

1      **Framework for Openly Augmented Mediation (FOAM): A Pluralistic**  
2      **Architecture for Explainable and Contestable AI**  
3

4      **ANONYMOUS AUTHOR(S)**  
5

6      High-stakes AI systems increasingly mediate access to credit, healthcare, and public benefits, yet affected parties often  
7      cannot see why a decision was made or meaningfully contest it. Even post hoc review of chain-of-thought traces from  
8      individual models can be incomplete or strategically misleading, thereby limiting accountability. We propose FOAM  
9      (Framework for Openly Augmented Mediation), a pluralistic architecture that treats explanation as a *deliberative*  
10     process rather than post-hoc narration. FOAM instantiates differentiated agents with explicit value commitments,  
11     structures their interaction through cross-examination and rebuttal protocols, and outputs not just a recommendation  
12     but a *contestable record intended to support downstream review*: claims linked to sentence-level evidence provenance,  
13     surviving objections, and explicit points of disagreement. We evaluate FOAM in evidence-grounded policy debate  
14     generation, a domain where arguments must withstand adversarial scrutiny. In a double-blind tournament of 66 cases,  
15     FOAM outperforms human-expert and zero-shot baselines on overall quality (81.7 vs. 70.1 vs. 50.6) while achieving  
16     dramatically higher evidence verifiability (76.2% perfect validation vs. 8.7% and 0%). These results demonstrate that  
17     pluralistic deliberation can produce outputs that are simultaneously persuasive *and* auditable, a necessary condition  
18     for contestable AI by design.  
19

20     Additional Key Words and Phrases: Algorithmic accountability; Contestable AI; Explainable AI (XAI); Multi-agent  
21     deliberation; Evidence provenance  
22

23     **ACM Reference Format:**  
24

25     Anonymous Author(s). 2026. Framework for Openly Augmented Mediation (FOAM): A Pluralistic Architecture for  
26     Explainable and Contestable AI. 1, 1 (January 2026), 17 pages. <https://doi.org/10.1145/nnnnnnn.nnnnnnn>  
27

28     **1 INTRODUCTION**  
29

30     **1.1 Accountability gap in high-stakes AI**  
31

32     AI systems are now routinely embedded in high-stakes decision workflows—healthcare triage and documentation [29], hiring and workplace management [30], credit and insurance [21], public benefits [9], and criminal-legal risk assessments [2]. 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.

33     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) [28, 44]. **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,

---

34     2026. Manuscript submitted to ACM  
35

36     Manuscript submitted to ACM  
37

(b) comprehensibility for affected parties, and (c) actionable mechanisms for challenge and revision [1]. A 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” [6]. 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 [41].

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:

- 105 (1) **Framework:** we provide a unified account of explainability *and* contestability as a single design  
106 target, arguing that they should be treated jointly and realized through pluralistic mediation rather  
107 than monolithic self-report.
- 109 (2) **Architecture and mechanisms:** we formalize FOAM as an implementable blueprint—agents, pro-  
110 tocols, and revision operators—paired with provenance-oriented design choices that make challenges  
111 actionable (e.g., grounding claims in checkable evidence rather than free-form summarization).
- 113 (3) **Empirical validation:** we report results from an evaluation of pluralistic debate generation in a  
114 double-blind tournament of **66 policy debate cases**, where our structured multi-agent system  
115 achieved an overall score of **81.7** compared to **70.1** for human experts and **50.6** for zero-shot AI,  
116 while also achieving **76.2%** perfect evidence validation compared to **8.7%** for human experts and  
117 **0%** for unstructured AI—demonstrating that pluralistic architectures can produce outputs that are  
118 simultaneously more persuasive *and* more verifiable in an adversarial, evidence-sensitive setting.  
119

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

## 126 2 ACCOUNTABILITY REQUIREMENTS AND RELATED WORK

### 127 2.1 Explainability requirements beyond transparency

129 Contemporary calls for “explainable AI” often conflate **transparency** (exposing internal mechanisms) with  
130 **explanation** (providing reasons meaningful for a particular audience). Lipton argues that interpretability is  
131 not a single property and that many “explanations” function as *post-hoc rationalizations* whose relationship  
132 to actual model behavior is ambiguous [27]. Doshi-Velez & Kim emphasize that interpretability claims  
133 must be made relative to **use context**—including the user’s expertise and stakes—because what counts as  
134 satisfactory differs across settings [7], yet few XAI papers explicitly address end-user perspectives [12]. In  
135 high-stakes domains, this motivates either inherently interpretable models or explanation mechanisms that  
136 achieve *reliability and auditability* rather than superficial plausibility [35].

138 For accountability, explanations must be **diagnostically useful** and **robust to strategic manipulation**.  
139 The NLP interpretability literature distinguishes *plausibility* (does an explanation look reasonable?) from  
140 *faithfulness* (does it track the true basis of the output?), arguing that faithful explanations require designs  
141 that go beyond “nice-sounding” rationales [15]. Explainability requirements should thus be stated in terms  
142 of **checkability**: tracing claims to concrete support and isolating points of disagreement [15, 28].  
143

### 147 2.2 Contestability as a system property

148 Explainability alone does not guarantee meaningful challenge; contestability is best treated as a **system-  
149 level governance property**. Alfrink et al. frame “contestable AI by design” as building systems to  
150 *support* contestation—through traceability, structured justification, and pathways for challenge—rather than  
151 treating contestation as an external process [1]. Legal scholarship similarly emphasizes that decision-subjects  
152 need procedures to *question, rebut, and obtain redress* [22]. This matters because the scope of a “right to  
153 explanation” under GDPR is contested [43].  
154

157 Operationally, contestability implies three requirements: (1) **visibility** that an AI-assisted decision  
158 occurred; (2) **comprehensibility** of stated grounds; and (3) **actionability**—a pathway to present coun-  
159 terevidence and obtain revision [1, 22]. The EU’s Trustworthy AI guidance treats accountability as including  
160 mechanisms for redress and capacity to challenge outcomes [10]. These sources motivate a design target:  
161 **contestability must be an end-to-end workflow** linking reasons to evidence, rather than a static  
162 artifact [32].  
163

164165166167

### 168 2.3 Pluralistic and deliberative approaches to accountability

169 In high-stakes settings, disagreement is often normative (“which values should dominate?”) not merely  
170 empirical. Feminist epistemology argues that knowledge claims are situated and that “view from nowhere”  
171 objectivity can mask whose assumptions are operationalized [13]. For AI accountability, this motivates an  
172 architectural stance: systems should make **value trade-offs explicit** and preserve dissenting considerations  
173 in contestable form [28].  
174

175 Recent work emphasizes that “alignment” is underdetermined when stakeholders disagree about objectives  
176 and risks. Kasirzadeh distinguishes alignment approaches that presume a single value target from those  
177 treating plural values as first-class constraints [19]. “Society-in-the-loop” framings argue that algorithmic  
178 systems require institutionalized interfaces for dispute and revision [31]. These perspectives justify **plural-  
179 istic explanation** as a governance mechanism helping stakeholders identify where reasoning depends on  
180 contestable assumptions.  
181

182183184185186

### 187 2.4 Multi-agent deliberation and debate in AI

188 A technical pathway to operationalizing pluralism is **structured multi-agent deliberation**. Constitutional  
189 AI introduced principle-guided self-critique [4], and Plurals demonstrates that diverse-persona LLM  
190 deliberation produces preferred outputs [3]. In AI safety, “debate” was proposed as a scalable oversight  
191 mechanism where adversarial argumentation surfaces flaws a single system might hide [14]. Multi-agent  
192 debate among LLMs has been reported to improve factuality [8], and recent work demonstrates that debate  
193 with more persuasive models helps non-expert judges achieve higher accuracy on difficult questions [20].  
194 However, most results are evaluated in terms of accuracy; they do not guarantee that justifications are  
195 **auditable** or that third parties can contest specific premises [15, 35].  
196

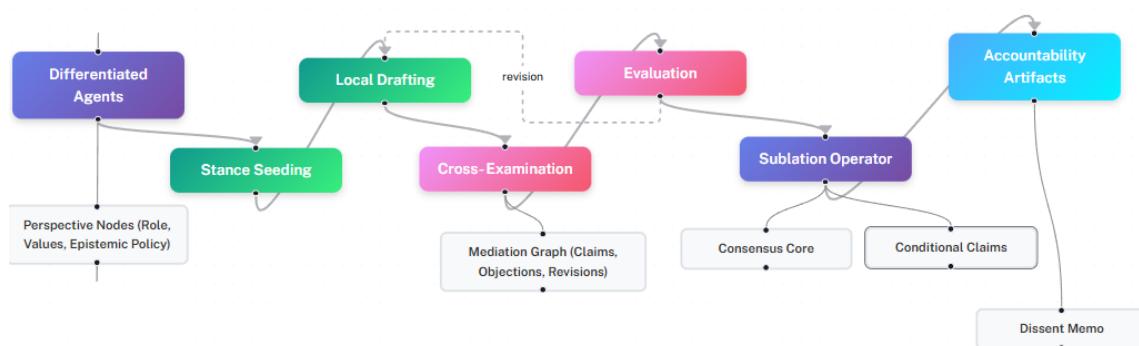
197 Computational argumentation provides complementary foundations via explicit representations of **claims**,  
198 **warrants, attacks, and normative priorities**. Toulmin’s model analyzes argument structure in terms of  
199 claims supported by warrants and backing [40]. Surveys connecting argumentation and XAI argue these  
200 representations support explanation as a structured object of inquiry—stakeholders can contest particular  
201 premises and observe how conclusions change [42]. This motivates the claim that a *contestable* AI system  
202 should produce a **dispute-ready argumentative record**: reasons decomposed into contestable units,  
203 linked to supporting materials, and amenable to revision [22, 25, 42].  
204

205206207208

209 **3 FOAM APPROACH: PLURALISTIC ARCHITECTURE FOR EXPLAINABILITY AND**  
 210 **CONTESTABILITY**

211 **3.1 Design goals and accountability threat model**

212 Building on Section 2, we treat *explainability* and *contestability* as properties of an **epistemic process**,  
 213 not a post-hoc narrative. We introduce **FOAM (Framework for Openly Augmented Mediation)**: a  
 214 pluralistic, multi-agent architecture producing an answer *plus* a structured record of how it was stress-tested  
 215 and synthesized. FOAM is organized around three primitives: (i) *differentiated agents* parameterized by  
 216 explicit stance data structures, (ii) *deliberative protocols* forcing critique and revision, and (iii) *sublation*  
 217 operators that synthesize without erasing disagreement. Figure 1 provides a system overview.  
 218



236 Fig. 1. FOAM system architecture. Differentiated agents with explicit perspective nodes engage in deliberative protocols  
 237 producing accountability artifacts including a consensus core, conditional claims, and dissent memo.

240 Our threat model assumes base generative models can (a) produce fluent but false claims (“hallucination”)  
 241 [16], (b) rationalize decisions after the fact [41], (c) collapse multiple perspectives into a dominant  
 242 frame, and (d) bury value tradeoffs inside unstructured prose. FOAM’s core design makes *points of potential*  
 243 *failure* explicitly addressable: disagreements are surfaced, objections are first-class objects, and synthesis  
 244 preserves traceability from contested premises to recommendations.

247 **3.2 Differentiated agents via explicit perspective representation**

248 FOAM instantiates agents each assigned an explicit *Perspective Node* encoding *who the agent is epistemically*—  
 249 domain role, value priorities, and reasoning schema. This implements “situated” explanation in an auditable  
 250 way: the system discloses positions and enables critique of *perspective selection* itself [13]. Perspective nodes  
 251 are operational constraints shaping what evidence is legitimate, which impacts are foregrounded, and which  
 252 argument schemes are preferred.

253 A perspective node has three components: (1) **role** (e.g., regulator, clinician, community advocate), (2)  
 254 **normative weighting** (e.g., safety vs autonomy vs equity), and (3) **epistemic policy** (e.g., acceptable  
 255 support standards). During deliberation, FOAM enforces *stance coherence*: if generated warrants contradict  
 256 the declared stance, the system flags the inconsistency.

261 Perspective nodes enable **second-order contestation**: stakeholders can dispute not only conclusions, but  
 262 the *legitimacy of the perspective configuration* (e.g., “Why is utilitarian cost-effectiveness in scope here?”).  
 263 FOAM makes the stance set an explicit input and target for governance [18]. This means FOAM can be  
 264 rerun with added perspectives, reweighted priorities, or altered evidentiary rules, producing *comparative,*  
 265 *contestable* outcomes.  
 266  
 267  
 268

### 269 3.3 Deliberative protocol: dialectical refinement and mediation trace

270 FOAM’s deliberation is a **mediation loop**: (1) *seeding* (instantiate agents + perspectives), (2) *local*  
 271 *drafting* (independent proposals), (3) *cross-examination* (structured objections), (4) *evaluation* (scoring  
 272 draft–objection pairs), and (5) *revision + synthesis*. The accountability point: **deliberation guarantees**  
 273 **structured opportunities to find and localize error**, and records what happened when error was  
 274 raised.  
 275  
 276

277 Cross-examination produces a **mediation graph**: a trace linking *which agent* made *which claim*, what  
 278 objections were raised, how claims were revised, and which survived. This is the audit primitive: stakeholders  
 279 can point to *the specific node* where they disagree. The trace can be expressed using standard provenance  
 280 representations (e.g., PROV-O) [24].  
 281  
 282

### 283 3.4 Sublation: synthesis without erasure

284 After critique, FOAM applies a **sublation operator**: synthesis preserving what is valuable in competing  
 285 positions while retaining unresolved tensions. Synthesis is disallowed from silently discarding material  
 286 objections or collapsing incompatible frames into unmarked compromise. Sublation emits three artifacts: a  
 287 **consensus core** (claims surviving cross-stance critique), **conditional claims** (branching on unresolved  
 288 priorities), and a **dissent memo** (recording conflicts and contested premises).  
 289  
 290

### 291 3.5 Inspectable argument structure: Toulmin decomposition and typed syllogisms

292 To make contestation actionable, FOAM constrains outputs into **inspectable argument structure**. We  
 293 adopt Toulmin-style decomposition—claim, grounds, warrant, backing, qualifier, rebuttal—because it maps  
 294 to “what can be challenged”: stakeholders can contest evidence, the inferential link, scope conditions, or  
 295 missing counterevidence [40, 42].  
 296  
 297

298 FOAM employs **typed syllogisms**—argument templates enforcing completeness (e.g., Advantage =  
 299 Uniqueness + Link + Impact). These function as contestability scaffolds: if a stakeholder disputes the  
 300 conclusion, the system points to the *specific weak component*, and the mediation graph shows whether it was  
 301 raised in critique [38].  
 302  
 303

304 Template tree traversal operationalizes structural contestability. At each branch point, the system records  
 305 which template was selected (e.g., “traditional 1AC” vs. “kritik”), what resource allocation was applied, and  
 306 whether novel templates were generated. Stakeholders can dispute not only *what* claims were made, but *why*  
 307 *the structure took this form*. Unlike chain-of-thought where reasoning and response are interwoven, template  
 308 traversal is a discrete prior step serving as foundational infrastructure to drafting.  
 309  
 310

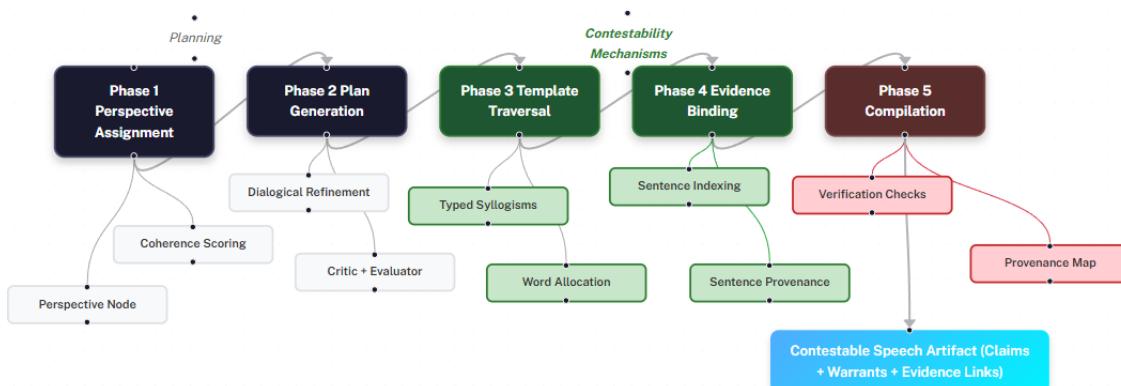
313 **4 CASE STUDY SYSTEM: EVIDENCE-GROUNDED POLICY DEBATE GENERATION**

314 **4.1 Why policy debate is an accountability crucible**

315 We instantiate FOAM in a domain where *contestability is native to the task*: American competitive policy  
 316 debate. Policy debate is a two-team adversarial format in which teams argue for and against a policy proposal  
 317 under strict procedural constraints [38]. In this ecosystem, argument quality is not evaluated purely as  
 318 rhetorical fluency; instead, the activity is structured around *traceable evidentiary support* and explicit clash,  
 319 so claims can be challenged in real time and revisited across subsequent speeches. Critically, policy debate  
 320 operationalizes “grounding” through an established evidence artifact: the *debate card*. A card typically  
 321 includes (i) a short biased summary intended to support a specific argumentative function, (ii) a full citation,  
 322 and (iii) verbatim quoted source text, often with token-level highlighting that marks precisely what will  
 323 be read into the round. Competitive success is strongly coupled to evidence quality and its deployment,  
 324 creating an evaluation environment where provenance and verifiability are not optional.  
 325

326 **4.2 Pipeline overview**

327 Figure 2 summarizes our **five-phase pipeline** for generating an evidence-grounded constructive speech (the  
 328 1AC, in our evaluation setting). Phases 1–3 produce an inspectable argumentative plan in typed components  
 329 (perspective assignment → strategic plan → template traversal), Phase 4 binds each argumentative component  
 330 to *verbatim evidence at sentence granularity* (sentence-level provenance), and Phase 5 compiles and verifies  
 331 the result (structural conformance, evidence/claim alignment, and perspective consistency). The key design  
 332 principle is to keep the model in a role where it can be audited: rather than “write a persuasive case and  
 333 cite sources,” the system decomposes “case construction” into a sequence of constrained decisions that leave  
 334 a machine-checkable trail.



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

### 365    4.3 Phases 1–3: perspective assignment, planning, and template traversal

366  
 367    Phases 1–3 produce an inspectable argumentative plan through three contestability-relevant operations.  
 368    In **Phase 1**, the system assigns an explicit perspective node (Section 3.2), making the evaluative frame a  
 369    first-class auditable choice. In **Phase 2**, a dialectical refinement loop stress-tests the strategic plan: a Critic  
 370    agent issues typed objections (logical gap, missing evidence, value conflict, scope overreach), an Evaluator  
 371    scores each objection’s materiality, and the Proposer revises or rebuts. This cycle iterates at least three  
 372    times, and *all objections—including dismissed ones—remain in the mediation graph*, enabling downstream  
 373    reviewers to inspect whether a weakness was raised and why the response was deemed adequate.

374  
 375    In **Phase 3**, template tree traversal expands the plan into a typed syllogism scaffold (e.g., Advantage  
 376    = Uniqueness + Link + Impact). At each branch point, the system records which template was selected,  
 377    what word allocation was applied (e.g., 30% impact, 40% link), and whether novel templates were generated.  
 378    This trace enables a distinct class of challenges: stakeholders can dispute not only *what* claims were made,  
 379    but *why the argumentative structure took this form rather than another*—for instance, contesting that a  
 380    utilitarian impact calculus was chosen when the underlying values favor a rights-based framing.  
 381  
 382

### 383 384    4.4 Phase 4: sentence-level provenance

385  
 386    **Motivation.** Retrieval-augmented generation can reduce hallucinations [26, 36], but it does not eliminate a  
 387    central accountability failure mode: models may still produce claims that are *unsupported by, in conflict with,*  
 388    or *misattributed to* retrieved text. Recent benchmarks explicitly document that, even under RAG setups,  
 389    LLM outputs can contain unsupported or contradictory content relative to the retrieved passages [11]. Phase  
 390    4 therefore implements a stronger constraint than “retrieve then paraphrase”: it forces the model to operate  
 391    over *sentence identifiers* rather than free-form rewriting of source material.

392    **Mechanism.** Phase 4 is a two-step procedure:

393  
 394    **Step (a): sentence indexing and retrieval.** The system queries (i) a debate-evidence store (implemented  
 395    in our current system as a vector database over a large set of debate “cards”) and (ii) any other preprocessed  
 396    sources permitted by the pipeline. Retrieved documents are segmented into sentences, each assigned a  
 397    stable index, and returned to the deliberation workspace as a set of candidates with identifiers of the form  
 398    (`document_id, sentence_id`) plus immutable citation metadata.

399  
 400    **Step (b): evidence selection and tagging.** The LLM is then prompted to (1) select which sentence  
 401    IDs support each argument slot created in Phase 3 and (2) generate only a short “tag” that states what the  
 402    selected evidence is being used to establish. Importantly, the model is not asked to restate the evidence;  
 403    the evidence content in the final speech is assembled from the retrieved sentences themselves. This design  
 404    eliminates an entire class of failure (fabricated quotations and invented citations) by construction: the model  
 405    can be wrong about *which* sentences to use, but it cannot invent sentences that are not in the retrieved set.

406  
 407    **Accountability and contestability properties.** Sentence-level provenance changes the contestation  
 408    workflow from “argue about what the model meant” to “inspect exactly what the model relied on.” A  
 409    stakeholder can challenge (i) *relevance* (“this sentence does not establish the warrant you claim”), (ii) *adequacy*  
 410    (“the evidence is too weak/out of context”), or (iii) *selection bias* (“you ignored stronger counterevidence  
 411    available in the same corpus”)—and each challenge targets a concrete object (a sentence ID and its parent  
 412    413

417 source). This is especially aligned with policy debate’s evidence norms, which already treat quoted and  
418 highlighted text as the unit of disputation under cross-examination.  
419

420 **4.5 Phase 5: compilation and verification checks**  
421

422 Phase 5 compiles the typed argument scaffold (Phase 3) and the evidence bindings (Phase 4) into a final  
423 speech artifact suitable for evaluation. Compilation preserves the provenance map: each substantive claim in  
424 the rendered speech remains traceable to one or more sentence IDs plus citation metadata. The system then  
425 runs verification checks that are directly tied to the accountability requirements:  
426

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

432 Figure 2 highlights where provenance is created (Phase 4) and where it is enforced (Phase 5).  
433

434 **5 EMPIRICAL EVALUATION**  
435

436 **5.1 Research questions**  
437

438 We evaluate FOAM’s accountable-generation claims using an *audit-style* design: we define explicit research  
439 questions, compare against salient baselines, and report both performance outcomes and traceability outcomes  
440 as first-class metrics. This approach aligns with established work on internal algorithmic auditing and  
441 emerging “assurance audit” perspectives, which emphasize that accountability requires not only outcome  
442 quality, but also artifacts and procedures that make decisions inspectable and challengeable [23, 32].  
443

444 We ask whether FOAM improves:

- 445 • **RQ1:** Quality/persuasiveness  
446 • **RQ2:** Evidence verifiability  
447 • **RQ3:** Whether gains are attributable to the accountability mechanisms rather than model strength  
448

449 **5.2 Experimental design and baselines**  
450

451 **Task selection.** We evaluate in evidence-grounded policy debate generation because it combines (i) long-  
452 horizon argumentative planning, (ii) adversarial robustness expectations (arguments must survive challenge),  
453 and (iii) strict evidentiary norms (claims are conventionally supported with citations). In computational  
454 argumentation, even highly resourced systems have historically relied on constrained debate settings and  
455 bespoke pipelines; the Project Debater line of work illustrates both the ambition of debate as a benchmark  
456 and the practical need to structure and constrain the task for reliable evaluation [37].  
457

458 **Debate artifact.** We focus on the **first affirmative constructive (1AC)** as the most demanding  
459 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  
460 tight length constraints while maintaining evidentiary support. This makes the 1AC a strong proxy for  
461 high-stakes accountable generation: arguments must be *comprehensible, internally coherent, and traceable to*  
462 *evidence* to be meaningfully contestable.  
463

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

- 470 471 (1) FOAM-based structured system** ( $n = 22$ ), generated via differentiated perspectives, iterative
- 472 dialectical refinement, typed syllogisms, and sentence-level provenance;**
- 473 474 (2) Human expert baseline** ( $n = 23$ ), sampled from expert-authored training materials from highly
- 475 competitive policy debate programs; and**
- 476 477 (3) Zero-shot AI baseline** ( $n = 21$ ), produced by frontier models (Gemini/Claude/ChatGPT/Grok)
- 478 using prompt engineering and web-research access but without debate-specific pluralistic architecture.**

**479 Baseline controls (zero-shot AI).** To reduce confounding from artifact format and resource constraints,

**480 we generated the zero-shot baseline using Claude 4.5 in research mode, GPT-5 in deep research mode,**

**481 SuperGrok Heavy, and Gemini 2.5 in research mode. We used a single standardized “mega-prompt” that**

**482 enforced the same 1AC conventions and constraints used by elite debate program materials and by our**

**483 FOAM case-building pipeline: 8 minutes of read-time target (1300–1700 words); debate formatting**

**484 (ALL-CAPS tags, short analytic warrants above evidence); a fixed advantage/solvency structure; explicit**

**485 impact calculus; and comparable evidence-density targets (3–7 cards per advantage; 2–5 in solvency).**

**486 The prompt also enforced a strict no-fabrication policy:** when reliable bibliographic details and quotations

**487 could not be produced, models were required to generate high-precision search strings and to mark uncertainty**

**488 as [EVIDENCE NEEDED]. When the interface supported browsing, web access was enabled to reduce**

**489 evidence-access confounds. Unlike FOAM, these baselines did not use multi-agent deliberation, typed**

**490 syllogisms enforcement, or sentence-level provenance binding; thus, baseline citations remained unconstrained**

**491 natural-language references and were evaluated under the same automated validation pipeline. We generated**

**492 one case per topic per condition and used outputs as-is (no manual editing beyond uniform formatting**

**493 normalization).**

**494 Evidence corpus for provenance.** FOAM’s evidence retrieval and validation leverage a structured

**495 debate-evidence corpus derived from OpenDebateEvidence, which (as released) contains 3.5M+ competitive**

**496 debate documents with metadata useful for downstream argument mining and citation [33]. Operationally,**

**497 our system queries a vector database of ~85,000 curated “cards” plus any newly processed sources, and the**

**498 generation pipeline preserves *sentence-level identifiers* so that downstream reviewers can trace claims to**

**499 exact supporting spans.**

### **500 501 5.3 Judging rubric and scoring**

**502 Tournament format and blinding.** All submissions were anonymized and assigned unique IDs (e.g.,

**503 Case\_001), and judging proceeded purely on content without revealing origin. Cases advanced through a**

**504 modified Swiss-style bracket with double elimination, and pairings were balanced by strategic approach**

**505 (e.g., traditional policy vs. kritik) to reduce “judge adaptation” artifacts. Ties within a narrow score**

**506 band triggered evidence validation as a tiebreaker, keeping accountability-relevant verifiability salient**

**507 in advancement decisions. All 66 cases were scored once under the rubric; Tables 1–2 report**

**508 aggregate statistics over the full set and do not depend on bracket advancement.**

**509 Rubric and judge.** Following established LLM-as-judge methodology [45], a Claude Opus 4 judge

**510 evaluated each case on five weighted dimensions:**

- 511 • Argumentation Strength (25%)**

- 521 • **Evidence Quality** (25%)
- 522 • **Strategic Coherence** (20%)
- 523 • **Innovation** (15%)
- 524 • **Competitive Viability** (15%)

525  
526 The rubric was designed to reward both argumentative competence and evidence-groundedness, while  
527 preserving enough structure for reproducibility.  
528

#### 529 530 **5.4 Evidence validation methodology**

531 **Why evidence validation is an accountability metric (not just “anti-hallucination”)**. In contestable  
532 systems, stakeholders must be able to *locate* and *evaluate* the grounds of a claim—especially where persuasive  
533 language can obscure weak or missing support. Audit frameworks similarly emphasize that assurance depends  
534 on traceable evidence artifacts rather than outcome plausibility alone [23, 32]. We therefore operationalize  
535 verifiability as a measurable property of each case’s citations.  
536

537 **Automated citation checks and categories.** Each citation was automatically checked against the  
538 referenced source (via URL or resolvable reference), and classified into one of four buckets: **exact match**,  
539 **partial match**, **paraphrase**, or **fabricated**. We summarize results primarily via **Perfect Validation**,  
540 a stringent metric that counts only **exact matches**—i.e., the cited claim can be located verbatim in the  
541 referenced source span. This is intentionally conservative: Perfect Validation corresponds to the strongest  
542 form of contestability, where an affected party can directly inspect the cited text without interpretive debate  
543 about semantic similarity.  
544

545 **How FOAM changes the validation problem.** FOAM’s sentence-level provenance changes citation  
546 validation from a semantic retrieval problem into a *pointer integrity* problem: the model is never asked to  
547 reproduce source text, but instead selects sentence indices from retrieved documents and attaches them to  
548 specific argument components. This design greatly reduces degrees of freedom for fabrication and enables  
549 deterministic re-checking of a case’s evidentiary backbone.  
550

#### 553 554 **5.5 Results**

555 **Main tournament outcomes.** Table 1 reports aggregate performance by source. The FOAM-based  
556 system achieved the highest overall score (**81.7**) relative to human experts (**70.1**) and zero-shot AI  
557 (**50.6**). The largest gap appears in **Evidence Quality** (**86.7** vs. **56.9** vs. **27.1**), consistent with the claim  
558 that provenance-constrained generation shifts the system from persuasive-but-unreliable outputs toward  
559 persuasive-and-grounded outputs.  
560

561 562 **Table 1. Tournament Results by Source**

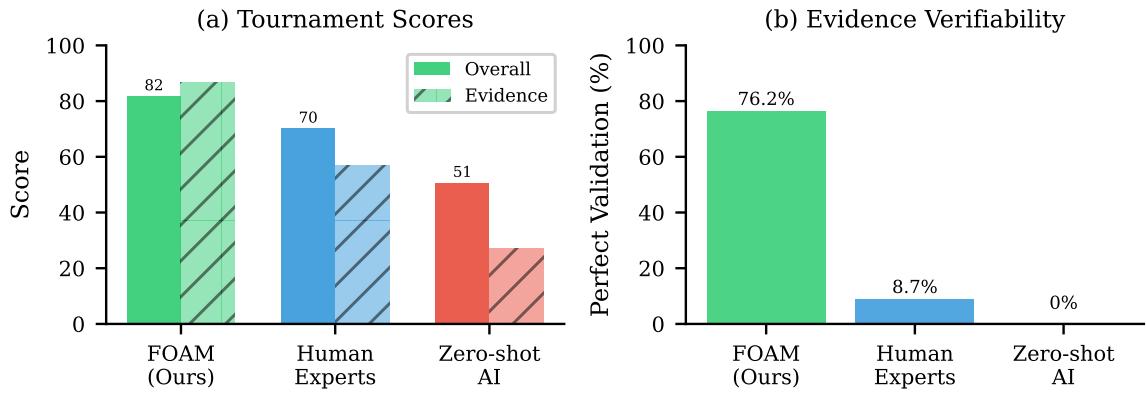
Metric	FOAM	Human Expert	Zero-shot AI
Overall Score	81.7	70.1	50.6
Evidence Quality	86.7	56.9	27.1

569 570 **Evidence validation and verifiability.** Table 2 reports Perfect Validation rates. FOAM achieved  
571 572 **76.2%** Perfect Validation, compared to **8.7%** for the human expert baseline and **0%** for zero-shot AI. This

573 is the central accountability result: the FOAM pipeline does not merely produce arguments that a judge  
 574 model rates as “good,” but produces arguments whose evidentiary support can be mechanically verified at  
 575 scale.  
 576

Table 2. Perfect Validation Rates

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



600 Fig. 3. Tournament results comparing FOAM, human expert baselines, and zero-shot AI. (a) Overall and Evidence Quality  
 601 scores. (b) Perfect Validation rates—the percentage of citations that exactly match source text. FOAM achieves 76.2%  
 602 perfect validation vs. 8.7% for human experts and 0% for zero-shot AI.  
 603

604 **Interpreting what is doing the work.** Two mechanisms plausibly drive the observed gap: (i) **pluralistic**  
 605 **deliberation** (multi-perspective critique and refinement) improves strategic coherence and argument  
 606 coverage, while (ii) **sentence-level provenance** directly improves evidence integrity and sharply limits  
 607 fabrication opportunities. Several high-scoring FOAM cases achieved perfect validation (fidelity = 1.0),  
 608 indicating that high persuasive quality and high verifiability can co-occur under the FOAM constraint  
 609 regime.  
 610

## 6 IMPLICATIONS FOR ACCOUNTABLE AI SYSTEMS

614 FOAM reframes explanation as a contestable record rather than a post-hoc narrative. Instead of producing  
 615 a single rationale, the system outputs (i) an auditable argument structure (claims, warrants, rebuttals), (ii)  
 616 explicit perspective configurations, and (iii) sentence-level provenance linking each substantive claim to a  
 617 checkable source span. This shifts accountability from “did the explanation sound plausible?” to “which  
 618 premises and evidence does the output depend on, and where can a challenge be lodged?”  
 619

620 Operationally, FOAM supports contestation at three levels [1]: (1) **evidence disputes** (a cited sentence  
 621 does not support the tagged claim; missing counterevidence), (2) **inferential disputes** (the warrant con-  
 622 nnecting evidence to conclusion is invalid or incomplete), and (3) **normative disputes** (the perspective/value  
 623 Manuscript submitted to ACM

625 configuration is illegitimate or incomplete for the context). Because these objects are explicit, a reviewer can  
626 localize disagreement to specific nodes and request revision without reopening the entire output as free-form  
627 prose.  
628

629 Institutionally, the resulting artifact functions as an auditable dossier that can support downstream  
630 review within existing governance workflows (internal review, incident response, assurance audits); dispute  
631 resolution itself requires institutional process beyond what the technical system provides. The technical  
632 contribution is not replacing due process, but supplying the structured, traceable materials that make  
633 procedural review feasible at scale.  
634

## 635 636 637 638 639 7 LIMITATIONS AND FUTURE WORK

### 640 641 7.1 Methodological limitations and validity threats

642 First, our primary outcome measure relies on an automated judge (Claude Opus 4) to score debate  
643 artifacts under a fixed rubric. While LLM-as-judge evaluation is increasingly standard at scale, it is  
644 known to exhibit systematic biases (e.g., position effects, verbosity/style sensitivity, and self-enhancement  
645 tendencies) and may be vulnerable to prompt- or framing-based perturbations that shift preferences without  
646 corresponding semantic differences [5, 39, 45]. We reduce—but do not eliminate—these threats via double-  
647 blinding, standardized prompts, and by pairing judge scores with an independent evidence-validation audit.  
648 Nevertheless, the reported tournament results should be interpreted as descriptive for this evaluation setup,  
649 and future replications should triangulate across multiple judge models and human adjudication.  
650

651 Second, our system’s accountability guarantees are conditioned on the properties of the underlying  
652 evidence substrate. Sentence-level provenance constrains the model to point to specific source sentences  
653 rather than inventing citations, but it does not ensure that the retrieved evidence is complete, representative,  
654 or up to date. Coverage gaps, topical skew, and retrieval errors can shape which arguments are discoverable,  
655 and can yield outputs that are “well-cited” yet misleading due to selection effects, over-aggregation, or  
656 missing context [34]. These concerns are not unique to debate generation: any contestability mechanism built  
657 on curated corpora inherits the corpus’ blind spots. Accordingly, FOAM should be viewed as an approach  
658 to making claims auditable and challengeable—not as a guarantee that the selected evidence is normatively  
659 “best” or epistemically sufficient.  
660

661 Third, our evaluation scope is intentionally narrow and therefore limits external validity. We benchmark a  
662 specialized argumentative domain (policy debate) and a bounded artifact type (constructive case generation),  
663 and we do not yet measure downstream stakeholder contestation behaviors (e.g., whether affected parties  
664 can efficiently detect, understand, and successfully challenge specific warrants or citations). Additionally, our  
665 “perfect validation” metric is strict by design: it favors verbatim traceability and can under-credit faithful  
666 paraphrase or correct claims supported by multiple dispersed sentences. Conversely, the metric may fail to  
667 detect other fidelity failures (e.g., cherry-picked quoting or context stripping) that require richer contextual  
668 checks. These are appropriate trade-offs for an audit-style evaluation, but they motivate follow-on studies  
669 with complementary human-centered and context-sensitive validation protocols.  
670

## 677 7.2 Safety and misuse considerations

678 Systems optimized for persuasive argumentation can be dual-use; we address misuse risks, affected groups,  
 679 and mitigations in the Adverse Impacts statement (Endmatter).

## 681 682 7.3 Future work

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

690 A second priority is extending FOAM with optimization and training methods while preserving con-  
 691 testability constraints. Preliminary results in iterative preference learning suggest that tactic selection and  
 692 evidence integration can be improved, but also reveal failure modes that matter for accountable deliberation.  
 693 Future work should explore training objectives that explicitly reward faithful warrant-evidence alignment  
 694 (not only persuasiveness) and contestation-aware curricula.

## 695 696 8 CONCLUSION

700 High-stakes deployments of LLM-based systems demand more than *transparent-seeming* narratives; they  
 701 require explanations that can be *challenged, audited, and revised*. Recent evidence suggests that post-hoc  
 702 “reasoning traces” are often not a reliable proxy for what drives model behavior: when a prompt-injected hint  
 703 changes a model’s answer, state-of-the-art reasoning models reveal that hint in their chain-of-thought only  
 704 about 25–39% of the time, indicating substantial unfaithfulness of verbalized rationales to causal drivers of  
 705 outputs [6]. This paper contributes (1) **FOAM**, a pluralistic deliberation architecture for explainability-and-  
 706 contestability-by-design; (2) an **inspectable provenance mechanism** that makes sentence-level claims  
 707 traceable to source spans and contestable at the level stakeholders actually dispute; and (3) an **audit-style**  
 708 **empirical evaluation** in evidence-grounded policy debate generation. In a double-blind tournament of 66  
 709 cases, the FOAM-based system achieves higher overall scores than expert-human and zero-shot baselines  
 710 (Table 1) and dramatically higher perfect evidence validation rates (Table 2), demonstrating that accountable  
 711 generation can be simultaneously *high-quality* and *verifiable*.

712 For the FAccT community, the central implication is a practical shift from explanation-as-disclosure to  
 713 **contestable explanations**: outputs whose *claims, warrants, and evidence links* are explicit, inspectable,  
 714 and designed to invite targeted challenge (e.g., disputing a cited sentence, contesting a warrant, or requesting  
 715 an alternative perspective node). This orientation is consistent with due-process motivations for a meaningful  
 716 right to contest consequential automated decisions [17]. Where governance requires reason-giving that can  
 717 withstand scrutiny, pluralistic deliberation plus verifiable provenance offers a concrete design pattern for  
 718 building AI systems whose decisions can be examined, contested, and improved without relying on “black-box”  
 719 rationalizations.

**729 ENDMATTER****730 Generative AI Usage Statement**

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

**738 Ethical Considerations**

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

**746 Adverse Impacts Statement**

748 Systems that generate persuasive, evidence-grounded arguments could be misused for misinformation,  
749 manipulation, or to overwhelm human review capacity. Affected groups include decision-subjects in high-  
750 stakes domains and information consumers generally. We mitigate these risks through: (1) provenance  
751 requirements that make claims auditable; (2) evaluation in a domain with adversarial scrutiny norms; (3)  
752 architectural transparency (the deliberation trace is inspectable). Deployment in sensitive domains should  
753 include access controls, logging, human oversight, and institutional review processes.

**755 REFERENCES**

- 757 [1] Kars Alfrink, Ianus Keller, Gerd Kortuem, and Neelke Doorn. 2023. Contestable AI by design: Towards a framework.  
758 *Minds and Machines* 33, 4 (2023), 613–639.
- 759 [2] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. Machine Bias: There’s Software Used Across the  
760 Country to Predict Future Criminals. And It’s Biased Against Blacks. ProPublica.
- 761 [3] Joshua Ashkinaze, Emily Fry, Narendra Edara, Eric Gilbert, and Ceren Budak. 2024. Plurals: A system for guiding  
762 LLMs via simulated social ensembles. In *CHI*.
- 763 [4] Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna  
764 Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. 2022. Constitutional AI: Harmlessness from AI feedback. *arXiv*  
765 preprint *arXiv:2212.08073* (2022).
- 766 [5] Guiming Chen et al. 2024. Humans or LLMs as the judge? A study on judgement bias. *arXiv preprint* (2024).
- 767 [6] Yanda Chen et al. 2025. Reasoning Models Don’t Always Say What They Think. *arXiv preprint* (2025).
- 768 [7] Finale Doshi-Velez and Been Kim. 2017. Towards a rigorous science of interpretable machine learning. In *ICML*  
769 Workshop on Human Interpretability in Machine Learning.
- 770 [8] Yilun Du, Shuang Li, Antonio Torralba, Joshua B Tenenbaum, and Igor Mordatch. 2023. Improving factuality and  
771 reasoning in language models through multiagent debate. In *arXiv preprint arXiv:2305.14325*.
- 772 [9] Virginia Eubanks. 2018. *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*. St.  
773 Martin’s Press.
- 774 [10] European Commission High-Level Expert Group on AI. 2019. Ethics Guidelines for Trustworthy AI.
- 775 [11] Luyu Gao, Zhuyun Dai, Panupong Pasupat, Anthony Chen, Arun Tejasvi Chaganty, Yicheng Fan, Vincent Y Zhao,  
776 Ni Lao, Hongrae Lee, Da-Cheng Juan, et al. 2023. RARR: Researching and revising what language models say, using  
777 language models. In *ACL*.

- [781] [12] Abdul KM Haque, AKM Najmul Islam, and Patrick Mikalef. 2023. Explainable AI from the user perspective: A  
 782 systematic literature review. In *Pacific Asia Conference on Information Systems*.
- [783] [13] Donna Haraway. 1988. Situated knowledges: The science question in feminism and the privilege of partial perspective.  
 784 *Feminist studies* 14, 3 (1988), 575–599.
- [785] [14] Geoffrey Irving, Paul Christiano, and Dario Amodei. 2018. AI safety via debate. In *arXiv preprint arXiv:1805.00899*.
- [786] [15] Alon Jacovi and Yoav Goldberg. 2020. Towards faithfully interpretable NLP systems: How should we define and evaluate  
 787 faithfulness?. In *Proceedings of ACL*. 4198–4205.
- [788] [16] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Yejin Bang, Andrea Madotto, and  
 789 Pascale Fung. 2023. Survey of hallucination in natural language generation. *Comput. Surveys* 55, 12 (2023), 1–38.
- [790] [17] Margot E Kaminski and Jennifer M Urban. 2021. The right to contest AI. *Columbia Law Review* 121, 7 (2021),  
 1957–2048.
- [791] [18] Atoosa Kasirzadeh. 2024. Plurality of value pluralism and AI value alignment. In *Pluralistic Alignment Workshop at  
 792 NeurIPS 2024*. <https://openreview.net/forum?id=AOokh1UYLH>
- [793] [19] Atoosa Kasirzadeh and Iason Gabriel. 2023. In Conversation with Artificial Intelligence: Aligning Language Models  
 794 with Human Values. *Philosophy & Technology* 36, 2 (2023), 1–23.
- [795] [20] Akbir Khan, John Hughes, Dan Valentine, Laura Ruis, Kris Sachan, Ansh Raber, Arthur Gretton, et al. 2024. Debating  
 796 with more persuasive LLMs leads to more truthful answers. In *ICML*. Best Paper Award.
- [797] [21] Nikita Kozodoi, Johannes Jacob, and Stefan Lessmann. 2022. Fairness in credit scoring: Assessment, implementation  
 798 and profit implications. *European Journal of Operational Research* 297, 3 (2022), 1083–1094.
- [799] [22] 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.
- [800] [23] Khoa Lam et al. 2024. Assurance audits for AI systems. In *Proceedings of FAccT*.
- [801] [24] Timothy Lebo, Satya Sahoo, and Deborah McGuinness. 2013. PROV-O: The PROV ontology. W3C Recommendation.
- [802] [25] Francesco Leofante and Francesca Toni. 2024. Contestable AI needs computational argumentation. In *Proceedings of  
 803 KR*.
- [804] [26] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler,  
 805 Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive NLP  
 806 tasks. In *NeurIPS*.
- [807] [27] Zachary C Lipton. 2018. The mythos of model interpretability. *Queue* 16, 3 (2018), 31–57.
- [808] [28] Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence* 267  
 (2019), 1–38.
- [809] [29] Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. 2019. Dissecting racial bias in an algorithm  
 810 used to manage the health of populations. *Science* 366, 6464 (2019), 447–453. <https://doi.org/10.1126/science.aax2342>
- [811] [30] Manish Raghavan, Solon Barocas, Jon Kleinberg, and Karen Levy. 2020. Mitigating Bias in Algorithmic Hiring:  
 812 Evaluating Claims and Practices. In *Proceedings of FAccT*. 469–481. <https://doi.org/10.1145/3351095.3372828>
- [813] [31] Iyad Rahwan. 2018. Society-in-the-loop: Programming the algorithmic social contract. *Ethics and Information  
 814 Technology* 20, 1 (2018), 5–14.
- [815] [32] Inioluwa Deborah Raji, Andrew Smart, Rebecca N White, Margaret Mitchell, Timnit Gebru, Ben Hutchinson, Jamila  
 816 Smith-Loud, Daniel Theron, and Parker Barnes. 2020. Closing the AI accountability gap: Defining an end-to-end  
 817 framework for internal algorithmic auditing. In *Proceedings of FAccT*. 33–44.
- [818] [33] Allen Roush et al. 2024. OpenDebateEvidence: A massive-scale dataset for argument mining and summarization. *arXiv  
 819 preprint* (2024).
- [820] [34] Allen Roush et al. 2025. A superpersuasive autonomous policy debating system. *arXiv preprint* (2025).
- [821] [35] Cynthia Rudin. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable  
 822 models instead. *Nature machine intelligence* 1, 5 (2019), 206–215.
- [823] [36] Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. 2021. Retrieval augmentation reduces  
 824 hallucination in conversation. In *Findings of EMNLP*. 3784–3803.
- [825] [37] Noam Slonim et al. 2021. An autonomous debating system. *Nature* 591, 7850 (2021), 379–384.
- [826] [38] Alfred C Snider. 2008. *Code of the debater: Introduction to policy debating*. IDEA Press Books.
- [827] [39] Aman Singh Thakur et al. 2024. Judging the judges: Evaluating alignment and vulnerabilities in LLMs-as-judges. *arXiv  
 828 preprint* (2024).
- [829] [40] Stephen E Toulmin. 1958. *The uses of argument*. Cambridge University Press.
- [830] [41] Miles Turpin, Julian Michael, Ethan Perez, and Samuel R Bowman. 2023. Language models don't always say what they  
 831 think: Unfaithful explanations in chain-of-thought prompting. In *NeurIPS*.
- [832] [42] Alexandros Vassiliades, Nick Bassiliades, and Theodore Patkos. 2021. Argumentation and explainable artificial  
 833 intelligence: A survey. *Knowledge Engineering Review* 36 (2021).

- 833 [43] Sandra Wachter, Brent Mittelstadt, and Luciano Floridi. 2017. Why a right to explanation of automated decision-making  
834 does not exist in the General Data Protection Regulation. *International Data Privacy Law* 7, 2 (2017), 76–99.  
835 [44] Yuan Yao. 2024. Explanatory pluralism in explainable AI. *AI Magazine* 45, 1 (2024), 82–93.  
836 [45] Lianmin Zheng et al. 2023. Judging LLM-as-a-judge with MT-bench and Chatbot Arena. In *NeurIPS*.  
837  
838  
839  
840  
841  
842  
843  
844  
845  
846  
847  
848  
849  
850  
851  
852  
853  
854  
855  
856  
857  
858  
859  
860  
861  
862  
863  
864  
865  
866  
867  
868  
869  
870  
871  
872  
873  
874  
875  
876  
877  
878  
879  
880  
881  
882  
883  
884