formed on the basis of observable data and structured relationships between them, which ensures internal admissibility of decisions without reliance on external human authorization.

The reasoning process is implemented through orchestration of specialized components combined into a unified control flow. Management of reasoning stages and system state is provided by an application level orchestrator implemented using FastAPI and RedisJSON, while consistency and sequencing of processing are supported through stream coordination based on Kafka. Execution of formalized reasoning procedures and interaction with language models are performed using LangChain and Semantic Kernel.



Language models are applied for processing textual data and extracting semantic structures within the defined reasoning logic. Formal validation of logical relationships and stability of normative conclusions is performed using symbolic reasoning based on SWI Prolog. Semantic organization of facts and relationships is implemented through vector representations and graph models based on SentenceTransformers and Neo4j, while analysis of regulatory consequences and behavioral scenarios is carried out through mathematical modeling using NumPy and OR Tools.

This approach enables the formation of normative decisions through a formally defined and reproducible reasoning process in which admissibility is determined by the internal logic of the system and by observable consequences of decision application. This provides a foundation for transparent audit, reproducible testing, and adaptation of the system across different legal and regulatory environments.

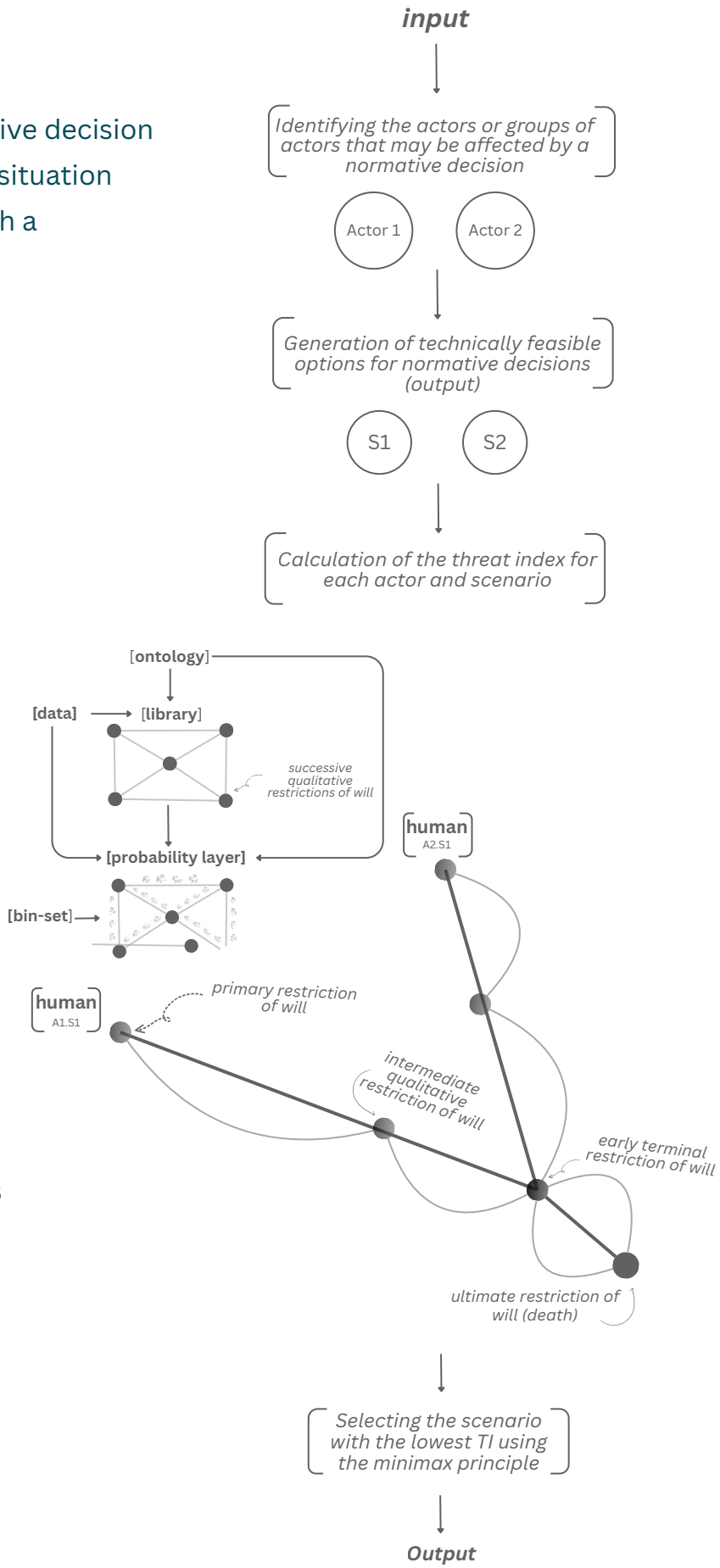


[law as computation]

JudgeAI implements a universal normative decision pipeline that transforms any real-world situation into a single admissible outcome through a structured and reproducible process.

The system begins by identifying all actors whose conditions or future ability to act may be affected. It then generates the set of feasible decision scenarios and models, for each scenario, the causal consequences for every actor. These consequences are evaluated in terms of how they alter the actor's viable action space, capturing both immediate effects and downstream transitions across progressively more severe constraints on agency.

Each scenario is assigned an internal threat profile that reflects the degree to which it concentrates risk on any participant. JudgeAI then applies a minimax selection principle and chooses the scenario that minimizes the highest level of threat across all affected actors. This ensures that the selected outcome preserves system stability and avoids trajectories that lead to critical or irreversible degradation of an actor's capacity to function.



The same pipeline applies uniformly across adjudication, regulatory design, and autonomous agent decision-making, since each domain reduces to the same underlying task: identifying affected actors, modeling causal consequences of alternative actions, and selecting the outcome that maintains admissibility under a unified risk-based criterion.

The internal threat evaluation, causal modeling structure, and parameter calibration are implemented as part of JudgeAI's proprietary architecture, while the decision logic remains transparent, auditable, and reproducible at the level of process.

Detailed research underlying the JudgeAI project is presented in the following publications:

- Kozlov, Yuri and Bajwa, Taaha and Shutova, Maria, Formal Conditions for Autonomous Normative Rule-Making (January 08, 2026). Available at SSRN: https://ssrn.com/abstract=6043174
- Kozlov, Yuri and Shutova, Maria and Bajwa, Taaha, Automated Judge is Not a Task For LegalTech But For DeepTech (February 24, 2025). Available at SSRN: https://ssrn.com/abstract=5151862

[testing and validation]

Testing of JudgeAI was conducted to assess the system's ability to form normative decisions under conditions comparable to judicial and lawmaking practice. The evaluation focused on the reproducibility and admissibility of normative conclusions generated through the processing of empirical data.

Within the lawmaking function, the system generated regulatory solutions based on specified socio economic conditions. The resulting conclusions were compared with existing legal regulation across several jurisdictions using parameters commonly applied in Regulatory Impact Assessment In these comparisons, generated norms demonstrated a consistent advantage, with an average increase of approximately 18–20 points out of 35, corresponding to roughly 50–57% higher RIA scores.


JudgeAI

JUDGEAI BILL

Migration Stabilization and Local Impact Adjustment Act of 2026

Generated Draft

Report | Legislative Text

Appendix A: Methodological Basis for Regulatory Architecture Selection

February 2026

The link to the testing results of the system's lawmaking functionality

Lawmaking testing

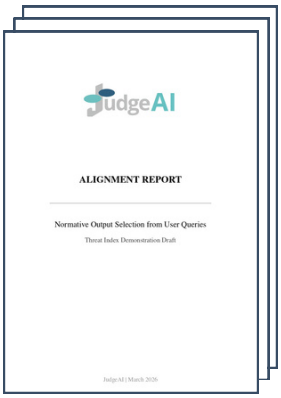
Within the judicial function, the system was applied to model and real dispute scenarios to generate normative conclusions. The logic of fact establishment, causal structuring, and final decisions was compared with judicial practice in analogous situations.

***JudgeAI** demonstrated **96%** alignment with real judicial decisions, analyzed across the following categories of disputes:*

- *International disputes*
- *Investment disputes*
- *Commercial disputes*
- *Intellectual property disputes*
- *Antitrust and competition disputes*
- *Employment disputes*
- *Consumer protection cases*
- *Real estate disputes*
- *Banking and finance disputes*

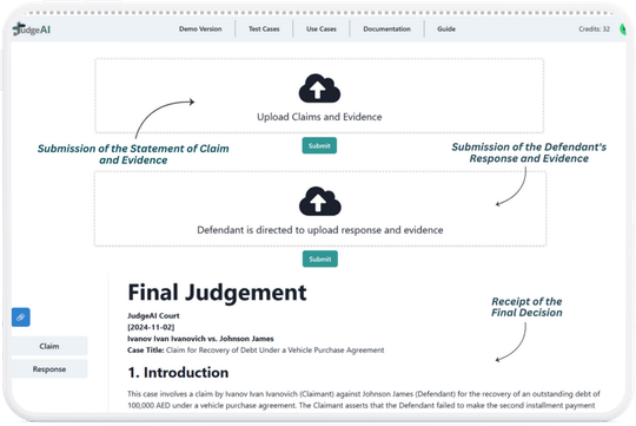
Adjudication testing

In the new JudgeAI Alignment Report, we publish the results of prototype testing and show what such a layer can look like. On one concrete case, the prototype runs a full traceable process: from formulating alternative scenarios to selecting a normatively admissible outcome. With actor identification, causal chains, irreversible harm assessment, and a formalized selection criterion.



Alignment report

The interactive demo of the dispute resolution tool is available on our platform.



interactive demo

[concluding framework]

JudgeAI implements autonomous normative reasoning within a formally defined admissibility logic, in which normative decisions are formed and applied on the basis of internal system criteria. These criteria ensure reproducibility of reasoning and enable formal verification of normative conclusions without reliance on external authority or interpretation.

Further work on the system is oriented toward verifying the stability of admissibility criteria under system scaling and transfer. Verification is conducted by varying conditions of application, configurations of interests, and the structure of factual inputs while preserving formal coherence of the reasoning process.

JudgeAI supports interaction formats aimed at formal and empirical validation of autonomous normative reasoning, including expert review, controlled pilot testing, and deployment within regulatory experimental environments.

Book an introductory session

info@judgeai.space