

The Talmudic Epistemic Stack

Structured Disagreement as a Methodology for Trustworthy, Updateable, Adversarially-Robust AI Knowledge Bases

Submission to the Future of Life Foundation Epistemic Stack Competition (“Lab Leaks, Black Holes, and Eggs”) · v1.0 · 2026 · CC-BY-4.0 Worked examples, full proofs, schemas and the 987-paper corpus index are appendix material in the accompanying repository ([epistemic_stack/](#)); curated entry points in its README.

Abstract. Modern AI knowledge tools treat knowledge as *answers*: they retrieve sources, emit a narrative or a single probability, and discard the reasoning. Such artifacts are hard to trust (provenance is a footnote), hard to update (revision means rebuilding), brittle under adversarial pressure (objections are not kept), and impossible to compound (the next investigator restarts). We argue that the **Talmud** — a continuously-edited legal-intellectual knowledge commons maintained, attacked, and extended by thousands of contributors for ~1,500 years — is the most battle-tested existing solution to exactly this problem, because it treats knowledge as **structured disagreement** and preserves the reasoning rather than collapsing it. We make that editorial discipline computable using **Dov Gabbay’s** logics (Labelled Deductive Systems, abstract argumentation, fibring, non-monotonic/defeasible reasoning, AGM belief revision, reactive/temporal logic) and his *Talmudic Logic Project*. The result is a methodology — the **Sugya Protocol** — and a dual-database (Neo4j + ArangoDB) substrate in which every claim is provenance-labelled, evaluated under several logics at once, and updated by versioning rather than overwriting. We connect each FLF criterion to a published theorem, and demonstrate the method on all three case studies. We are deliberately careful about the central claim: we do **not** assert “Judaism solves everything” (unfalsifiable, un-Talmudic); we defend the narrower, demonstrable claim that a *specific* Talmudic epistemic technology — disciplined preservation of structured disagreement with non-monotonic, provenance-aware revision — is a uniquely mature design pattern for AI epistemic infrastructure, and we log where our own argument is unresolved (*teiku*).

1. The problem and the thesis

FLF seeks “workflows and methodologies for using AI to produce reliable, trustworthy knowledge bases” that **compound**: “if structured analyses become reusable, refineable artifacts, every serious investigation enables future work ... by the same or different people.” The dominant AI paradigm fails this almost by definition. Retrieval-augmented generation yields a fluent narrative whose provenance is a list of links and whose reasoning is not a manipulable object. A single Bayesian probability is worse in a way the competition deliberately stresses: as we note in §7, six competent Bayesian analyses of the **same** COVID-origins evidence span **twenty-three orders of magnitude**. A system that emits one number has not represented the state of knowledge — it has hidden it.

Our thesis: the problem FLF poses was substantially solved, for legal and intellectual disputes, by the **editorial method of the Talmud**, and that method is now mechanizable. The Talmud is a *format* and a *process*, not merely an old text. Its defining move is to **preserve the structure of disagreement** — question (*kushya*), objection, answer (*terutz*), distinction (*chiluk*), and *machloket*: a disagreement recorded with both sides, their proponents, and their support, **never flattened to a winner** even when a practical ruling issues. It records honest non-resolution (*teiku*) as a first-class output. It is layered (Mishnah → Gemara → Rishonim → Acharonim); each layer cites and argues with the previous without overwriting it — an append-only, provenance-preserving, adversarially-edited knowledge commons maintained across fifteen centuries and a globally distributed contributor base. The bridge from “good format” to “buildable system” is **Dov Gabbay’s** formal program.

2. What FLF is buying (the criteria)

FLF splits the task into three layers — **ingestion** (attribute claims to sources with provenance), **structure** (inference + discourse structure; “similar but not identical” claims; evolution over time), and **assessment** (rhetorical-vs-evidential weight; correlated evidence treated as independent; **cruxes**; what’s missing; confidence accounting for out-of-model error; *distinguish what the debate settled from what it merely performed settling*). From these we use eleven dimensions throughout: **provenance, structure, assessment, adversarial robustness, scalability, interoperability, updateability, faithful uncertainty, crux discovery, reusability, and settled-vs-performed**. The last is the Talmudic competence *par excellence*: the genre exists to separate a binding ruling (*halacha*) from the reasoning theater around it, and *machloket* (live dispute) from *teiku* (unresolved) from *psak* (ruling).

3. The Talmud as a knowledge-commons format

Viewed as an information system, four structural properties stand out — each a FLF desideratum:

1. **Preserved disagreement (machloket).** Minority and rejected opinions are recorded *by name* with their reasoning, beside the accepted view (m. Eduyot 1:5–6, explicitly so a later court may rely on them). This is non-destructive editing of a contested base — exactly “distinguish what was settled from what merely performed settling,” and exactly what argument-mapping tools that drive to one winner lack.
2. **Layered provenance.** Each stratum cites and argues with its predecessors; a claim’s *authority* is a function of *who said it, when, in what chain of transmission*. Provenance is the primary key, not bolted-on metadata.
3. **Append-only revision.** New rulings are added as commentary literally surrounding the original; nothing is deleted. The Talmudic page (*daf*) is a working interoperable, multi-perspective format that compounded across centuries.
4. **Honest non-resolution (teiku).** Unresolvable disputes are recorded and retained as valuable, citable outputs — faithful uncertainty as an editorial norm.

Why a fifteen-century-old format? Longevity under adversarial pressure is evidence of robustness: the method survived dispersion, censorship, and continuous internal red-teaming (its whole genre is objection and counter-objection) while remaining coherent and extensible. This is the defensible core of the “tradition that survived adverse selection may help us survive powerful AI” intuition — not metaphysics, but the observation that a *specific epistemic technology* has demonstrated extreme durability and is worth porting.

Honest scope. A dedicated adversarial check (§10) found that the specific “Judaism-survived-adverse-selection → useful-for-RSI-AI” thesis is articulated in the literature by *others* (e.g., Zibulevsky 2024; Kalman/FLI 2023), not as a documented institutional position of any single lab; AE Studio’s public alignment work is the “Neglected Approaches” agenda and its AI-antisemitism research. We therefore treat the survival argument as a *motivating hypothesis with independent support* and rest the case on the formal logic and the case studies, which stand on their own.

4. Making it computable: Gabbay’s apparatus

The Talmudic Logic Project (Gabbay, Schild & David, *Logica Universalis* 13(4):425–442, 2019) has formalized specific Talmudic rules we reuse: *kal vachomer* (a fortiori) via **matrix abduction** (Abraham, Gabbay & Schild, *Studia Logica* 92(3):281–364, 2009); Talmudic deontic logic (*AI & Law* 19, 2011); temporal/contrary-to-time conditionals (*AI & Law* 20, 2012); Talmudic public-announcement logic that deletes *accessibility links*

rather than worlds, i.e. reversible update (*J. Applied Logic* 11, 2013); the *Klal uPrat* set rules (*HPL* 32, 2011); and **uncertainty rules** — *rov* (majority), *chazakah* (presumption) — formalizing *safek* (Gabbay & Koppel, *HPL* 32(1), 2011).

Gabbay’s general program supplies the rest. **Labelled Deductive Systems** attach a label (assumptions, sources, reasoning path) to every formula — provenance as a carried object (*Labelled Deduction*, Kluwer 1996). **Abstract argumentation** (Dung, *Artif. Intell.* 77(2):321–357, 1995) models a debate as an attack graph whose *extensions* are the defensible sets; Gabbay’s extensions type each attack by its logical relation, allow attacks-on-attacks (meta-disagreement), add graded *support*, and require a *dialectical tier* in which a claim must answer objections before it stands — the **kushya** → **terutz** cycle as a soundness condition. **Fibring** combines logics without reducing one to another — the formal license for “never collapse the dimensions into one score” — and even fibres argumentation frames with Bayesian and neural networks. **Non-monotonic/defeasible** consequence makes conclusions retractable (the essence of updateability) and **AGM** belief revision (*Frontiers in Belief Revision*, Kluwer 2001) gives the rationality postulates for principled updating. **Reactive Kripke semantics** lets traversing one edge switch others, giving context-aware, self-modifying assessment with completeness theorems.

A real substrate. This provenance algebra runs over a large corpus of Jewish texts the author maintains in **Neo4j** and **ArangoDB** — millions of OCR’d English volumes plus thousands of hand-scanned curated works spanning every period (Biblical → Tannaitic → Amoraim → Geonim → Rishonim → Acharonim → modern), with mathematics and psychology holdings that inform the assessment layer. The repository’s `torah.yaml` already encodes the authority tiers (Biblical = 10, Talmudic = 9, Rishonim = 8 ...) and Hebrew/Aramaic/Yiddish/English fibring; the same machinery that ranks a Rishon against an Acharon ranks a peer-reviewed paper against a preprint against an anonymous web claim. (The live indices were not reachable from the build sandbox; artifacts are built to load into them — §10.)

5. The Sugya Protocol

The primary object is a **Sugya**: an append-only argument graph storing the *moves*, not a bare answer. Node types and their roles:

Node	Talmudic name	Role
Question / Claim / Evidence	—	what’s asked; assertions (defeasibly labelled); data for/against
Source / Authority	—	who said it; how independent; what incentive
Kushiya / Terutz	objection / resolution	strongest attack; strongest answer
Chiluk	distinction	shows two “contradictory” claims hold in different cases
Machloket / Position	preserved disagreement	both sides kept, never flattened
NafkaMina / Safek / Teiku	consequence / uncertainty / unresolved	downstream stakes; typed uncertainty; honest open question

Invariant: nothing is deleted. Updates set `valid_to` and add a **REVISES** edge to a new version; every node carries an LDS label `L=(truth, completeness, sources, priority)` and a `talmudic_status` ∈ {accepted, disputed, minority, majority, practical_ruling, unresolved, suspended}.

Seven operations, each a graph mutation backed by a logic: **kushya** (typed attack) · **terutz** (defense/reinstatement) · **chiluk** (matrix-completion distinction) · **machloket** (paraconsistent preservation) · **nafka mina** (defeasible consequence) · **safek** (typed uncertainty; factual/model/source/measurement/adversarial) · **teiku** (logged non-resolution).

Agents, single-responsibility and deterministically orchestrated (so judges can re-run on a new case): *Mishnah* (extract claims+provenance, no interpretation) → *Gemara* (kushya/terutz/chiluk) → *Tosafot* (cross-link, find repeated assumptions) → *Eidim* (source authority; mark correlated evidence) → *Stam* (missing evidence; new questions) → *Machloket* (reduce to ranked **cruxes**) → *Daat* (incentives, meta-level only). **Crux priority** = $\text{leverage} \times \text{uncertainty} / (\text{tractability} + 1)$ — the assessment payload telling the next investigator where effort buys the most belief-movement.

6. Why the protocol satisfies the criteria (formal core)

Each is a rigorous link to a published result, not a novel theorem (full statements in the repository paper §6).

- **P1 — Preserved disagreement is safe.** Two mutually-attacking arguments yield a Dung grounded extension that accepts *neither* (both *undecided*) — natively representing *teiku/safek*; and **paraconsistent restricted-access** inference fires classical rules only over consistent subsets, so a contradiction is *localized*, not globally explosive (*ex falso* blocked). A practical ruling is a separate node, leaving the minority standing — the Mishnaic norm.
- **P2 — Updateability = non-monotonic consequence + AGM, as append-only versioning.** Adding evidence can retract a conclusion; `valid_to+REVISES` implement contraction-then-expansion while keeping the prior state for provenance, satisfying Success/Consistency; **independence/factorization** makes minimal change *local*, so revision in one subgraph does not perturb independent parts — the property that makes updateability scale.
- **P3 — Chiluk dissolves false contradictions by completion, not fiat.** In the case×attribute matrix (matrix abduction), an apparent “ $A \wedge \neg A$ ” is resolved by exhibiting a distinguishing column, re-indexing to $A|C$, $\neg A|\neg C$ — both consistent; “same unless distinguished” is *indiscernibility by default*, the same rule the repo’s entity-resolution uses.
- **P4 — Faithful uncertainty is a fibred vector, and interoperability is provable.** Fibring keeps classical/defeasible/temporal/authority/probabilistic verdicts as distinct components (probabilistic via PrAF; calculus may be probability, possibility, or Dempster–Shafer, distinguishing *ignorance* from *risk*). And **Dung’s complete extensions coincide exactly with three-valued stable models of logic programs** (Caminada & Gabbay) — so the same accepted/rejected/undecided verdict is recoverable whether a future investigator runs a debate-graph tool or a rule engine.
- **P5 — Eidim de-correlation prevents double-counting.** Independence/factorization plus substructural (resource-counting) proof underwrite a `CORRELATED_WITH {ρ}` discount, so two dependent sources do not contribute as two independent ones — FLF’s explicit ask.
- **P6 — Reactive accessibility gives context-aware assessment with a sound/complete basis.** Reactive Kripke models (edges that switch other edges), with completeness theorems and a proven correspondence to argumentation networks, back a graph whose accepted set is recomputed as context and time change.

7. Worked examples (full Sugyot in the appendix repository)

COVID-19 origins (open, adversarial). The 2023 Rootclaim debate ended 2–0 for zoonosis, yet six Bayesian analyses of the same evidence span ~23 OOM — proof that one number hides the state of knowledge. Our Sugya keeps both positions as a *Machloket*, records the judges’ verdict as a *separate practical_ruling* node (settled-vs-performed), and reduces the dispute to ranked cruxes (market clustering vs. ascertainment bias; the A/B two-lineage argument; the furin cleavage site; DEFUSE). The *Eidim* discount stops the many shared-data market papers from counting as independent confirmations; typed *Safek* shows the 23-OOM spread is

model uncertainty (priors), not data. (Citations: Pekar et al., *Science* 2022; Worobey et al. 2022; Weissman 2024; reply arXiv:2405.08040; Alexander, ACX 2024.)

LHC micro black holes (closed, defense-in-depth). The Sugya encodes three independent backstops — Hawking evaporation → cosmic-ray survival → white-dwarf/neutron-star survival (Giddings & Mangano, *Phys. Rev. D* 78:035009, 2008) — each a Terutz catching the previous layer’s failure, and makes the *weakest links explicit* as Safek: Hawking radiation is unobserved (so the case is built not to need it), and the neutral/slow-accretion corner is least-closed. (LSAG, *J. Phys. G* 35:115004, 2008; Hawking 1975; Dimopoulos & Landsberg, *PRL* 87:161602, arXiv:hep-ph/0106295.)

Eggs / dietary cholesterol (messy, individual). Central work is chiluk: five distinctions dissolve the apparent contradiction between harm cohorts (Zhong, *JAMA* 2019) and null cohorts (Drouin-Chartier, *BMJ* 2020; PURE, *AJCN* 2020) — dietary cholesterol ≠ serum LDL; average ≠ individual (hyper-responders/ApoE4); surrogate ≠ hard endpoint; association ≠ causation; LDL-mediated ≠ TMAO-independent. The grounded extension leaves both headline claims *undecided* (correct: preserved Machloket + Teiku), with a logged, conditional bottom line rather than a false universal verdict.

8. Self-audit against the criteria

Criterion	Mechanism	Backing
Provenance	ASSERTED_BY + LDS sources; PROV-O projection	§4
Structure	typed support/attack/distinguish/answer edges	§4, P4
Assessment	grounded extension + Crux Map + Eidim + Daat	§5
Adversarial robustness	mandatory kushya; non-explosive base; self-review	P1, P5
Scalability	local (factorized) revision; complexity-aware fibring; shared graph stores	P2, §4
Interoperability	one JSON-Schema → Neo4j/ArangoDB/JSON-LD; Dung ≡ 3-valued stable models	P4
Updateability	append-only versioning + AGM REVISES	P2
Faithful uncertainty	typed Safek; fibred multi-logic verdict	P4
Crux discovery	ranked Crux nodes	§5
Reusability	CC-BY graph + machine-readable schema	§5
Settled-vs-performed	Machloket + separate practical_ruling; talmudic_status	P1, §3

9. Relation to prior art

Prior art	Limitation	Talmudic-stack addition
Toulmin (1958)	per-argument; resolves to one qualified claim	adopt rebuttal as kushya edge
IBIS (Kunz & Rittel 1970), Kialo	drive to a winning position; no authority/independence model	machloket + Eidic de-correlation
Walton schemes (2008)	defeasible but not a versioned provenance store	schemes as typed kushya/terutz on a graph
Bayesian / Rootclaim	single posterior hides analyst disagreement (23 OOM)	probability is one fibred component, not the whole
AI safety via debate (Irving et al. 2018)	designed to produce a winner	preserve the loser’s standing argument
Nanopublications, PROV-O (2010/2013)	provenance as static metadata	provenance as defeasible, authority-ranked, revisable layer

Four things none of the above combines, which this approach adds: (1) **preserved disagreement as standing, citable authority**; (2) **non-monotonic updateability with AGM guarantees**; (3) **fibred multi-logic verdicts** (no forced scalar); (4) **provenance-as-authority** (the LDS label is the trust computation, ranked on the same scale as a fifteen-century corpus).

10. Limitations and adversarial self-review (kushya on ourselves)

- **Overclaim.** “Judaism solves all the issues” is unfalsifiable; we do not defend it. We defend the narrower claim (abstract) and the title’s “solves” is shorthand for “supplies a battle-tested pattern that meets.”
- **Is the framing load-bearing or ornament?** §6 shows each operation maps to a *distinct* mechanism a generic argument map lacks (chiluk = matrix completion; machloket = paraconsistent non-explosion + Dung undecided; teiku = typed output; authority = LDS label on a real corpus). *Residual safek*: one could perhaps reach the same constructs without the vocabulary — we agree the constructs are what matter; the vocabulary is a guide to which to include.
- **Attribution.** The Judaism-survival thesis is attributed to the independent literature, not a lab (§3); verified.
- **Corpus not queried live.** The Synology Neo4j/ArangoDB indices were unreachable from the build sandbox; schemas/tiers/graphs are built to load into them, and the 987-paper logic corpus was mined directly. Logged, not hidden.
- **Case-study facts.** Adversarial verification caught and fixed real errors pre-publication (market–WIV distance >10 mi not 6; Dimopoulos–Landsberg arXiv id; Keys-equation form; Gabbay–Woods imprints); remaining hedged items are marked *Safek*.
- **Teiku (open questions, recorded not papered over).** (i) Does preserving disagreement serve a decision-maker who must act, or abdicate? We give an Executive-Summary interface with a bottom line *and* the structure, but the trade-off is real. (ii) At what graph size does Sugya readability degrade, and what summarization avoids re-flattening? (iii) Can the fibred verdict be made legible to non-logicians without collapsing it?

11. Conclusion

FLF wants knowledge bases that are trustworthy, updateable, adversarially robust, and *compounding*. The Talmud is the longest-running working example of such a system; its discipline — preserve the structured disagreement, never flatten it, revise by adding not overwriting, record honest non-resolution — is what makes knowledge compound across contributors and centuries; and Gabbay’s program makes that discipline computable. We instantiated it as the Sugya Protocol over a dual-database substrate, connected each criterion to a published result, and demonstrated it on an open dispute, a closed case, and a messy empirical question. *The deliverable is the graph and the method, not an answer; the answer, in the Talmudic sense, is the preserved structure within which future investigators continue the argument rather than restart it.*

Key references

Gabbay, Schild & David (2019), *Logica Universalis* 13(4):425–442 · Abraham, Gabbay & Schild (2009), *Studia Logica* 92(3):281–364 (matrix abduction / kal vachomer) · Abraham, Gabbay & Schild (2011), *AI & Law* 19; (2012) 20 · Abraham, Belfer, Gabbay & Schild (2013), *J. Applied Logic* 11(1):63–90 · Gabbay & Koppel (2011), *Hist. & Phil. of Logic* 32(1):63–69 · Dung (1995), *Artif. Intell.* 77(2):321–357 · Caminada & Gabbay (2009), *Studia Logica* 92(3) (complete extensions $\equiv$ 3-valued stable models) · Gabbay (1996) *Labelled Deduction*; (1999) *Fibring Logics*; (2008) *Analysis and Synthesis of Logics* · Gabbay & Rodrigues, eds. (2001) *Frontiers in Belief Revision* · Gabbay (2013) *Reactive Kripke Semantics* · **COVID:** Pekar et al., *Science* 2022 (abp8337); Worobey et al. 2022; Weissman 2024; reply arXiv:2405.08040; Alexander, ACX 2024 · **LHC:** LSAG, *J. Phys. G* 35:115004 (2008); Giddings & Mangano, *Phys. Rev. D* 78:035009 (2008); Hawking, *Commun. Math. Phys.* 43:199 (1975); Dimopoulos & Landsberg, *PRL* 87:161602 (2001) · **Eggs:** Zhong et al., *JAMA* 2019;321:1081; Drouin-Chartier et al., *BMJ* 2020;368:m513; Dehghan et al., *AJCN* 2020;111:795; Carson et al., *Circulation* 2020;141:e39; Wang et al., *Nature* 2011;472:57 · **Context/prior art:** Sourbut & Goldhaber (2025); Berg, Carauleanu & Rosenblatt (2023); Toulmin (1958); Kunz & Rittel (1970); Walton, Reed & Macagno (2008); Irving, Christiano & Amodei (2018, arXiv:1805.00899); Groth et al. (2010); W3C PROV-O (2013).

Full proofs, worked Sugya graphs (Neo4j + ArangoDB loaders, Crux/Uncertainty maps), schemas, and the complete 987-paper corpus index with 154 curated Gabbay-program references are in the accompanying repository under `epistemic_stack/`.