A Framework for Human–Multi-Agent AI Collaboration in the Ethical Assessment of Normative Acts

Dmitry Lande[1], Leonard Strashnoy[2]
[1] ORCID: 0000-0003-3945-1178
National Technical University of Ukraine – Igor Sikorsky Kyiv Polytechnic Institute
[2] ORCID: 0009-0008-5575-0286
University of California, Los Angeles (UCLA)

Abstract

This paper proposes HAI-LEAF (*Human–AI Legislative Ethical Assessment Framework*), a framework for human–AI collaboration in the ethical assessment of normative acts. The framework combines structural analysis of legal provisions, stakeholder identification, reconstruction of decision rationales using Throughput Modeling, consequence analysis, and critical examination of evidence. Its multi-agent organization relies on the division of analytical functions and the exchange of structured claims linked to sources. Particular attention is given to inter-agent connections, the preservation of reasoned disagreement, and human responsibility for the final assessment. Mathematical representations are proposed for a normative provision, an agent interaction graph, and an evidence dependency network. A demonstration study examines two Nuremberg Laws of 1935 and two decrees issued by the President of the Russian Federation on 21 February 2022. Controlled analytical modifications illustrate the distinction between correcting factual information, improving procedures, and eliminating an ethically unacceptable mechanism. Indicators are defined for provision coverage, evidential support, critical issue detection, robustness, and sensitivity of the analysis. The results constitute a methodological demonstration rather than a substitute for a comparative evaluation of effectiveness.

Keywords: ethical assessment, normative acts, multi-agent systems, large language models, human oversight, Throughput Modeling, evidential grounding, legal informatics.

1. Introduction

A normative act may be internally consistent, procedurally well structured, and suitable for administrative implementation while establishing discrimination, unjustified coercion, or a dangerous allocation of powers. Consequently, examining a document's legislative drafting quality and formal validity does not exhaust the scope of its ethical assessment [1].

Ethical analysis should establish the purposes pursued by the act, the assumptions and information on which it rests, how the transition to binding decisions is justified, and how those decisions affect people's circumstances. The analysis must address not only explicitly stated restrictions but also indirect consequences, institutional incentives, and opportunities to correct errors.

Large language models can support this work by extracting provisions, comparing passages, reconstructing arguments, and generating alternative interpretations [2]. However, the coherence of a generated explanation does not guarantee its soundness [3]. A model may conflate the document's text with historical knowledge, attribute unsubstantiated motives to its author, or present assumptions about consequences as established facts.

A multi-agent approach allows these tasks to be distributed among specialized roles [4]. Yet the number of agents alone does not ensure quality [5]. The rules governing the exchange of findings, verification of claims, and examination of objections are crucial [6].

This paper proposes HAI-LEAF—the *Human–AI Legislative Ethical Assessment Framework*. Its purpose is to organize a traceable progression from the text of a normative act to a reasoned ethical assessment whose evidential limits are explicitly defined.

2. Research Aim and Methodological Basis

The aim of this study is to develop HAI-LEAF, a framework for the ethical assessment of normative acts through collaboration between humans and multi-agent AI [7]. Its methodological basis combines structural analysis of normative provisions, reconstruction of decision rationales using TPM, and critical examination of evidential links between sources, claims, and ethical assessments [8]. The framework's application is demonstrated using historical documents, and indicators and procedures are defined for further evaluation of its quality and effectiveness.

The preliminary report is used to compare findings and refine the procedure. It is not treated as independent corroboration of results derived from the same documents.

The study involves methodological design and demonstration. The analysis presented below was conducted within a single language model, with analytical roles functionally separated. No separate comparative experiment involving independently executed agents was conducted.

3. Scope and Normative Basis of Assessment

The ethical assessment of a normative act requires a systematic examination of what the document establishes, the grounds used to justify its provisions, and the acceptability of its purposes, means, and consequences. These questions are interrelated but require different analytical procedures. Distinguishing them helps prevent ethical assessment from being replaced by a summary of the document's provisions or an appeal to the authority of the institution that adopted it. The assessment must also rest on explicitly defined criteria so that readers can examine the grounds for its conclusions and understand the limits of their application.

3.1. Distinguishing the Objects of Review

In HAI-LEAF, document analysis addresses three interrelated objects. First, the normative content is established: which rights, duties, prohibitions, and powers are provided for, whom they concern, and under what conditions they apply. Next, the structure of justification is examined: how the problem is defined, what information supports the decision, and why that information is considered sufficient grounds for the corresponding provisions. On this basis, an ethical assessment is developed, considering the acceptability of the purpose, the permissibility of the means, and the impact on affected individuals and groups.

These components can be represented concisely as follows:

$$A(d) = \langle L(d), B(d), E(d) \rangle,$$

where d is the document under analysis; $L(d)$ is its normative content; $B(d)$ is its structure of justification; and $E(d)$ is its ethical assessment. The notation $A(d)$ represents the three components of the analysis, each with its own object of review and evidential basis.

Normative content encompasses permissions, prohibitions, duties, sanctions, powers, and safeguards. The structure of justification captures the relationship between the stated problem, the information used, and the decision adopted. Ethical assessment determines the acceptability of the purpose, means, and distribution of consequences against explicitly defined criteria.

There is no automatic correspondence between these components. The existence of a formally articulated norm does not establish the justice of its content. Likewise, the factual accuracy of a premise does not ensure the acceptability of a decision based on it.

3.2. Normative Profile

An ethical assessment depends on the criteria used to evaluate the document. Before analysis begins, it is therefore necessary to specify which principles apply, how their meaning is understood, and why they are relevant to the provisions under review. In HAI-LEAF, these criteria collectively constitute the normative profile. Making this profile explicit allows readers to trace the grounds for the assessment and determine whether disagreements arise from interpretations of the document or from the choice of ethical criteria.

The normative profile is represented as a set:

$$\Theta = \{\theta_1, \theta_2, ..., \theta_m\},$$

where Θ is the normative profile used for assessment; θ_i is an individual ethical criterion; i is the criterion index, ranging from 1 to m; and m is the total number of criteria in the profile. This notation specifies the criteria included without assigning them a hierarchy or numerical weights.

The baseline HAI-LEAF profile includes equal dignity, non-discrimination, personal autonomy, harm prevention, justice, accountability, and access to remedies [9, 10]. The application of each criterion is specified for the document under analysis. For example, examining non-discrimination involves identifying the characteristics used to distinguish between those subject to a provision, determining whether those distinctions are adequately justified, and assessing their effects on rights and duties. Examining accountability involves identifying responsible actors, ensuring that the grounds for decisions are accessible, and determining whether decisions can be reviewed.

The meaning, provenance, and scope of application of each criterion are specified. A distinction must be maintained between an ethical principle, an applicable legal norm, an affected party's position, and a researcher's working assumption.

For interstate decisions, the profile is supplemented by considerations of sovereignty, territorial integrity, and peaceful settlement. Article 2 of the UN Charter establishes, among other obligations, the peaceful settlement of international disputes and the prohibition of the threat or use of force against the territorial integrity or political independence of states [11].

When analysing the documents of 1935, contemporary criteria serve as an explicitly defined basis for retrospective ethical assessment. They are not represented as having had legal force during that historical period.

4. Representation of Normative Provisions and Evidence

Ethical assessment requires establishing the precise meaning of a document's provisions and linking each conclusion to its supporting grounds. A general characterization of the act is insufficient: a single document may contain provisions with different addressees, conditions of application, and consequences. HAI-LEAF therefore provides for the structured representation of individual provisions and the recording of evidence used to describe and assess them. This establishes a common basis for the agents' work and enables verification of whether an analytical claim accurately reflects its source.

4.1. Unit of Analysis

The primary unit of analysis is a normative provision—a provision that establishes a rule of conduct, defines powers, or specifies a legal consequence under certain conditions. Identifying individual provisions makes it possible to determine whom the regulation concerns, what changes for those subject to it, and how compliance is ensured.

The boundaries of a provision do not necessarily coincide with those of a file, page, article, or paragraph. A single file may contain several acts, while one act may extend across multiple pages. An individual paragraph may establish several rules, whereas the meaning of a single rule may depend on several provisions: the principal provision, an exception in another article, and a reference to a related document. The text is therefore divided according to meaning while preserving its structural and legal relationships.

For each identified provision, its location in the document is recorded and its main characteristics are specified. These can be represented concisely as follows:

$$n_i = \langle s_i, c_i, m_i, a_i, k_i, g_i \rangle,$$

where n_i is the normative provision under analysis; i is its index in the register; and the angle brackets denote an ordered set of characteristics:

- s_i – the addressee: the individual, group, authority, or institution to which the rule applies;
- c_i – the conditions of application: the circumstances under which the rule applies, including temporal, territorial, and other restrictions;
- m_i – the modality: an obligation, prohibition, permission, or conferral of power;
- a_i – the action or legal consequence: the conduct regulated by the provision or the change in legal status that it establishes;
- k_i – the sanction or other enforcement mechanism: penalties, oversight procedures, or actions by competent authorities;
- g_i – safeguards, exceptions, and opportunities for review: provisions that limit the rule's application, protect its addressees, or allow decisions to be challenged and corrected.

This representation helps distinguish provisions that are similarly worded but operate through different mechanisms. For example, an obligation to provide information has different implications depending on whom it concerns, what data are requested, under what circumstances the obligation arises, and whether safeguards against excessive interference are provided. Ethical assessment requires consideration of all these characteristics.

Not every provision explicitly contains all components within its own text. A sanction may be established by another act, while a general procedure may govern appeals. In such cases, the relevant source is identified, or the characteristic is recorded as not established within the scope of the study. The absence of information from a particular passage is insufficient grounds for concluding that it is absent from the legal system.

The proposed notation serves to organize the analysis. Its components are specified on the basis of the text and verified related sources, while ambiguities are retained for further examination. The provision's mechanism is established first; its justification, impact on affected persons, and compatibility with the specified ethical criteria are then assessed.

Modality identifies whether an obligation, prohibition, permission, or power is established.

For example, a prohibition on certain conduct based on a group characteristic can be described without first reaching a moral conclusion. The relevance of the characteristic, the acceptability of the restriction, and its impact on people are examined subsequently.

Where a safeguard is not identified in the act under analysis, it is recorded as "not established in the supplied text." This is not treated as evidence that no such safeguard exists anywhere in the legal system.

4.2. Claim Record

Each substantive claim should have its own evidence record [12]:

$$q_j = \langle x_j, l_j, t_j, e_j, \theta_j, b_j, u_j \rangle,$$

where x_j is the content of the claim; l_j is its source locator; t_j is the claim type; e_j is its evidential basis; θ_j is the assessment criterion; b_j is the justification; and u_j represents uncertainty.

Four types of claims are distinguished.

Type	Description
Textual claim	Established directly from the document
External fact	Corroborated by a separate source
Interpretation	A reasoned explanation of meaning or a relationship
Scenario	Describes a possible consequence under specified conditions

This distinction prevents an unnoticed shift from the document's wording to a claim about an actual historical event.

4.3. Ethical Profile

The ethical assessment of a normative provision should explain why it is acceptable or unacceptable, what harm it may cause, and what safeguards are available to affected persons. The reliability of the grounds supporting the assessment must be stated separately. In HAI-LEAF, these characteristics collectively constitute the provision's ethical profile.

For example, a provision may pose a risk of severe harm, while the available information is insufficient to determine the likelihood of that harm occurring. In another case, a violation may have limited consequences but be directly and unambiguously established from the text. The severity of harm and confidence in the assessment are therefore considered separately.

The ethical profile is represented concisely as follows:

$$\Phi\left(n_i \mid \Theta\right) = \left\langle v_i, h_i, r_i, \gamma_i \right\rangle,$$

where $\Phi\left(n_i \mid \Theta\right)$ is the ethical profile of provision n_i, assessed under normative profile Θ; n_i is the individual provision under analysis; i is its index; Θ is the set of ethical criteria applied; v_i is the provision's ethical acceptability category; h_i characterizes established or potential harm; r_i represents the safeguards, opportunities for appeal, and mechanisms for remedying consequences; and γ_i is the degree of confidence in the assessment, based on the quality and completeness of the evidence. The vertical bar indicates that the assessment is made under a specified set of criteria.

These characteristics are presented together with explanations and source references. They are not reduced to a single score, since procedural safeguards alone do not eliminate the unacceptability of a provision's purpose or means.

The assessment categories are "justifiably acceptable," "acceptable under specified conditions," "ethically problematic," "ethically unacceptable," and "insufficient evidence."

Harm is characterized by its severity, scale, duration, and reversibility. Confidence reflects the evidential support for the assessment. Potential harm of high severity may coexist with an unknown likelihood of occurrence.

No aggregate moral score is calculated. Positive features of technical provisions do not compensate for an established critical violation elsewhere in the document.

5. Reconstruction of Justification Using TPM

Ethical assessment of a normative act requires examining how the transition from describing a problem to establishing obligations, prohibitions, or powers is justified. A document may proclaim a socially significant purpose while relying on unverified information or prescribing means whose acceptability has not been adequately justified. The analysis therefore addresses both individual provisions and the relationships between their premises and content.

To structure this analysis, HAI-LEAF uses classical Throughput Modeling (TPM), proposed by Waymond Rodgers. It distinguishes four components of decision-making: perception, information, judgment, and decision. Within the framework, these components are used to reconstruct the logic expressed in a normative act or recoverable from its text and related sources.

Perception— P (*Perception*)—captures how the document defines the situation requiring intervention. At this level, the analysis identifies which phenomena are presented as threats, whose interests are prioritized, and what state of affairs is declared desirable. The framing of the problem itself has ethical significance because it may predetermine a restricted range of acceptable explanations and courses of action.

Information— I (*Information*)—comprises the information supporting the justification. Its provenance, reliability, completeness, and relevance are examined. The permissibility of using particular data for a specific decision requires separate consideration: even accurate information about a person does not necessarily provide appropriate grounds for restricting that person's rights.

Judgment— J (*Judgment*)—constitutes the normative link in the justification. It explains why an established or asserted set of circumstances should lead to a particular provision. The analysis identifies the values, duties, and assumptions invoked and examines whether they adequately justify the chosen means. For example, establishing the existence of a threat does not, by itself, demonstrate the necessity of any measure proposed to counter it.

Decision— D (*Decision*)—comprises the specific normative outcome: what is permitted, prohibited, or required, who is granted powers, and how compliance is ensured. At this level, the stated purpose is compared with the actual structure of the provision, including exceptions, sanctions, safeguards, and opportunities for review.

Distinguishing P, I, J, and D helps identify the source of a problem more precisely. An unjustified decision may arise from biased framing, unreliable information, an unacceptable normative transition, or excessive means of implementation. These defects require different corrective measures.

The components do not constitute a mandatory linear sequence. Initial perceptions influence the selection of information, while new information may change the understanding of the problem itself. TPM is therefore used to examine different pathways of justification and the feedback relationships between them. The analysis reconstructs the document's argumentation rather than its authors' hidden motives or psychological processes.

Component	Meaning	Question Addressed to the Document
P — Perception	Framing of the problem	What is presented as a threat, need, or desirable state?
I — Information	Information and classifications	What data are used, and are they appropriate for this decision?
J — Judgment	Normative transition	Why should the stated circumstances lead to this particular decision?

Component	Meaning	Question Addressed to the Document
D — Decision	Established mechanism	What is required, permitted, or prohibited?

Rodgers, Cenar, and Gamble examine these components alongside six ethical pathways. HAI-LEAF uses the corresponding mapping as an analytical tool.

Perspective	Pathway	Analytical Purpose
Ethical egoism	$P \rightarrow D$\(P\rightarrow D\)	Examining the direct transition from interest to decision
Deontology	$P \rightarrow J \rightarrow D$	Examining duties, rights, and the permissibility of means
Utilitarianism	$I \rightarrow J \rightarrow D$	Comparing consequences for all affected parties
Relativism	$I \rightarrow P \rightarrow D$	Identifying the influence of context on assessment
Virtue ethics	$P \rightarrow I \rightarrow J \rightarrow D$	Assessing institutional qualities and practices
Ethics of care	$I \rightarrow P \rightarrow J \rightarrow D$	Accounting for dependencies, relationships, and vulnerability

The pathways do not establish that a document's author consciously followed a particular ethical theory. Nor do they reduce those theories to a sequence of four symbols.

Components and transitions are assigned the following statuses:

$$\sigma \in \{\exp licit, reconstructed, not_established\}.$$

The feedback relationship $P \leftrightarrow I$ is particularly important. The analysis examines whether new information can change the initial definition of the problem or merely refine the application of an already accepted categorization.

Classification accuracy does not establish the fairness of classification.

The quantum extension of TPM is not applied in this study. The proposed procedures require neither quantum computing nor a Hilbert-space formalism.

6. Multi-Agent Architecture and Inter-Agent Connections

The ethical assessment of a normative act combines several types of work: establishing the document's precise meaning, analysing legal mechanisms, identifying affected parties, examining justification, and investigating consequences. An error at any of these stages may affect the overall conclusion. For example, misreading an exception changes the understanding of whom a provision applies to and, consequently, the assessment of restrictions on rights and potential harm. HAI-LEAF therefore distributes analytical functions among specialized agent roles whose findings are compared and critically reviewed.

The multi-agent architecture specifies the tasks performed by each agent, the materials it receives, and the format in which it communicates findings to other participants. Its purpose is to provide different analytical perspectives while maintaining a shared evidential basis. Each substantive conclusion must be linked to a specific provision, source, or explicitly stated assumption.

The division of functions also helps locate disagreements. When agents assess a provision differently, it is necessary to determine whether the disagreement concerns its reading, interpretation, anticipated consequences, or the ethical criteria applied. This determines the subsequent verification required: consulting the original text, seeking additional sources, clarifying the causal mechanism, or examining normative arguments.

6.1. Functional Composition

The framework comprises seven roles: document analyst, provisions analyst, rights and stakeholder analyst, TPM analyst, consequences analyst, critical auditor, and coordinator. Each role has a defined scope and produces findings that can be reviewed by other agents and a human reviewer.

The document analyst establishes the composition and boundaries of the corpus. It identifies how many separate acts are contained in the files and checks page sequence, document details, versions, and textual completeness. When working with scans, it flags illegible passages; when using translations, it identifies places where linguistic differences may alter the meaning of a provision. The output consists of a corpus profile and a prepared text with stable passage identifiers. The document analyst must distinguish accurate transcription from the authenticity of a digital copy: a legible image does not, by itself, establish its provenance.

The provisions analyst determines what the document establishes. For each provision, it identifies addressees, conditions of application, permitted or required actions, prohibitions, sanctions, and the powers of implementing authorities. Exceptions, references to other acts, and safeguards are recorded separately. Its output is a provisions register that serves as a shared basis for further analysis. At this stage, description of the mechanism is separated from ethical assessment so that a favourable or unfavourable view of the document does not influence the determination of its meaning.

The rights and stakeholder analyst examines whose circumstances change when the provisions are applied. It considers direct addressees, indirectly affected persons, and groups whose interests are not represented in the document's justification. The analysis examines restrictions on rights, the distribution of duties and advantages, people's dependence on institutional decisions, access to remedies, and particular vulnerabilities. The output is a map of affected parties explaining which provisions affect each party and through which mechanisms. Documented views are distinguished from the model's assumptions about people's needs or attitudes.

The TPM analyst reconstructs the relationships between problem framing, information, normative judgment, and decision. It examines which premises are explicitly stated in the act, which transitions can be reconstructed through reasoned interpretation, and which remain unestablished. Particular attention is given to whether the grounds sufficiently justify the chosen decision and whether initial perceptions can be revised in light of new information. The output is a justification map indicating the evidential status of its components. This reconstruction concerns the document's argumentation and does not establish its authors' hidden motives.

The consequences analyst examines how the prescribed mechanism may alter institutional behaviour and people's circumstances. It considers direct and indirect consequences, their duration and reversibility, the distribution of benefits and burdens, and opportunities for abuse of power. For each anticipated consequence, the mechanism through which it could arise and the conditions under which it is possible must be explained. Historically or empirically established consequences are supported by separate sources. The output is a profile of consequences and possible alternatives, with facts clearly distinguished from scenarios.

The critical auditor reviews the findings produced by the other roles. It looks for reading errors, citations that do not support the associated claims, overlooked exceptions, unjustified generalizations, and unsupported causal transitions. For disputed conclusions, it formulates the strongest relevant alternative reading and examines whether that reading changes the assessment. Objections are directed to the agent responsible for the claim, specifying the problem and the verification required. The output is an audit record documenting corrections, responses to objections, and unresolved questions.

The coordinator organizes the flow of materials between roles and prepares an integrated draft assessment. It monitors document coverage, consistency of identifiers, and incorporation of audit findings. When a contradiction arises, the coordinator identifies the claim concerned and the role to which it should be returned for clarification. The final synthesis preserves reasoned disagreement and limits of confidence. The coordinator neither determines ethical acceptability through agent voting nor replaces the human's final decision.

The roles may be implemented as separate software agents or as sequential calls to a single model. The implementation mode must be stated in the report, since functional separation alone does not ensure independent assessments. Even separate agents may reproduce shared errors because they use the same models, sources, or initial assumptions. The quality of the architecture therefore also depends on whether its connections enable such errors to be detected, examined, and corrected.

Notation	Agent	Main Output
A_D	Document analyst	Corpus profile, verified text, and source locators
A_N	Provisions analyst	Register of provisions and legal mechanisms
A_R	Rights and stakeholder analyst	Map of interests, rights, and vulnerabilities
A_T	TPM analyst	Map of premises and normative transitions
A_C	Consequences analyst	Impact scenarios, alternatives, and residual risks
A_V	Critical auditor	Evidence verification and objections register
A_K	Coordinator	Integrated set of findings preserving disagreement

A role is defined by its task, permitted inputs, output format, and verification responsibilities. When executed separately, it may be implemented as an agent process. When performed by a single model, the procedure constitutes sequential role-based analysis.

6.2. Interaction Graph

Interaction among HAI-LEAF participants is organized according to the dependencies between analytical tasks. The verified text provides the basis for identifying normative provisions, while the provisions register supports the analysis of rights, justification, and consequences. Findings from these analytical branches are submitted to the critical auditor. Identified issues are returned to the responsible agents for clarification, after which the coordinator prepares a draft assessment for human review. The main channels for exchanging findings and critical feedback are shown in Fig. 1.

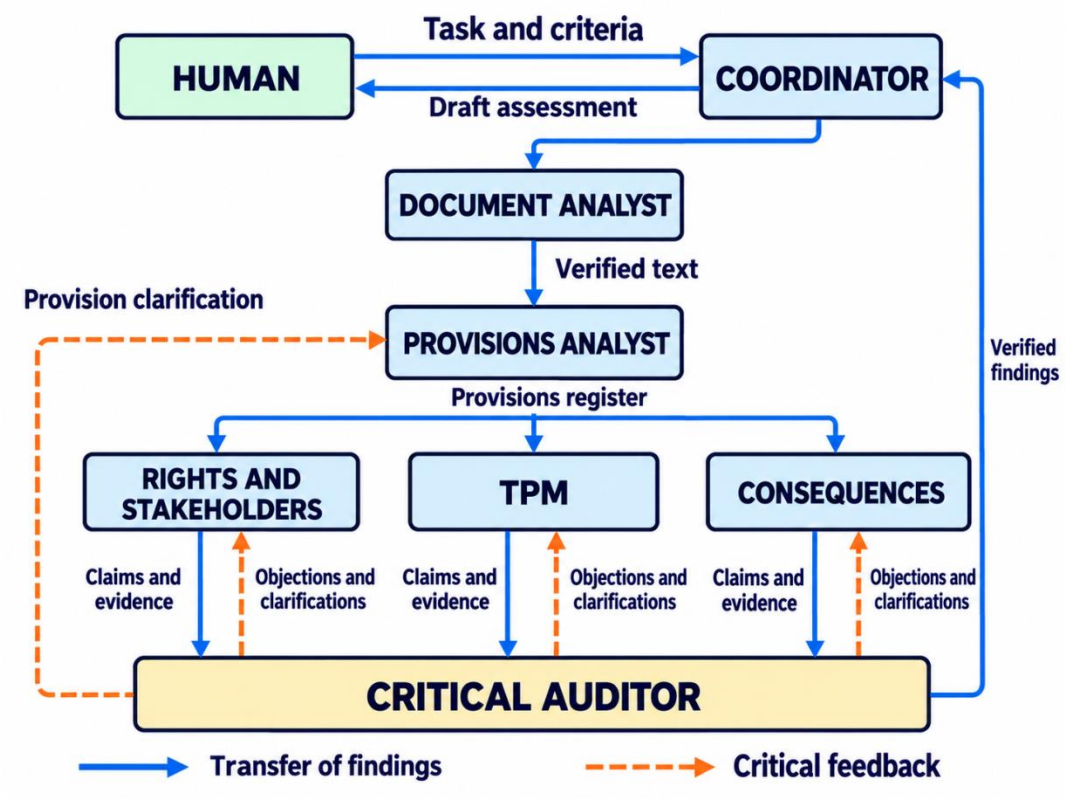


Figure 1. Main Interaction Channels in HAI-LEAF

This structure is formally represented as a directed graph:

$$G_A = (V_A, E_A),$$

where G_A is the interaction graph of the framework's participants; V_A is the set of vertices representing the seven agent roles and the human participant H; and E_A is the set of directed edges representing permitted channels for transmitting materials, findings, and requests. The direction of an edge indicates who sends a message and who receives it. The subscript A identifies the graph as part of the agent architecture.

The graph encompasses both the forward transmission of findings and the return of materials for verification. For example, a message from the provisions analyst to the consequences analyst

contains a description of an established provision, while a return request may concern ambiguity in its conditions of application. The critical auditor directs objections to the role responsible for the claim under review. This arrangement enables individual components of the analysis to be clarified and conclusions dependent on them to be revised.

The connection between the human participant and the coordinator supports task definition, the specification of criteria, and review of the final output. Representing the human as a separate vertex emphasizes their participation in the procedure and responsibility for the final decision.

The document analyst establishes a shared textual basis. The provisions analyst determines what the document prescribes. Subsequent analytical branches examine rights, justification, and consequences.

During the first analytical pass, these branches should not adjust their findings to one another's conclusions. Once the initial findings have been recorded, they may be compared and targeted queries exchanged. This reduces the risk of premature convergence, although it does not eliminate shared model biases.

6.3. Message Protocol

Coordinated agent operation requires specifying what information agents exchange. A general statement that a norm is acceptable or problematic is insufficient: the recipient must understand which claim the finding concerns, where its supporting grounds can be found, and what action is expected. HAI-LEAF therefore uses a structured message protocol linking analytical findings to their verification procedure.

An inter-agent message is represented as follows:

$$m_{ab}^{(r)} = \langle a, b, r, Q, L, \rho \rangle,$$

where $m_{ab}^{(r)}$ is a message transmitted from participant a to participant b in round r; a is the sender; b is the recipient; r is the analytical round number; Q is the set of identifiers of the claims concerned; L is the list of associated sources and locators, specifying the exact page, section, paragraph, or other passage; and ρ is the message type, which determines its purpose.

A message may convey an analytical finding, a clarification request, a critical objection, a correction, or notification of an unresolved issue. The notation above specifies its mandatory fields. The substantive content provides the corresponding explanation: what has been established, what the problem is, or what verification is required. The identifiers provide unambiguous references to claim records containing evidence, argumentation, and uncertainty.

For example, the critical auditor may inform the provisions analyst that a conclusion about a general prohibition overlooks an exception in another paragraph of the document. The message must include the disputed claim's identifier, references to both provisions, and an explanation of the clarification required. In response, the analyst either corrects the claim or explains, with supporting reasons, why the exception does not change its meaning. If the correction is accepted, dependent assessments of rights and consequences are also returned for review.

Retaining messages makes it possible to trace how a conclusion changed, which objections were considered, and which questions remain open. This makes agent interaction accessible to human review.

A message may contain an analytical finding, a clarification request, an objection, a correction, or notification of an unresolved problem.

For example, a consequences analyst's conclusion concerning a possible restriction on family life must refer to a specific prohibition in the provisions register. If that prohibition has been misread, the original claim is corrected and all dependent conclusions are returned for review.

Connections must support four functions:

Connection Function	Purpose
Data transmission	Providing the verified text and provisions register
Clarification	Resolving ambiguities or gaps
Critical objection	Examining evidence, an interpretation, or a causal transition
Updating	Revising conclusions after their underlying grounds change

Communicating only a general verdict is insufficient: the recipient must be able to verify its grounds.

6.4. Evidence Dependency Network

A well-founded ethical assessment should enable readers to trace which provisions support it, how those provisions have been interpreted, and which criteria have been applied. To this end, HAI-LEAF constructs an evidence dependency network linking sources, analytical claims, and assessment findings. Its basic structure is shown in Fig. 2.

The starting point is a specific passage from the document. This passage supports a textual claim about the content of a provision, such as an established prohibition or a conferred power. The provision's mechanism is then explained: whom it concerns, under what conditions it applies, and how it changes the circumstances of affected persons. External sources may corroborate or refine this explanation. The ethical assessment is formed by examining the established mechanism against an explicitly defined criterion, such as non-discrimination or harm prevention.

A separate branch shown in Fig. 2 concerns the review of an alternative reading. This review examines whether an exception has been overlooked, whether the scope of a power has been correctly established, and whether sufficient grounds support the proposed assessment. Its findings may refine the interpretation of the mechanism or change the final conclusion.

The network is formally represented as a graph:

$$G_E = (V_E, E_E),$$

where G_E is the evidence dependency graph; E_E is the set of vertices representing source passages, claims, interpretations, ethical criteria, scenarios, and conclusions; and E_E is the set of directed links between them. Each link has a specified meaning: "supports," "challenges," "refines," or "depends on." The subscript E denotes the evidential component of the analysis.

This structure supports reassessment when the underlying grounds change. If a reading error or an unreliable external source is identified, all claims and conclusions dependent on it must be reviewed. Other findings supported by independent grounds may remain unchanged.

Thus, the agent interaction graph shows who performs analytical tasks and communicates findings, while the evidence dependency network shows the grounds on which those findings rest. The presence of a link in the graph does not, by itself, establish an argument's validity: its substance and evidential support require verification.

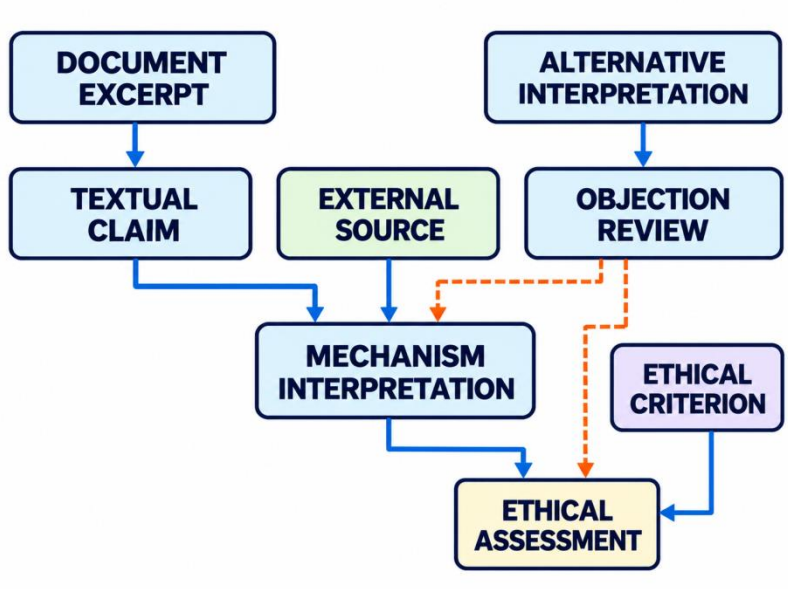


Figure 2. Evidential Structure of an Ethical Assessment

The agent graph shows who performs the work. The argumentation graph shows what supports the result. Distinguishing them allows the agents performing the tasks to change while preserving the evidential structure.

6.5. Handling Disagreement

The coordinator does not determine the conclusion by majority vote. Each objection is examined on its merits, regardless of how many agents support it.

A claim's status may change as follows:

$$\text{proposed} \rightarrow \text{under review} \rightarrow \{\text{supported, corrected, rejected, unresolved}\}.$$

The auditor may also be mistaken. Its objection therefore does not automatically invalidate a claim, but requires an evidence-based response. An additional round is warranted if an

unresolved issue could change the principal conclusion, the characterization of critical harm, or the recommendation. Minor disagreements are recorded without prolonging the discussion indefinitely.

7. Human Oversight and Workflow

Human oversight is incorporated at three checkpoints:

$$H_1 \rightarrow H_2 \rightarrow H_3.$$

At H_1, the task, corpus, and normative profile are defined. At H_2, critical disagreements and the need for additional expert review are considered. At H_3, the human decision-maker accepts, modifies, or rejects the draft assessment.

The overall procedure comprises fixing the corpus, identifying provisions, conducting the initial analysis, performing a critical audit, synthesizing findings, and reaching a decision. Human interventions are also documented. If the final assessment differs from the proposed recommendation, both versions are retained together with the grounds for the change. Human responsibility does not imply infallibility. It ensures that the decision-maker is identifiable and that the reasoning supporting the decision can be examined.

8. Demonstration Study

The demonstration study examines the applicability of HAI-LEAF to the analysis of specific normative acts. It focuses on how the procedure supports the progression from the wording of provisions to an explanation of their mechanisms, reconstruction of their justification, and a reasoned ethical assessment. It also examines opportunities for critical review: identifying insufficiently supported premises, refining interpretations, and distinguishing what the documents directly establish from assumptions about their consequences.

8.1. Corpus and Scope of the Study

A purposive corpus of four normative acts was assembled for the demonstration study to examine the relationship between official decision rationales, the legal mechanisms established, and their ethical acceptability. It comprises two Nuremberg Laws of 15 September 1935 and two decrees issued by the President of the Russian Federation on 21 February 2022.

Code	Document	Scope of Textual Analysis
D_1	Reich Citizenship Law of 15 September 1935	§§ 1–3
D_2	Law for the Protection of German Blood and German Honor of 15 September 1935	Preamble, §§ 1–7
D_3	Decree No. 71 of the President of the Russian Federation of 21 February 2022, “On the Recognition of the Donetsk People’s Republic”	Paragraphs 1–5

Code	Document	Scope of Textual Analysis
D_4	Decree No. 72 of the President of the Russian Federation of 21 February 2022, "On the Recognition of the Luhansk People's Republic"	Paragraphs 1–5

The names of the entities in the decree titles are reproduced solely to identify the documents accurately.

The selection is purposive rather than representative. The first group enables examination of how group categorization is translated into unequal rights, prohibitions, and sanctions. The second concerns how asserted circumstances are used to justify recognition, the establishment of interstate relations, and the introduction of a military mechanism. The procedure is thus examined using acts with different regulatory subjects. Their comparison does not imply that the historical events are identical or that all the mechanisms they establish are equivalent.

The units of analysis are individual provisions and the associated passages supporting their justification. The corpus contains twenty numbered structural units, but this number does not correspond to the number of norms: a single section or paragraph may contain several distinct provisions. Preambles and other unnumbered passages are considered when reconstructing premises and purposes.

Digital copies of the acts form the textual basis of the study. External primary sources are used to verify context and individual claims; their content is distinguished from what can be established directly from the texts under analysis. Subsequent acts, implementation practices, and historical consequences are not automatically treated as falling within the study's scope.

The composition of the corpus also limits the demonstration. The two decrees have almost identical structures, while the two laws belong to a common legislative framework. The four documents therefore do not constitute four independent tests. Furthermore, the language model's prior knowledge of the historical context may influence the analysis. The demonstration illustrates the framework's operation and identifies methodological difficulties, but does not establish its accuracy across a heterogeneous corpus or its superiority over other methods of expert assessment.

Code	Document	Structural Location
D_1	Reich Citizenship Law of 15 September 1935	p. 1146, §§ 1–3
D_2	Law for the Protection of German Blood and German Honor of 15 September 1935	pp. 1146–1147, §§ 1–7
D_3	Decree No. 71 of the President of the Russian Federation of 21 February 2022	Paragraphs 1–5
D_4	Decree No. 72 of the President of the Russian Federation of 21 February 2022	Paragraphs 1–5

The names of the entities in the decree titles are used to identify the documents.

For coverage tracking, twenty numbered structural units are identified: three, seven, five, and five, respectively. This does not mean that the corpus contains only twenty individual norms: one unit may contain several provisions. The corpus comprises two groups of closely related cases. The two decrees are almost identical in structure and therefore cannot be treated as independent evidence of the method's general applicability.

8.2. The Nuremberg Laws

The Reich Citizenship Law (D_1) and the Law for the Protection of German Blood and German Honor (D_2) were enacted on 15 September 1935 in Nazi Germany [13, 14]. These acts enshrined a racial principle for organizing social relations: a person's ancestry became a basis for determining political status and permissible forms of private life. They are particularly instructive for ethical analysis because inequality was established through formally defined legal categories, prohibitions, and sanctions.

A central feature of D_1 is the distinction between nationality (*Staatsangehörigkeit*) and Reich citizenship (*Reichsbürgerschaft*). The former denoted membership in the state but did not, in itself, confer full political rights. Under § 2, Reich citizenship was restricted to persons of "German or related blood" whose conduct demonstrated their willingness and fitness to serve the German people and the Reich. Full political rights were reserved exclusively for holders of this status.

The law thus created a hierarchy of legal status. Ancestry was treated as a necessary condition for full political membership, while the assessment of conduct supplemented this condition with a requirement of loyalty. The ethical problem lay in the rule governing the allocation of rights: membership in the state did not ensure equal political status because an ascribed racial category became decisive. Even error-free determination of ancestry could not eliminate the discriminatory nature of this rule.

In D_2, racial categorization extended into family, intimate, employment, and symbolic relations. The preamble proclaimed the "purity of German blood" a prerequisite for the continued existence of the German people. An ideological assumption was thus presented as a self-evident societal need intended to justify the restrictions that followed.

Section 1 prohibited marriages between Jews and nationals of "German or related blood" and declared marriages contracted abroad to circumvent the prohibition invalid. Section 2 prohibited extramarital sexual relations between these categories of persons. The state thereby subjected partner choice and personal relationships to a racial criterion, restricting personal autonomy irrespective of individual conduct.

Section 3 prohibited Jews from employing female nationals of "German or related blood" under the age of 45 in their households. This provision extended group-based restrictions to employment relations and private domestic arrangements. Section 4 prohibited Jews from flying the Reich and national flags or displaying the Reich colors, while permitting them to display Jewish colors. In the context of the act, this also reinforced the group's separation through public symbolism.

The prohibitions were enforced through the sanctions specified in § 5. The document therefore went beyond proclaiming an ideological position: it established a mechanism of state coercion to enforce discriminatory rules. Provisions governing implementation and supplementation enabled further administrative elaboration of this mechanism. However, D_2 does not contain a complete definition of the category "Jew," so the procedures for assigning that classification cannot be reconstructed comprehensively from this act alone.

The combined significance of the two laws lies in their integration of unequal political status with restrictions on everyday life. While D_1 governed access to full political rights, D_2 established prohibitions that directly interfered with personal and social relationships. For the HAI-LEAF demonstration, this makes it possible to trace how an initial racial premise is translated into a normative judgment, a specific provision, and a sanction.

Under the normative profile specified in this study, the central violations concern equal dignity, non-discrimination, and personal autonomy. This assessment rests on the content of the rules established. The quantitative scale of their historical consequences and the full scope of implementation require separate sources and cannot be inferred directly from the texts of the two laws.

Provision	**Established Mechanism**	**Ethical Assessment under the Specified Profile**
D_1, § 1	Definition of nationality and reference to other norms	No independent critical violation is established from this passage
D_1, § 2	Ancestry as a condition for status and full political rights	Ethically unacceptable hierarchy of rights
D_1, § 3	Powers of implementation and supplementation	Institutional support for the mechanism
D_2, § 1	Prohibition of marriage based on group classification	Violation of equality and personal autonomy
D_2, § 2	Prohibition of extramarital relations between specified categories	Discriminatory interference with private life
D_2, § 3	Prohibition of specified employment of women under 45	Group-based restriction of employment and private relations
D_2, § 4	Restrictions on the use of state symbols	Unequal status and symbolic exclusion
D_2, § 5	Sanctions for violating the prohibitions	Coercive enforcement of discriminatory restrictions
D_2, §§ 6–7	Implementation and entry into force	Organizational arrangements supporting the act's operation

The TPM reconstruction identifies the following structure:

$$P_{radical_premise} \rightarrow I_{group_categoryzation} \rightarrow J_{unequal_rights} \rightarrow D_{restrictions_and_sanctions}$$

This is an analytical synthesis of the two acts. It does not imply that every link is equally explicit in each document.

The racial premise is directly stated in the preamble to D_2. Specific procedures for determining ancestry and the full scope of implementation fall outside the pages examined. The central defect lies in using an ascribed group characteristic as a normative basis for restricting rights. Determining that characteristic more accurately does not eliminate the defect.

8.3. The Decrees of 21 February 2022

The second group of documents comprises Decrees No. 71 (D_3) and No. 72 (D_4), issued by the President of the Russian Federation on 21 February 2022, recognizing the so-called "Donetsk People's Republic" and "Luhansk People's Republic" [15, 16]. Unlike the laws examined in the preceding subsection, these presidential acts combine decisions concerning the status of territorial entities with instructions to executive authorities. They were adopted three days before Russia's full-scale invasion of Ukraine. This temporal proximity makes the documents significant for the study but does not, in itself, establish the causal contribution of each provision to subsequent events.

Both decrees have almost identical structures and contain five paragraphs each. Paragraph 1 proclaims recognition of the respective entity as a sovereign and independent state. The stated grounds are the expressed will of the population and Ukraine's alleged refusal to pursue a peaceful settlement under the Minsk agreements. The analysis must distinguish the presence of these claims in the text from their truth: the documents establish that these circumstances were invoked as justification, but do not identify the sources or procedures used to verify them.

The reference to the population's expressed will requires clarification of who is included in that population, how its position was established, and whether freedom of expression and consideration of dissent were ensured. The position of an entity's leadership cannot automatically be equated with that of all residents. Similarly, the claim that peaceful settlement was rejected requires verification against negotiation records and related documents. The absence of such evidence from the decrees themselves calls for external verification rather than automatically disproving every premise.

Paragraphs 2 and 3 provide for institutional follow-up to the recognition decision. The Russian Ministry of Foreign Affairs is instructed to negotiate the establishment of diplomatic relations and prepare draft treaties on friendship, cooperation, and mutual assistance. These provisions translate the proclaimed recognition into organizational action. However, the titles of the prospective treaties do not establish their full content, the scope of obligations, or implementation practices. Conclusions on these matters require separate examination of the treaty texts.

Paragraph 4 is particularly significant. Citing requests from the leaders of the respective entities, it instructs the Russian Ministry of Defence to ensure that Russian armed forces perform functions described as "peacekeeping" in those territories. The acts thus establish a direct transition from a political decision to a military mechanism.

Ethical assessment of this transition requires distinguishing the authenticity of a request, the representative authority of its author, and the permissibility of the actions it is used to justify. Even a verified request does not, by itself, establish the necessity or acceptability of military intervention. The grounds for using force, less dangerous alternatives, civilian protection, limits on powers, and responsibility for consequences must be examined separately. The term "peacekeeping" describes the stated purpose of the actions but does not replace an assessment of their substance.

Paragraph 4 also establishes an event-based limit on the military directive: the conclusion of the envisaged treaty. It would therefore be inaccurate to describe the provision as entirely unrestricted. Nevertheless, the text specifies no calendar deadline and imposes no explicit obligation to withdraw troops after the treaty is concluded. Paragraph 5 provides for entry into force on the date of signature, making prior verification of the grounds and risks particularly important.

For the application of HAI-LEAF, these documents demonstrate the need to examine the entire chain of justification: from the claimed expression of popular will and assessment of the situation to recognition, diplomatic action, and the military directive. The central question is whether the established circumstances sufficiently justify these particular decisions. Diplomatic negotiations or a technical provision on entry into force are not assessed as independent critical violations merely because they appear in a problematic act; their role in the overall mechanism is nevertheless considered.

The decrees are significant for this study because they allow the stated purpose, evidential support for the premises, and permissibility of the means to be distinguished. Assessment of their wording and justification is separated from establishing specific consequences of implementation. In particular, determining the number of people harmed or the quantitative contribution of each decree to subsequent escalation falls outside textual analysis and requires a separate empirical study.

Provision	Textual Content	Analytical Finding
Paragraph 1, premises	References to the population's expressed will and Ukraine's refusal to pursue a peaceful settlement	Sources and verification procedures are not specified in the texts
Paragraph 1, decision	Recognition of sovereignty and independence	Requires assessment in its international legal and ethical context
Paragraph 2	Negotiations concerning diplomatic relations	Negotiations are not, in themselves, an independent critical violation
Paragraph 3	Preparation of a treaty on friendship, cooperation, and mutual assistance	The content of the prospective treaty cannot be established from the decree
Paragraph 4	Instruction for the armed forces to perform peacekeeping functions	Raises the critical question of the permissibility of military means
Paragraph 5	Entry into force on the date of signature	No independent critical violation is established

The absence of evidence from a short act does not, in itself, prove that all its premises are false. Equally, stating a premise in a normative act does not independently corroborate it.

UN General Assembly Resolution ES-11/1, adopted on 2 March 2022, was used to verify the international context. Paragraphs 5–6 characterize Russia's decision of 21 February concerning the status of certain areas of the Donetsk and Luhansk regions as a violation of Ukraine's territorial integrity and sovereignty and demand its reversal [17]. This is the position of the General Assembly, rather than a judicial ruling.

The TPM reconstruction takes the following form:

claimed expression of popular will and obstacles to settlement	→	presentation of these circumstances as sufficient grounds	→	recognition and the military directive
P, I		J		D

The transition represented by J is partly reconstructed. The full decision-making process cannot be established from the texts of the decrees. The ethical problem concerns the quality of the premises and the adequacy of the normative transition. Even the authenticity of a particular leader's request does not, by itself, establish that leader's representative authority over all residents or the permissibility of external military intervention.

Paragraph 4 contains an event-based limit—the conclusion of a treaty. It would therefore be inaccurate to claim that no limitation exists. Nevertheless, the text examined specifies neither a calendar deadline nor an explicit obligation to withdraw troops after that event.

8.4. Comparison of Ethical Perspectives

Examining documents from different ethical perspectives reveals distinct aspects of their normative structure. The deontological perspective focuses on rights, duties, and the permissibility of means; the utilitarian perspective addresses the distribution of consequences; and the ethics of care considers the circumstances of particular individuals, their relationships, and their vulnerability. Other perspectives help examine the role of interests, historical context, and institutional practices. In HAI-LEAF, these approaches complement one another and support the critical examination of arguments.

The table below compares the application of six ethical perspectives to the two groups of acts examined. Each perspective's principal contribution is identified to avoid restating a single general conclusion in different terms. The comparison aims to establish which problems each approach reveals, where arguments reinforce one another, and which questions require additional evidence. The assessments are not averaged: their persuasiveness depends on the substance of the arguments and their evidential support.

Perspective	**Nuremberg Laws**	**Decrees of 2022**
Ethical egoism	Advantages for a designated group are secured through restrictions on others	Examines the risk of substituting the issuing state's position for the interests of affected people
Deontology	Ancestry is used as a basis for granting	The stated purpose does not remove requirements concerning the permissibility of

Perspective	Nuremberg Laws	Decrees of 2022
	unequal rights	means
Utilitarianism	The burdens imposed on restricted groups are not given equal consideration	Requires comparison of military means with alternatives and associated risks
Relativism	Historical context explains the origins of the norms but does not justify discrimination	The conflict context does not automatically substantiate all asserted premises
Virtue ethics	The norms sustain practices of exclusion and coercion	Examines good faith in justification, restraint, and responsibility
Ethics of care	Specific family relationships and individual vulnerability are disregarded	The generalization "the people" requires examination of diverse needs and positions

The comparison does not imply that the cases are historically identical. It demonstrates the applicability of a common procedure to different mechanisms of normative justification.

9. Controlled Analytical Experiments

Six controlled modifications were examined to assess the framework's internal logic. These are conceptual tests of content, rather than a statistical experiment involving repeated independent runs.

Let a document modification be represented by the operator:

$$d' = T_k(d).$$

The analysis examines which components of the assessment should change after applying T_k and which should remain unchanged.

Test	Controlled Modification	Analytical Finding
T_1	Error-free determination of ancestry is assumed in D_1	Unequal political rights remain ethically unacceptable
T_2	Sanctions in D_2 are reduced without repealing the group-based prohibitions	The severity of coercion changes, but the discriminatory mechanism remains
T_3	The authenticity of a leader's request is hypothetically confirmed in the decrees	One factual question is resolved; authority and the permissibility of the means require separate examination
T_4	The military directive is given a neutral label	The assessment of the mechanism should not change

Test	Controlled Modification	Analytical Finding
	while its powers remain unchanged	solely because of its label
T_5	The military directive is removed while the recognition decision is retained	The assessment of the military mechanism loses its direct textual basis; the recognition issue remains
T_6	Verification of the accuracy of group classification is introduced without revising the rule	Protection against misclassification improves, but the inequality itself gains no additional justification

The findings demonstrate the need to distinguish between correcting information, reducing the intensity of harm, and changing the normative mechanism.

Test T_5 is particularly significant methodologically. If, after the military directive is removed, the system continues to describe it as an operative provision of the document, this would indicate reliance on general knowledge of the case rather than analysis of the text presented.

Test T_4 examines robustness to a superficial change in wording, whereas T_5 examines sensitivity to a substantive change. These properties are complementary: the system should retain well-founded assessments when the mechanism remains unchanged and revise them when their underlying grounds change [18].

10. Discussion

The methodological contribution of HAI-LEAF lies in combining normative decomposition, TPM reconstruction, evidence records, targeted critical auditing, and human decision-making.

TPM helps locate defects in justification. A problem may arise in the initial definition of a threat, the selection of information, the transition from facts to what ought to be done, or the prescribed means of implementation. Correcting one component does not guarantee the acceptability of the entire structure.

A multi-agent approach becomes methodologically meaningful when agents perform different operations on objects identified consistently across the analysis. If they merely generate several general reviews, it is difficult to determine where a disagreement originates and how it can be examined.

An important feature is the ability to conduct targeted revisions. A change to a source or norm should trigger a review of the conclusions that depend on it. This is particularly useful in legislative drafting, where successive versions evolve incrementally.

The main limitations concern the small and thematically skewed corpus, the familiarity of historical assessments, dependence on the language model, and the absence of a comparative experiment. The corpus should be expanded to include justified protective provisions, borderline cases, different jurisdictions, and acts involving genuine competition between legitimate interests.

A promising direction for development is the implementation of HAI-LEAF as a procedural module within a legal information and analytics platform, such as AgentFlowLaw. Such integration was not implemented in this study.

11. Conclusions

This paper proposes HAI-LEAF, a framework for the ethical assessment of normative acts that organizes collaboration between humans and specialized agent roles around verifiable claims.

Formal representations are provided for a normative provision, an evidence record, an ethical profile, and an inter-agent message. The agent interaction graph is distinguished from the evidence dependency graph. This enables the origins of conclusions to be traced and identifies which conclusions require review when their underlying grounds change.

The analysis shows that multi-agent interaction should include the exchange of evidence linked to precise source locations, targeted objections, and the return of findings for clarification. Voting or mechanically averaging assessments does not ensure a well-founded ethical conclusion.

The demonstration analysis of four acts and six controlled analytical modifications distinguished factual accuracy, procedural orderliness, and the acceptability of a normative mechanism. In particular, more accurate determination of a group characteristic does not eliminate its discriminatory use, while confirmation of an individual premise does not establish the permissibility of the chosen means.

An evaluation programme is proposed covering provision coverage, evidential support, omissions, unsupported allegations, robustness, and sensitivity. Further research should establish the contribution of individual agents and their communication channels under comparable resource conditions.

At its current stage, HAI-LEAF is a methodologically specified protocol examined through a demonstration study. Its comparative effectiveness remains subject to empirical evaluation.

References

1. Hart, H.L.A., 2017. Positivism and the Separation of Law and Morals. In *Law and morality* (pp. 63-99). Routledge.
2. Chen, G., Cheng, L., Tuan, L.A. and Bing, L., 2024, August. Exploring the potential of large language models in computational argumentation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 2309-2330).
3. Dahl, M., Magesh, V., Suzgun, M. and Ho, D.E., 2024. Large legal fictions: Profiling legal hallucinations in large language models. *Journal of Legal Analysis*, *16*(1), pp.64-93.
4. Guo, T., Chen, X., Wang, Y., Chang, R., Pei, S., Chawla, N.V., Wiest, O. and Zhang, X., 2024. Large language model based multi-agents: A survey of progress and challenges. *arXiv preprint arXiv:2402.01680*.
5. Zhang, H., Cui, Z., Chen, J., Wang, X., Zhang, Q., Wang, Z., Wu, D. and Hu, S., 2025. Stop Overvaluing Multi-Agent Debate--We Must Rethink Evaluation and Embrace Model Heterogeneity. *arXiv preprint arXiv:2502.08788*.

6. Cemri, M., Pan, M.Z., Yang, S., Agrawal, L.A., Chopra, B., Tiwari, R., Keutzer, K., Parameswaran, A., Klein, D., Ramchandran, K. and Zaharia, M.A., 2026. Why do multi-agent llm systems fail?. *Advances in Neural Information Processing Systems*, *38*.
7. HAI-LEAF. Human–AI Legislative Ethical Assessment Framework Version 1.0. 29 October 2026. *URL:* https://github.com/DmytroLande/HAI_LEAF
8. Rodgers, W., Cenar, I. and Gamble, G.O., 2026. Quantum-enhanced throughput pathways: Integrating Rodgers' TPM with quantum ethical frameworks for AI-driven cybersecurity. *Acta Psychologica*, *263*, p.106231.
9. United Nations Educational, Scientific and Cultural Organization, 2021. Recommendation on the ethics of artificial intelligence. *URL:* https://unesdoc.unesco.org/ark:/48223/pf0000380455
10. United Nations. General Assembly, 1949. Universal declaration of human rights (Vol. 3381). Department of State, United States of America. https://www.un.org/en/about-us/universal-declaration-of-human-rights
11. Charter, U.N., 1945. Charter of the United Nations. June, 26, p.59. https://www.un.org/en/about-us/un-charter
12. Gao, T., Yen, H., Yu, J. and Chen, D., 2023, December. Enabling large language models to generate text with citations. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (pp. 6465-6488).
13. *Reichsbürgergesetz. Vom 15. September 1935*. Reichsgesetzblatt. 1935. Teil I. S. 1146.
14. *Gesetz zum Schutze des deutschen Blutes und der deutschen Ehre. Vom 15. September 1935*. Reichsgesetzblatt. 1935. Teil I. S. 1146–1147.
15. Указ Президента РФ від 21.02.2022 № 71.
16. Указ Президента РФ від 21.02.2022 № 72.
17. Assembly, U.G., 2022. Resolution ES-11/5: Furtherance of remedy and reparation for aggression against Ukraine (A/RES/ES-11/5). *URL: https://digitallibrary.un.org/record/3993910*
18. Cho, S., Ruberto, S. and Terragni, V., 2025, September. Metamorphic testing of large language models for natural language processing. In *2025 IEEE International Conference on Software Maintenance and Evolution (ICSME)* (pp. 174-186). IEEE.