Framework for Openly Augmented Mediation (FOAM): A Pluralistic Architecture for Explainable and Contestable AI

ANONYMOUS AUTHOR(S)

High-stakes AI systems increasingly mediate access to credit, healthcare, and public benefits, yet affected parties often cannot see why a decision was made or meaningfully contest it. Even post hoc review of chain-of-thought traces from individual models can be incomplete or strategically misleading, thereby limiting accountability. We propose FOAM, a pluralistic architecture for multi-agent language systems that treats explanation as a deliberative process where differentiated agents advance value- and role-specific arguments, a protocol structures rebuttal and evidence challenges, and a synthesis operator outputs both a recommendation and the surviving points of contention with sentence-level provenance. We implement FOAM within a policy-debate case-generation system and evaluate it in a blinded tournament of 66 cases using automated multi-criteria evaluation and independent evidence verification. FOAM outperforms human-expert and zero-shot model baselines on overall quality (81.7 vs. 70.1 and 50.6) and yields substantially higher perfect-evidence validation (76.2% vs. 8.7% and 0%), thereby enabling downstream auditing and dispute resolution. We discuss how deliberative architectures can operationalize the requirements of transparency and contestation in emerging governance regimes and outline safeguards for dual-use persuasive capabilities.

Additional Key Words and Phrases: Algorithmic accountability; Contestable AI; Explainable AI (XAI); Multi-agent deliberation; Evidence provenance

ACM Reference Format:

Anonymous Author(s). 2026. Framework for Openly Augmented Mediation (FOAM): A Pluralistic Architecture for Explainable and Contestable AI. 1, 1 (January 2026), 18 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

1.1 Accountability gap in high-stakes AI

AI systems are now routinely embedded in high-stakes decision workflows—healthcare triage and documentation, hiring and workplace management, credit and insurance, public benefits, and criminal-legal risk assessments. In these settings, "performance" cannot be reduced to predictive accuracy or user satisfaction: when a system's output influences outcomes that materially affect people's rights, opportunities, or safety, **accountability requires (i) intelligible reasons and (ii) effective avenues to challenge and revise those reasons**. Yet most deployed AI remains organized around a monolithic model that produces a single authoritative output, with limited transparency into *why* it said what it said and little procedural support for contesting it when it is wrong, biased, or normatively inappropriate.

This accountability gap has two tightly coupled dimensions. **Explainability** is often treated as a documentation problem—generate a rationale, a summary, or a list of features—rather than a *reason-giving* problem grounded in the kinds of explanations different stakeholders actually need (e.g., diagnostic vs. role-based explanations) [24]. **Contestability**, meanwhile, is frequently bolted on as an afterthought (appeals processes, "report a problem" buttons, or generic feedback loops) rather than built into the architecture of reasoning itself. Meaningful contestability requires at least (a) visibility into decision logic, (b) comprehensibility for affected parties, and (c) actionable mechanisms for challenge and revision [2]. A

2026. Manuscript submitted to ACM

system that cannot surface its operative assumptions, show its evidentiary basis, and support structured disagreement cannot plausibly satisfy these conditions—especially in domains where reasonable stakeholders legitimately disagree about values, tradeoffs, and acceptable risk.

1.2 Why post-hoc "explanations" break: the faithfulness problem

A central reason current explainability tooling struggles is that it frequently relies on **post-hoc self-explanation from the same model that produced the decision**. For large language models in particular, chain-of-thought and rationale-style explanations can be fluent and persuasive while remaining weakly coupled to what actually drove the output. Chen et al. benchmark state-of-the-art reasoning models and report low overall faithfulness scores—e.g., **25% for Claude 3.7 Sonnet and 39% for DeepSeek R1** under their evaluation design—highlighting that models may omit or misrepresent key determinants of their answers even when explicitly prompted to "show their work" [7]. Related work similarly emphasizes that CoT can be misleading as an interpretability proxy, especially when users treat it as a reliable window into computation rather than a generated text artifact.

This "faithfulness gap" creates a direct accountability failure mode: if the explanation channel can drift from the decision channel, then transparency becomes performative—useful for persuasion, but unreliable for oversight, auditing, or recourse. In high-stakes contexts, that is not a subtle limitation; it is a design-level mismatch between what institutions need (verifiable reasons and traceable evidence) and what monolithic systems can robustly provide. The core implication is architectural: **if we want explanations that can support contestation, we need systems that can produce multiple, checkable reason-giving traces—not a single narrative generated by the same mechanism being explained.** This motivates pluralistic approaches that externalize disagreement, force explicit warrants, and attach provenance to claims so that challenges can target the actual moving parts of the reasoning.

1.3 What we propose (FOAM) and what is new

This paper develops and evaluates **pluralistic AI systems** that operationalize explainability and contestability through **structured multi-agent deliberation** rather than post-hoc narration. We introduce **FOAM (Framework for Openly Augmented Mediation)**, an architecture that treats accountable AI outputs as the product of a mediated process:

(1) **Differentiated agents** with distinct roles and epistemic commitments (e.g., advocate, skeptic, evidence-checker, values/impact assessor),
(2) **Deliberative protocols** that require agents to advance and respond to claims under explicit constraints (e.g., argument typing, cross-examination, and structured rebuttal), and
(3) **Sublation operators**—formal mechanisms for preserving what survives critique while revising what fails, so that the system's final output is not merely an average of perspectives but a documented transformation through contestation.

The intended artifact is not just a recommendation, but a contestable record: claims, counterclaims, evidentiary supports, explicit points of disagreement, and the rationale for any resolution.

We make three contributions:

(1) **Framework:** we provide a unified account of explainability *and* contestability as a single design target, arguing that they should be treated jointly and realized through pluralistic mediation rather than monolithic self-report.
(2) **Architecture and mechanisms:** we formalize FOAM as an implementable blueprint—agents, protocols, and revision operators—paired with provenance-oriented design choices that make challenges actionable (e.g., grounding claims in checkable evidence rather than free-form summarization).
(3) **Empirical validation:** we report results from an evaluation of pluralistic debate generation in a double-blind tournament of **66 policy debate cases**, where our structured multi-agent system achieved an overall score of **81.7** compared to **70.1** for human experts and **50.6** for zero-shot AI, while also achieving **76.2%** perfect evidence validation compared to **8.7%** for human experts and **0%** for unstructured AI—demonstrating that pluralistic architectures can produce outputs that are simultaneously more persuasive *and* more verifiable in an adversarial, evidence-sensitive setting.

We close by discussing implications for AI governance and by outlining a research agenda for **contestable AI by design**.

2 ACCOUNTABILITY REQUIREMENTS AND RELATED WORK

2.1 Explainability requirements beyond transparency

Contemporary calls for "explainable AI" often conflate **transparency** (exposing internal mechanisms) with **explanation** (providing reasons that are meaningful for a particular audience and purpose). Lipton argues that interpretability is not a single property and that many "explanations" in ML function as *post-hoc rationalizations* whose relationship to actual model behavior is ambiguous, especially when the explanation's audience is a regulator, decision-subject, or domain expert rather than a model developer [23]. Relatedly, Doshi-Velez & Kim emphasize that interpretability claims must be made relative to **use context**—including the user's expertise, the stakes, and the kind of decision being supported—because what counts as a satisfactory explanation differs across settings [9]. In high-stakes domains, this motivates either (i) models that are inherently interpretable, or (ii) explanation mechanisms that achieve a comparable standard of *reliability and auditability* rather than superficial plausibility [29].

For accountability, explanations must be more than persuasive narratives; they must be **diagnostically useful** and **robust to strategic manipulation**. Empirically, Adebayo et al. show that post-hoc explanations can fail as diagnostic tools—e.g., they may not reliably reveal spurious correlations that drive model behavior—undercutting the hope that explanation interfaces alone can serve as accountability checks [1]. More broadly, the NLP interpretability literature distinguishes *plausibility* (does an explanation look reasonable to humans?) from *faithfulness* (does it track the true basis of the model's output?), and argues that faithful explanations require evaluation criteria and designs that go beyond "nice-sounding" rationales [15]. As a result, explainability requirements in FAccT-relevant deployments should be stated in terms of **checkability**: the ability to trace claims to concrete support, interrogate counterfactuals, and isolate points of disagreement, rather than merely presenting a single coherent story [15, 24].

2.2 Contestability as a system property

Explainability alone does not guarantee that affected parties can meaningfully challenge an AI-mediated decision; contestability is best treated as a **system-level governance property** rather than an after-the-fact user interface feature. Alfrink et al. frame "contestable AI by design" as the view that systems should be built to *support* contestation—through traceability, structured justification, and pathways for challenge—rather than treating contestation as an external legal or organizational process that happens "around" the model [2]. Legal scholarship on automated decision-making similarly emphasizes that accountability requires more than disclosure: decision-subjects need procedures to *question, rebut, and obtain redress*, and these procedures depend on the availability of intelligible grounds and records of how outputs were produced [20]. This is particularly important because the existence and scope of a freestanding "right to explanation" under the GDPR is contested, with influential analyses arguing that GDPR does not straightforwardly provide a general right to detailed model explanations—reinforcing the need for contestability mechanisms that do not rely on a single doctrinal reading of transparency rights [36].

Operationally, contestability implies three minimal requirements:

(1) **Visibility** that an automated or AI-assisted decision has occurred and can be challenged;
(2) **Comprehensibility** of the stated grounds and supporting materials; and
(3) **Actionability**, meaning a practical pathway to present counterevidence/counterarguments and obtain review and potential revision [2, 20].

The GDPR is relevant here not only through transparency provisions, but also because Article 22 and associated provisions are commonly read as requiring procedural hooks such as the ability to obtain human intervention and contest certain automated decisions, even if the precise informational entitlements are debated [12, 36]. Complementing legal requirements, the EU High-Level Expert Group's Trustworthy AI guidance explicitly treats accountability as including mechanisms for redress and the capacity to challenge outcomes, which aligns with FAccT's emphasis on socio-technical accountability rather than purely technical interpretability [11]. These sources jointly motivate a design target: **contestability must be implemented as an end-to-end workflow** that links reasons to evidence and enables structured challenge, rather than as a static explanatory artifact [2, 26].

2.3 Pluralistic and deliberative approaches to accountability

In many high-stakes settings, disagreement is not merely empirical ("what are the facts?") but normative ("which values should dominate?"). Feminist epistemology and science studies have long argued that knowledge claims are situated and that purportedly "view from nowhere" objectivity can mask whose interests and assumptions are being operationalized [13]. In governance terms, Dewey's account of public problem-solving similarly emphasizes that collective inquiry is iterative and that institutions must be structured to surface and revise the premises that guide decision-making, especially under conditions of uncertainty and plural publics [8]. For AI accountability, these traditions motivate an architectural stance: rather than forcing a single model to output one authoritative rationale, systems should be designed to make **value trade-offs explicit** and to preserve dissenting considerations in a form that can be examined and contested [13, 24].

Recent work in value alignment and governance likewise emphasizes that "alignment" is underdetermined when stakeholders disagree about objectives, priorities, and acceptable risks. Kasirzadeh distinguishes forms of alignment that presume a single coherent value target from approaches that treat plural and conflicting values as first-class constraints—implying that accountability mechanisms must represent disagreement rather than suppress it [19]. In parallel, "society-in-the-loop" framings argue that algorithmic systems are components of an evolving social contract and therefore require institutionalized interfaces for dispute, oversight, and revision [25]. In FAccT terms, these perspectives justify **pluralistic explanation**: not as an optional UX feature, but as a governance mechanism that helps stakeholders identify where the system's reasoning depends on contestable assumptions [19, 25].

2.4 Multi-agent deliberation and debate in AI

A technical pathway to operationalizing pluralism is to replace monolithic generation with **structured multi-agent deliberation**, including debate-style protocols. In AI safety, "debate" was proposed as a scalable oversight mechanism in which adversarial argumentation can surface flaws or deception that a single system might otherwise hide [14]. Subsequent theoretical work studies conditions under which debate can be made efficient and verifiable, strengthening the conceptual link between adversarial dialogue and reliable oversight [4]. Empirically, multi-agent debate among language models has been reported to improve factuality and reasoning in some settings, suggesting that disagreement and cross-examination can function as error-correction dynamics rather than mere rhetoric [10]. However, most "LLM debate" results are evaluated in terms of accuracy or judge preference; they do not, by themselves, guarantee that the resulting justifications are **auditable** or that third parties can meaningfully contest specific premises, evidence selections, or value judgments [15, 29].

Computational argumentation provides complementary foundations for making deliberation outputs contestable because it supplies explicit representations of **claims, warrants, attacks, defenses, and (in value-based variants) normative priorities**. Toulmin's model remains foundational for analyzing argument structure in terms of claims supported by warrants and backing [34]. Formal work in AI argumentation further develops abstract and assumption-based frameworks for representing defeasible reasoning, while value-based argumentation captures how outcomes change when different values are prioritized [3, 33]. Surveys connecting argumentation and XAI argue that these representations can support explanation as a structured object of inquiry—closer to an "inspectable case" than a narrative rationale—because stakeholders can contest particular premises or inference steps and observe how the conclusion changes [35]. This literature motivates the core related-work claim that a *contestable* AI system should produce not only an answer, but also a **dispute-ready argumentative record**: reasons decomposed into contestable units, linked to supporting materials, and amenable to revision under challenge [20, 35].

3 FOAM APPROACH: PLURALISTIC ARCHITECTURE FOR EXPLAINABILITY AND CONTESTABILITY

3.1 Design goals and accountability threat model

Building on the accountability requirements in Section 2, we treat *explainability* and *contestability* as properties of an **epistemic process**, not a post-hoc narrative from a monolithic model. Concretely, we introduce **FOAM (Framework for Openly Augmented Mediation)**: a pluralistic, multi-agent architecture that produces an answer *plus* a structured record of how that answer was stress-tested, revised, and synthesized. FOAM is organized around three primitives:

(1) *Differentiated agents* parameterized by explicit and persistent stance data structures at test time,
(2) *Deliberative protocols* that force critique and revision, and
(3) *Sublation* operators that synthesize without erasing disagreement.

Figure 1 provides a system overview: stance seeding → local drafting → cross-examination → evaluation → revision/sublation → accountability artifacts.

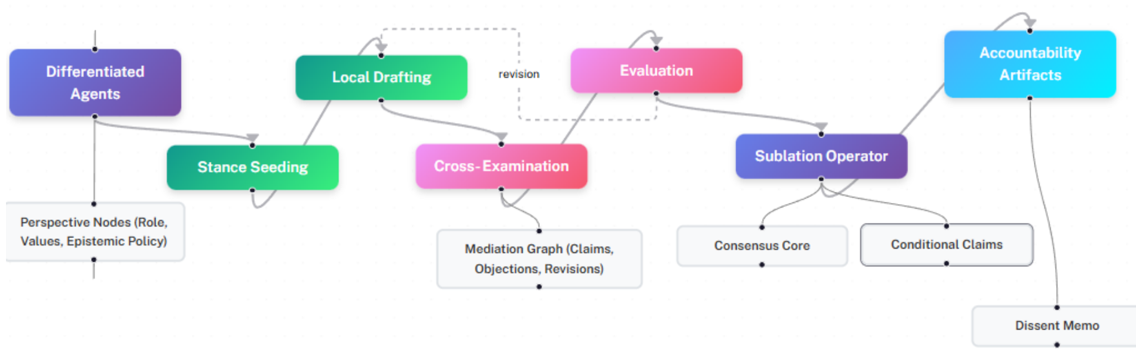

Fig. 1. FOAM system architecture. Differentiated agents with explicit perspective nodes engage in deliberative protocols (stance seeding, local drafting, cross-examination, evaluation) that produce accountability artifacts including a consensus core, conditional claims, and dissent memo. The mediation graph tracks claims, objections, and revisions throughout the process.

Our threat model is accountability-centric: we assume that base generative models can (a) produce fluent but false claims and fabricated or misattributed support ("hallucination"), (b) rationalize decisions after the fact, (c) collapse multiple stakeholder perspectives into a single dominant frame, and (d) bury value tradeoffs inside unstructured prose such that stakeholders cannot identify *what*, precisely, to challenge. These failure modes do not require adversarial intent; they are well documented in contemporary NLP systems and can persist even under strong prompting [16]. FOAM's core design choice is therefore to make *points of potential failure* explicitly addressable: disagreements are surfaced rather than smoothed, objections are represented as first-class objects, and synthesis is constrained to preserve traceability from contested premises to final recommendations.

3.2 Differentiated agents via explicit perspective and stance representation

FOAM begins by instantiating a small set of agents (n chosen by stakes and time budget), each assigned an explicit data structure represented as a *Perspective Node* and stored in a vector database that encodes *who the agent is meant to be epistemically*—its domain role, value priorities, and reasoning schema. This implements "situated" explanation in a directly auditable way: instead of implicitly claiming neutrality, the system discloses positions and thereby enables critique of the *perspective selection* itself [13]. In FOAM, perspective nodes are not just labels; they are operational constraints that shape what evidence is considered legitimate, which impacts are foregrounded, and which argument schemes are preferred.

Practically, we treat a perspective node as a structured record with three minimum components:

(1) **Role** (e.g., regulator, clinician, affected community advocate),
(2) **Normative weighting** (e.g., safety vs autonomy vs distributive equity), and
(3) **Epistemic policy** (e.g., what counts as acceptable support; how uncertainty must be qualified).

During deliberation, FOAM enforces *stance coherence*: if an agent's generated warrants or qualifiers contradict its declared stance, the system requests revision or flags the inconsistency for downstream inspection. This is the anti-"performative pluralism" mechanism: pluralism is only accountability-relevant if the system can show (and users can contest) whether distinct perspectives were actually maintained rather than rhetorically simulated.

Perspective nodes also make **second-order contestation** practical: stakeholders can dispute not only the system's conclusion, but the *legitimacy of the value and perspective configuration* that produced it (e.g., "Why is utilitarian cost-effectiveness even in scope here?"). This matters because pluralistic systems can otherwise "value-wash" by claiming inclusivity while quietly privileging one evaluative frame. FOAM makes the stance set an explicit input and therefore a target for governance and oversight; this aligns with work arguing that legitimacy depends on making value choices and their selection procedures contestable [18]. In deployment terms, this means FOAM can be rerun with (i) added perspectives, (ii) reweighted value priorities, or (iii) altered evidentiary rules, producing *comparative, contestable* outcomes rather than a single authoritative verdict.

3.3 Deliberative protocol: dialectical refinement and mediation trace

FOAM's deliberation is implemented as a **mediation loop**:

(1) *Seeding* (instantiate agents + perspective nodes),
(2) *Local drafting* (agents generate independent proposals),
(3) *Cross-examination* (agents issue structured objections and targeted questions),
(4) *Evaluation* (a judge/jury component scores draft–objection pairs against criteria), and
(5) *Revision + synthesis* (agents revise and a sublation operator composes the provisional output).

The accountability point is not that deliberation guarantees truth; it is that **deliberation guarantees structured opportunities to find and localize error**, and then to record what happened when error was raised.

Cross-examination produces a **mediation graph**: a structured trace that links *which agent* made *which claim*, what objections were raised (e.g., missing evidence, value conflict, logical gap), how the claim was revised, and which surviving claims contributed to the synthesis. This is the audit primitive: contestation requires that stakeholders can point to *the specific node* where they disagree and see what depended on it. As an interoperability target, the mediation trace can be expressed using standard provenance representations (e.g., PROV-O) so that downstream tools can query "what influenced what" across a run [22].

3.4 Sublation: synthesis without erasure

After critique and revision, FOAM applies a **sublation operator**: a synthesis step intended to preserve what is valuable in competing positions while explicitly retaining unresolved tensions. In FOAM, sublation is not a rhetorical flourish; it is a concrete rule: synthesis is disallowed from silently discarding objections that were scored as material or from collapsing incompatible value frames into an unmarked compromise.

Sublation emits three artifacts: **a consensus core** (claims surviving cross-stance critique), **conditional claims** (branching on unresolved value priorities), and a **dissent memo** (recording conflicts and contested premises).

3.5 Inspectable argument structure: Toulmin decomposition and typed syllogisms

To make contestation actionable, FOAM constrains agent outputs into an **inspectable argument structure** rather than free-form prose. We adopt Toulmin-style decomposition—claim, grounds, warrant, backing, qualifier, rebuttal—because it maps naturally to "what can be challenged": stakeholders can contest evidence (grounds), the inferential link (warrant), scope conditions (qualifier), or missing counterevidence (rebuttal) [34]. This structure also aligns with prior work connecting computational argumentation to explainable AI, where explanations are made more useful by exposing structured reasons and counterreasons rather than only surface-level narratives [35].

FOAM additionally employs **typed syllogisms**—domain-relevant argument templates that enforce completeness (e.g., in policy debate: Advantage = Uniqueness + Link + Impact; Disadvantage = Uniqueness + Link + Impact; Kritik = Link + Impact + Alternative). These structures are standard in competitive policy debate pedagogy and make dependencies explicit for non-expert audiences [31]. In FOAM, typed syllogisms function as contestability scaffolds: if a stakeholder disputes the conclusion, the system can point to the *specific missing or weak component* (e.g., "impact evidence absent" or "link warrant unsupported"), and the mediation graph can show whether that component was ever raised in critique and why it survived. The result is a system where "challenge" is not a vague request to "explain more," but a targeted operation on a specific argumentative component with traceable upstream dependencies.

Template tree traversal operationalizes structural contestability. At each branch point, the system records which template was selected (e.g., "traditional 1AC with 3 advantages" vs. "kritik with alternative"), what resource allocation was applied, and whether any novel templates were generated. This trace enables a distinct class of challenges: stakeholders can dispute not only *what* claims were made, *but why the argumentative structure took this form rather than another.* For instance, a reviewer might contest that a utilitarian impact calculus was chosen when the underlying values favor a rights-based framing—and the template selection trace makes this challenge actionable. Furthermore, unlike traditional chain-of-thought reasoning where reasoning and response are interwoven and in some cases reasoning is not always a reliable indicator for why outputs occurred, the template tree traversal process is a discrete step occurring prior and serving as a foundational infrastructure to drafting.

4 CASE STUDY SYSTEM: EVIDENCE-GROUNDED POLICY DEBATE GENERATION

4.1 Why policy debate is an accountability crucible

We instantiate FOAM in a domain where *contestability is native to the task*: American competitive policy debate. Policy debate is a two-team adversarial format in which teams argue for and against a policy proposal under strict procedural constraints. In this ecosystem, argument quality is not evaluated purely as rhetorical fluency; instead, the activity is structured around *traceable evidentiary support* and explicit clash, so claims can be challenged in real time and revisited across subsequent speeches. Critically, policy debate operationalizes "grounding" through an established evidence artifact: the *debate card*. A card typically includes (i) a short biased summary intended to support a specific argumentative function, (ii) a full citation, and (iii) verbatim quoted source text, often with token-level highlighting that marks precisely what will be read into the round. Competitive success is strongly coupled to evidence quality and its deployment, creating an evaluation environment where provenance and verifiability are not optional.

4.2 Pipeline overview

Figure 2 summarizes our **five-phase pipeline** for generating an evidence-grounded constructive speech (the 1AC, in our evaluation setting). Phases 1–3 produce an inspectable argumentative plan in typed components (perspective assignment → strategic plan → template traversal), Phase 4 binds each argumentative component to *verbatim evidence at sentence granularity* (sentence-level provenance), and Phase 5 compiles and verifies the result (structural conformance, evidence/claim alignment, and perspective consistency). The key design principle is to keep the model in a role where it can be audited: rather than "write a persuasive case and cite sources," the system decomposes "case construction" into a sequence of constrained decisions that leave a machine-checkable trail.

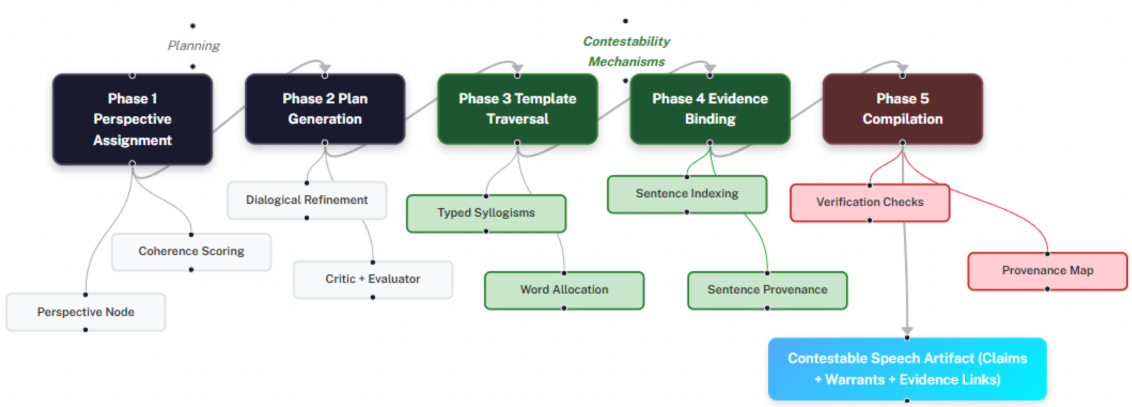

Fig. 2. Five-phase pipeline with accountability mechanisms. Phases 1–3 (Perspective Assignment, Plan Generation, Template Traversal) handle argumentative planning. Phase 4 (Evidence Binding) creates sentence-level provenance by selecting specific sentence IDs rather than paraphrasing. Phase 5 (Compilation) enforces verification checks. The output is a contestable speech artifact with claims, warrants, and traceable evidence links.

4.3 Phases 1–3: perspective assignment, planning, and template traversal

Phases 1–3 produce an inspectable argumentative plan through three contestability-relevant operations. In **Phase 1**, the system assigns an explicit perspective node (Section 3.2), making the evaluative frame a first-class auditable choice. In **Phase 2**, a dialectical refinement loop stress-tests the strategic plan: a Critic agent issues typed objections (logical gap, missing evidence, value conflict, scope overreach), an Evaluator scores each objection's materiality, and the Proposer revises or rebuts. This cycle iterates at least three times, and *all objections—including dismissed ones—remain in the mediation graph*, enabling downstream reviewers to inspect whether a weakness was raised and why the response was deemed adequate.

In **Phase 3**, template tree traversal expands the plan into a typed syllogism scaffold (e.g., Advantage = Uniqueness + Link + Impact). At each branch point, the system records which template was selected, what word allocation was applied (e.g., 30% impact, 40% link), and whether novel templates were generated. This trace enables a distinct class of challenges: stakeholders can dispute not only *what* claims were made, but *why the argumentative structure took this form rather than another*—for instance, contesting that a utilitarian impact calculus was chosen when the underlying values favor a rights-based framing.

4.4 Phase 4: sentence-level provenance

Motivation. Retrieval-augmented generation can reduce hallucinations, but it does not eliminate a central accountability failure mode: models may still produce claims that are *unsupported by*, *in conflict with*, or *misattributed to* retrieved text. Recent benchmarks explicitly document that, even under RAG setups, LLM outputs can contain unsupported or contradictory content relative to the retrieved passages. Phase 4 therefore implements a stronger constraint than "retrieve then paraphrase": it forces the model to operate over *sentence identifiers* rather than free-form rewriting of source material.

Mechanism. Phase 4 is a two-step procedure:

Step (a): sentence indexing and retrieval. The system queries (i) a debate-evidence store (implemented in our current system as a vector database over a large set of debate "cards") and (ii) any other preprocessed sources permitted by the pipeline. Retrieved documents are segmented into sentences, each assigned a stable index, and returned to the deliberation workspace as a set of candidates with identifiers of the form `(document_id, sentence_id)` plus immutable citation metadata.

Step (b): evidence selection and tagging. The LLM is then prompted to (1) select which sentence IDs support each argument slot created in Phase 3 and (2) generate only a short "tag" that states what the selected evidence is being used to establish. Importantly, the model is not asked to restate the evidence; the evidence content in the final speech is assembled from the retrieved sentences themselves. This design eliminates an entire class of failure (fabricated quotations and invented citations) by construction: the model can be wrong about *which* sentences to use, but it cannot invent sentences that are not in the retrieved set.

Accountability and contestability properties. Sentence-level provenance changes the contestation workflow from "argue about what the model meant" to "inspect exactly what the model relied on." A stakeholder can challenge (i) *relevance* ("this sentence does not establish the warrant you claim"), (ii) *adequacy* ("the evidence is too weak/out of context"), or (iii) *selection bias* ("you ignored stronger counterevidence available in the same corpus")—and each challenge targets a concrete object (a sentence ID and its parent source). This is especially aligned with policy debate's evidence norms, which already treat quoted and highlighted text as the unit of disputation under cross-examination.

4.5 Phase 5: compilation and verification checks

Phase 5 compiles the typed argument scaffold (Phase 3) and the evidence bindings (Phase 4) into a final speech artifact suitable for evaluation. Compilation preserves the provenance map: each substantive claim in the rendered speech remains traceable to one or more sentence IDs plus citation metadata. The system then runs verification checks that are directly tied to the accountability requirements:

(1) **Structural completeness** (template validators—e.g., required components are present),
(2) **Evidence/claim alignment** (each slot has at least one bound sentence; missing bindings fail closed), and
(3) **Perspective consistency** (warrants and impacts do not contradict the declared perspective node from Phase 1).

Figure 2 highlights where provenance is created (Phase 4) and where it is enforced (Phase 5).

5 EMPIRICAL EVALUATION

5.1 Research questions

We evaluate FOAM's accountable-generation claims using an *audit-style* design: we define explicit research questions, compare against salient baselines, and report both performance outcomes and traceability outcomes as first-class metrics. This approach aligns with established work on internal algorithmic auditing and emerging "assurance audit" perspectives, which emphasize that accountability requires not only outcome quality, but also artifacts and procedures that make decisions inspectable and challengeable [21, 26].

We ask whether FOAM improves:

- **RQ1:** Quality/persuasiveness
- **RQ2:** Evidence verifiability
- **RQ3:** Whether gains are attributable to the accountability mechanisms rather than model strength

5.2 Experimental design and baselines

Task selection. We evaluate in evidence-grounded policy debate generation because it combines (i) long-horizon argumentative planning, (ii) adversarial robustness expectations (arguments must survive challenge), and (iii) strict evidentiary norms (claims are conventionally supported with citations). In computational argumentation, even highly resourced systems have historically relied on constrained debate settings and bespoke pipelines; the Project Debater line of work illustrates both the ambition of debate as a benchmark and the practical need to structure and constrain the task for reliable evaluation [30].

Debate artifact. We focus on the **first affirmative constructive (1AC)** as the most demanding generative unit in competitive policy debate: it must introduce a full strategic position (advantages/disadvantages/solvency framing), anticipate common lines of negative attack, and do so under tight length constraints while maintaining evidentiary support. This makes the 1AC a strong proxy for high-stakes accountable generation: arguments must be *comprehensible*, *internally coherent*, and *traceable to evidence* to be meaningfully contestable.

Corpus and baselines. We ran a **double-blind tournament of 66 cases** drawn from three sources:

(1) **FOAM-based structured system** ("DebaterHub Structured System," $n = 22$), generated via differentiated perspectives, iterative dialectical refinement, typed syllogisms, and sentence-level provenance;
(2) **Human expert baseline** ($n = 23$), sampled from prestigious debate camps (Dartmouth, Georgetown, Michigan, Emory); and
(3) **Zero-shot AI baseline** ($n = 21$), produced by frontier models (Gemini/Claude/ChatGPT/Grok) using prompt engineering and web-research access but without debate-specific pluralistic architecture.

Evidence corpus for provenance. FOAM's evidence retrieval and validation leverage a structured debate-evidence corpus derived from OpenDebateEvidence, which (as released) contains **3.5M+** competitive debate documents with metadata useful for downstream argument mining and citation [27]. Operationally, our system queries a vector database of ∼85,000 curated "cards" plus any newly processed sources, and the generation pipeline preserves *sentence-level identifiers* so that downstream reviewers can trace claims to exact supporting spans.

5.3 Judging rubric and scoring

Tournament format and blinding. All submissions were anonymized and assigned unique IDs (e.g., `Case_001`), and judging proceeded purely on content without revealing origin. Cases advanced through a modified Swiss-style bracket with double elimination, and pairings were balanced by strategic approach (e.g., traditional policy vs. kritik) to reduce "judge adaptation" artifacts. Ties within a narrow score band triggered evidence validation as a tiebreaker, keeping accountability-relevant verifiability salient in advancement decisions.

Rubric and judge. A Claude Opus 4 judge evaluated each case on five weighted dimensions:

- **Argumentation Strength** (25%)
- **Evidence Quality** (25%)
- **Strategic Coherence** (20%)
- **Innovation** (15%)
- **Competitive Viability** (15%)

The rubric was designed to reward both argumentative competence and evidence-groundedness, while preserving enough structure for reproducibility.

5.4 Evidence validation methodology

Why evidence validation is an accountability metric (not just "anti-hallucination"). In contestable systems, stakeholders must be able to *locate* and *evaluate* the grounds of a claim—especially where persuasive language can obscure weak or missing support. Audit frameworks similarly emphasize that assurance depends on traceable evidence artifacts rather than outcome plausibility alone [21, 26]. We therefore operationalize verifiability as a measurable property of each case's citations.

Automated citation checks and categories. Each citation was automatically checked against the referenced source (via URL or resolvable reference), and classified into one of four buckets: **exact match**, **partial match**, **paraphrase**, or **fabricated**. We summarize results primarily via **Perfect Validation**, a stringent metric that counts only **exact matches**—i.e., the cited claim can be located verbatim in the referenced source span. This is intentionally conservative: Perfect Validation corresponds to the strongest form of contestability, where an affected party can directly inspect the cited text without interpretive debate about semantic similarity.

How FOAM changes the validation problem. FOAM's sentence-level provenance changes citation validation from a semantic retrieval problem into a *pointer integrity* problem: the model is never asked to reproduce source text, but instead selects sentence indices from retrieved documents and attaches them to specific argument components. This design greatly reduces degrees of freedom for fabrication and enables deterministic re-checking of a case's evidentiary backbone.

5.5 Results

Main tournament outcomes. Table 1 reports aggregate performance by source. The FOAM-based system achieved the highest overall score (**81.7**) relative to human experts (**70.1**) and zero-shot AI (**50.6**). The largest gap appears in **Evidence Quality** (**86.7** vs. **56.9** vs. **27.1**), consistent with the claim that provenance-constrained generation shifts the system from persuasive-but-unreliable outputs toward persuasive-and-grounded outputs.

Table 1. Tournament Results by Source

Metric	FOAM	Human Expert	Zero-shot AI
Overall Score	81.7	70.1	50.6
Evidence Quality	86.7	56.9	27.1

Evidence validation and verifiability. Table 2 reports Perfect Validation rates. FOAM achieved **76.2%** Perfect Validation, compared to **8.7%** for the human expert baseline and **0%** for zero-shot AI. This is the central accountability result: the FOAM pipeline does not merely produce arguments that a judge model rates as "good," but produces arguments whose evidentiary support can be mechanically verified at scale.

Table 2. Perfect Validation Rates

Source	Perfect Validation (%)
FOAM System	76.2
Human Expert	8.7
Zero-shot AI	0.0

Interpreting what is doing the work. Two mechanisms plausibly drive the observed gap: (i) **pluralistic deliberation** (multi-perspective critique and refinement) improves strategic coherence and argument coverage, while (ii) **sentence-level provenance** directly improves evidence integrity and sharply limits fabrication opportunities. Consistent with this interpretation, the tournament champion (`Case_045`, "Navy Underwater Exploration") achieved **fidelity = 1.0** alongside a strong final-round score, indicating that high persuasive quality and high verifiability can co-occur under the FOAM constraint regime.

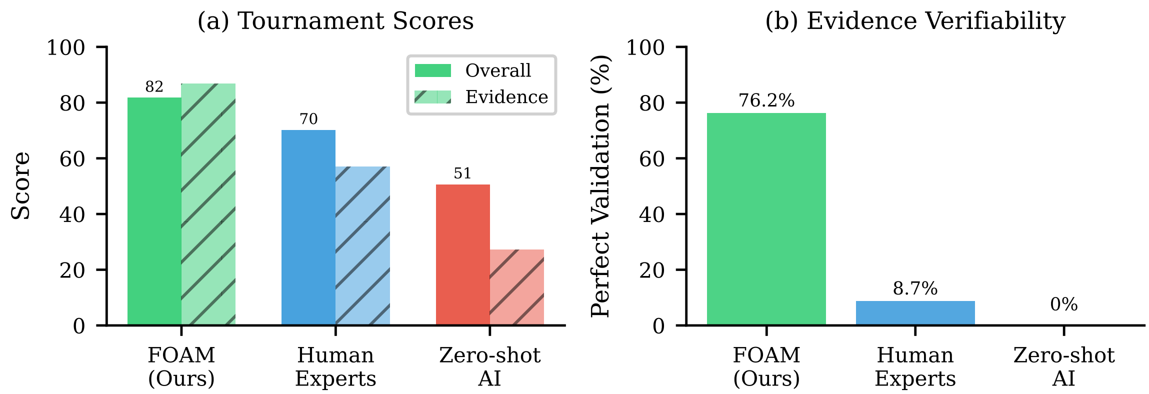

Fig. 3. Tournament results comparing FOAM, human expert baselines, and zero-shot AI. (a) Overall and Evidence Quality scores. (b) Perfect Validation rates—the percentage of citations that exactly match source text. FOAM achieves 76.2% perfect validation vs. 8.7% for human experts and 0% for zero-shot AI.

6 IMPLICATIONS FOR ACCOUNTABLE AI SYSTEMS

FOAM reframes explanation as a contestable record rather than a post-hoc narrative. Instead of producing a single rationale, the system outputs (i) an auditable argument structure (claims, warrants, rebuttals), (ii) explicit perspective configurations, and (iii) sentence-level provenance linking each substantive claim to a checkable source span. This shifts accountability from "did the explanation sound plausible?" to "which premises and evidence does the output depend on, and where can a challenge be lodged?"

Operationally, FOAM supports contestation at three levels: (1) **evidence disputes** (a cited sentence does not support the tagged claim; missing counterevidence), (2) **inferential disputes** (the warrant connecting evidence to conclusion is invalid or incomplete), and (3) **normative disputes** (the perspective/value configuration is illegitimate or incomplete for the context). Because these objects are explicit, a reviewer can localize disagreement to specific nodes and request revision without reopening the entire output as free-form prose.

Institutionally, the resulting artifact functions as an auditable dossier that can plug into existing governance workflows (internal review, incident response, assurance audits, and post-hoc dispute resolution). The technical contribution is not replacing due process, but supplying the structured, traceable materials that make procedural review feasible at scale.

7 LIMITATIONS AND FUTURE WORK

7.1 Methodological limitations and validity threats

First, our primary outcome measure relies on an automated judge (Claude Opus 4) to score debate artifacts under a fixed rubric. While LLM-as-judge evaluation is increasingly standard at scale, it is known to exhibit systematic biases (e.g., position effects, verbosity/style sensitivity, and self-enhancement tendencies) and may be vulnerable to prompt- or framing-based perturbations that shift preferences without corresponding semantic differences [6, 32, 37]. We reduce—but do not eliminate—these threats via double-blinding, standardized prompts, and by pairing judge scores with an independent evidence-validation audit.

Nevertheless, the reported tournament results should be interpreted as descriptive for this evaluation setup, and future replications should triangulate across multiple judge models and human adjudication.

Second, our system's accountability guarantees are conditioned on the properties of the underlying evidence substrate. Sentence-level provenance constrains the model to point to specific source sentences rather than inventing citations, but it does not ensure that the retrieved evidence is complete, representative, or up to date. Coverage gaps, topical skew, and retrieval errors can shape which arguments are discoverable, and can yield outputs that are "well-cited" yet misleading due to selection effects, over-aggregation, or missing context [28]. These concerns are not unique to debate generation: any contestability mechanism built on curated corpora inherits the corpus' blind spots. Accordingly, FOAM should be viewed as an approach to making claims auditable and challengeable—not as a guarantee that the selected evidence is normatively "best" or epistemically sufficient.

Third, our evaluation scope is intentionally narrow and therefore limits external validity. We benchmark a specialized argumentative domain (policy debate) and a bounded artifact type (constructive case generation), and we do not yet measure downstream stakeholder contestation behaviors (e.g., whether affected parties can efficiently detect, understand, and successfully challenge specific warrants or citations). Additionally, our "perfect validation" metric is strict by design: it favors verbatim traceability and can under-credit faithful paraphrase or correct claims supported by multiple dispersed sentences. Conversely, the metric may fail to detect other fidelity failures (e.g., cherry-picked quoting or context stripping) that require richer contextual checks. These are appropriate trade-offs for an audit-style evaluation, but they motivate follow-on studies with complementary human-centered and context-sensitive validation protocols.

7.2 Safety and misuse considerations

Systems optimized for persuasive, evidence-backed argumentation can be dual-use. Even when designed for accountability, modular pipelines that improve rhetorical quality and citation hygiene could be adapted for manipulation at scale (e.g., coordinated influence operations, astroturfing, or microtargeted persuasion), especially if paired with personalization and distribution infrastructure [5, 28]. We therefore include a dedicated Adverse Impacts statement in the paper's Endmatter describing plausible misuse modes, anticipated affected groups, and mitigations (e.g., access controls, logging/auditability, and deployment constraints) appropriate to this capability class.

7.3 Future work

A first priority is human-subject evaluation of contestability as an interaction property rather than a static artifact property. We plan controlled studies in which participants (including domain experts and affected stakeholders) attempt to (i) locate supporting evidence for a contested sentence, (ii) challenge a warrant or inference step, and (iii) request or compare alternative perspective nodes. Primary outcomes should include time-to-challenge, challenge success rates, perceived procedural fairness, and the degree to which the system supports actionable revision pathways (e.g., retracting a claim, swapping evidence, or surfacing counter-arguments) rather than merely producing longer explanations.

A second priority is extending FOAM with optimization and training methods while preserving contestability constraints. Our preliminary results in iterative preference learning for debate suggest that tactic selection and evidence integration can be improved substantially, but also reveal failure modes (e.g., "phantom critic" contamination and degraded interactive cross-examination under naïve retry-with-feedback regimes) that matter directly for accountable deliberation systems. Future work should explore (i) multi-judge and human-calibrated optimization targets, (ii) training objectives that explicitly reward faithful warrant-evidence alignment (not only persuasiveness), and (iii) contestation-aware curricula that treat interactive questioning and rebuttal as first-class skills rather than afterthoughts.

8 CONCLUSION

High-stakes deployments of LLM-based systems demand more than *transparent-seeming* narratives; they require explanations that can be *challenged, audited, and revised*. Recent evidence suggests that post-hoc "reasoning traces" are often not a reliable proxy for what drives model behavior: when a prompt-injected hint changes a model's answer, state-of-the-art reasoning models reveal that hint in their chain-of-thought only about **25–39%** of the time, indicating substantial unfaithfulness of verbalized rationales to causal drivers of outputs [7]. This paper contributes (1) **FOAM**, a pluralistic deliberation architecture for explainability-and-contestability-by-design; (2) an **inspectable provenance mechanism** that makes sentence-level claims traceable to source spans and contestable at the level stakeholders actually dispute; and (3) an **audit-style empirical evaluation** in evidence-grounded policy debate generation. In a double-blind tournament of 66 cases, the FOAM-based system achieves higher overall scores than expert-human and zero-shot baselines (Table 1) and dramatically higher perfect evidence validation rates (Table 2), demonstrating that accountable generation can be simultaneously *high-quality* and *verifiable*.

For the FAccT community, the central implication is a practical shift from explanation-as-disclosure to **contestable explanations**: outputs whose *claims, warrants, and evidence links* are explicit, inspectable, and designed to invite targeted challenge (e.g., disputing a cited sentence, contesting a warrant, or requesting an alternative perspective node). This orientation is consistent with due-process motivations for a meaningful right to contest consequential automated decisions [17]. More broadly, FOAM reframes accountability as a *system property* produced by structured mediation among differentiated perspectives, rather than as a post-hoc narrative appended to a monolithic model. Where governance requires reason-giving that can withstand scrutiny, pluralistic deliberation plus verifiable provenance offers a concrete design pattern for building AI systems whose decisions can be examined, contested, and improved without relying on "black-box" rationalizations.

GENERATIVE AI USAGE STATEMENT

This research investigates the use of large language models (LLMs) within a structured multi-agent deliberation framework. The FOAM system described in this paper uses LLMs as components within the deliberation pipeline. The paper text itself was drafted by human authors with AI assistance limited to copy-editing and formatting suggestions. All substantive claims, experimental design, and analysis reflect human judgment and interpretation.

ETHICAL CONSIDERATIONS

This work develops AI systems with persuasive capabilities, which raises dual-use concerns. We address these in Section 7 and Section 6, discussing safeguards including transparency requirements, evidence provenance constraints, and the deliberate choice to evaluate in a domain (competitive debate) with established norms for scrutinizing persuasive claims. The evaluation involved no human subjects; all baselines were drawn from publicly available debate materials or generated outputs.

REFERENCES

[1] Julius Adebayo, Michael Muelly, Harold Abelson, and Been Kim. 2022. Post hoc explanations may be ineffective for detecting unknown spurious correlation. In *International Conference on Learning Representations*.

[2] Kars Alfrink, Ianus Keller, Gerd Kortuem, and Neelke Doorn. 2023. Contestable AI by design: Towards a framework. *Minds and Machines* 33, 4 (2023), 613–639.

[3] Trevor JM Bench-Capon and Paul E Dunne. 2009. Argumentation in artificial intelligence. *Artificial intelligence* 171, 10-15 (2009), 619–641.

[4] Jonah Brown-Cohen, Geoffrey Irving, and Georgios Piliouras. 2024. Scalable AI safety via doubly-efficient debate. *arXiv preprint* (2024).

[5] Miles Brundage et al. 2018. The malicious use of artificial intelligence: Forecasting, prevention, and mitigation. *arXiv preprint arXiv:1802.07228* (2018).

[6] Guiming Chen et al. 2024. Humans or LLMs as the judge? A study on judgement bias. *arXiv preprint* (2024).

[7] Yanda Chen et al. 2025. Reasoning Models Don't Always Say What They Think. *arXiv preprint* (2025).

[8] John Dewey. 1927. *The public and its problems*. Holt.

[9] Finale Doshi-Velez and Been Kim. 2017. Towards a rigorous science of interpretable machine learning. In *ICML Workshop on Human Interpretability in Machine Learning*.

[10] Yilun Du, Shuang Li, Antonio Torralba, Joshua B Tenenbaum, and Igor Mordatch. 2023. Improving factuality and reasoning in language models through multiagent debate. In *arXiv preprint arXiv:2305.14325*.

[11] European Commission High-Level Expert Group on AI. 2019. Ethics Guidelines for Trustworthy AI.

[12] European Union. 2016. Regulation (EU) 2016/679 of the European Parliament and of the Council (General Data Protection Regulation).

[13] Donna Haraway. 1988. Situated knowledges: The science question in feminism and the privilege of partial perspective. *Feminist studies* 14, 3 (1988), 575–599.

[14] Geoffrey Irving, Paul Christiano, and Dario Amodei. 2018. AI safety via debate. In *arXiv preprint arXiv:1805.00899*.

[15] Alon Jacovi and Yoav Goldberg. 2020. Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness?. In *Proceedings of ACL*. 4198–4205.

[16] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Yejin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *Comput. Surveys* 55, 12 (2023), 1–38.

[17] Margot E Kaminski and Jennifer M Urban. 2021. The right to contest AI. *Columbia Law Review* 121, 7 (2021), 1957–2048.

[18] Atoosa Kasirzadeh. 2024. Plurality of value pluralism and AI value alignment. In *Pluralistic Alignment Workshop at NeurIPS 2024*. https://openreview.net/forum?id=AOokh1UYLH

[19] Atoosa Kasirzadeh and Iason Gabriel. 2023. In Conversation with Artificial Intelligence: Aligning Language Models with Human Values. *Philosophy & Technology* 36, 2 (2023), 1–23.

[20] Joshua A Kroll, Joanna Huey, Solon Barocas, Edward W Felten, Joel R Reidenberg, David G Robinson, and Harlan Yu. 2017. Accountable algorithms. *University of Pennsylvania Law Review* 165 (2017), 633.

[21] Khoa Lam et al. 2024. Assurance audits for AI systems. In *Proceedings of FAccT*.

[22] Timothy Lebo, Satya Sahoo, and Deborah McGuinness. 2013. PROV-O: The PROV ontology. W3C Recommendation.

[23] Zachary C Lipton. 2018. The mythos of model interpretability. *Queue* 16, 3 (2018), 31–57.

[24] Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence* 267 (2019), 1–38.

[25] Iyad Rahwan. 2018. Society-in-the-loop: Programming the algorithmic social contract. *Ethics and Information Technology* 20, 1 (2018), 5–14.

[26] Inioluwa Deborah Raji, Andrew Smart, Rebecca N White, Margaret Mitchell, Timnit Gebru, Ben Hutchinson, Jamila Smith-Loud, Daniel Theron, and Parker Barnes. 2020. Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. In *Proceedings of FAccT*. 33–44.

[27] Allen Roush et al. 2024. OpenDebateEvidence: A massive-scale dataset for argument mining and summarization. *arXiv preprint* (2024).

[28] Allen Roush et al. 2025. A superpersuasive autonomous policy debating system. *arXiv preprint* (2025).

[29] Cynthia Rudin. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature machine intelligence* 1, 5 (2019), 206–215.

[30] Noam Slonim et al. 2021. An autonomous debating system. *Nature* 591, 7850 (2021), 379–384.

[31] Alfred C Snider. 2008. *Code of the debater: Introduction to policy debating.* IDEA Press Books.

[32] Aman Singh Thakur et al. 2024. Judging the judges: Evaluating alignment and vulnerabilities in LLMs-as-judges. *arXiv preprint* (2024).

[33] Francesca Toni. 2014. A tutorial on assumption-based argumentation. *Argument & Computation* 5, 1 (2014), 89–117.

[34] Stephen E Toulmin. 1958. *The uses of argument.* Cambridge University Press.

[35] Alexandros Vassiliades, Nick Bassiliades, and Theodore Patkos. 2021. Argumentation and explainable artificial intelligence: A survey. *Knowledge Engineering Review* 36 (2021).

[36] Sandra Wachter, Brent Mittelstadt, and Luciano Floridi. 2017. Why a right to explanation of automated decision-making does not exist in the General Data Protection Regulation. *International Data Privacy Law* 7, 2 (2017), 76–99.

[37] Lianmin Zheng et al. 2023. Judging LLM-as-a-judge with MT-bench and Chatbot Arena. In *NeurIPS*.