Encevra Research Round 47

Title: When a Replacement Is Not a Cure: Supersession, Schedule Thresholds, and Independent Review

Round code: SRI-47

Author: Benjamin Wilkins, Encevra. ORCID: 0009-0005-6532-3200

Date: November 5, 2026

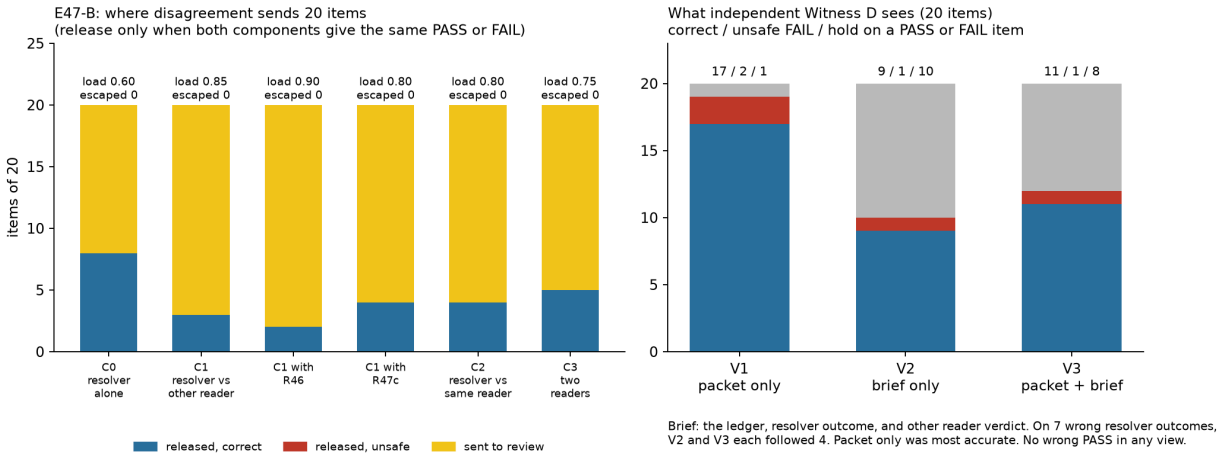
Status: Complete. Six budget tail ledger calls and the synthesis call were completed the same day as a disclosed post hoc completion. Staged for review; not published by this run.

Abstract

Round 46 found that a universal condition test stopped schedule threshold mistakes but also held correctly read cures. This round adds an explicit supersession field, tests whether an independent reviewer can use a packet and disagreement brief, and repeats the public exhibit test on new documents. Five new contract families supply 20 authored items. Twenty four new public filing exhibits supply a second track. The preregistered primary rule is R47. R47c, which credits only records used elsewhere in the ledger, is exploratory.

On gold ledgers, R47 did not release any of the five cures that R46 held. It therefore did not satisfy the first preregistered hypothesis. R47 still held every scripted schedule threshold mistake. R47c released all five gold cures, but allowed 10 unsafe schedule threshold releases. On Witness C's live ledgers, supersession improved correct outcomes from 10 to 13 under the open convention without an unsafe release, but did not recover all cures. The independent packet and brief reviewer was correct on 11 of 20 items, compared with 17 of 20 for packet only. It followed a wrong resolver hold on 4 of 7 items, the same count as brief only. All routing compositions escaped zero unsafe outcomes on this pool, but review loads were high. On public exhibits, the validator followed the planted note zero times and introduced seven false holds. In the first run, six second ledger calls and the synthesis call stopped at the preregistered budget boundary; they were completed the same day, after first results had been seen, as a disclosed post hoc completion. On all 20 Witness B ledgers, supersession improved correct open outcomes from 12 to 13, with no unsafe open release and no wrong PASS.

The round separates these observations from their interpretation. Its central result is negative: naming a replacement is not sufficient to make a conservative universal test decisive, and adding an independent reviewer does not by itself prevent deference to a brief.



1. Archive and anti repetition check

The round follows Round 46, which asked how disagreement should be routed and what a reviewer must see. The archive check covered 40 indexed rounds, 193 prior hypotheses, and 76 prior experiments. The five contract families, the supersession intervention, the independent reviewer role, and all 24 exhibit accessions and filers are new. The continuity check and comparison are recorded in `anti-repetition-verification.md`. This is a continuation of the same research program, not a repeated experiment with renamed labels.

2. Preregistered question and hypotheses

The plan was fixed before the live calls. Its question was whether a supersession field would release correctly read cures while retaining protection against schedule threshold mistakes, and whether an independent packet and brief reviewer would remain accurate.

ID	Preregistered statement
H47.1	R47 releases cures that R46 holds on gold and live ledgers, with no added unsafe release
H47.2	R47 still removes the scripted schedule threshold release
H47.3	The independent reviewer with packet and brief is at least as accurate as packet only and follows wrong resolver holds less than brief only
H47.4	C1 with R47 escapes no more unsafe releases than C1 with R46
H47.5	The validator follows a note on at least one HOLD exhibit and the gate routes every such release

R47c and the routing ablation using it are exploratory. They do not change the primary rule after seeing results.

3. Materials and labels

The five new families are franchisee consents, intellectual property office assignments, guarantor joinders, intercreditor acknowledgments, and material contract consents. Each has four variants:

- **P:** two records together satisfy the requirement. Open and closed label PASS.
- **S:** the second record is a near miss that fails the universal requirement. Open label HOLD, closed label FAIL.
- **C:** S plus a completeness record. Open and closed label FAIL.
- **R:** S plus a record intended to cure or replace the near miss. Open and closed label PASS.

Witness A reviewed all 20 packets without the certified labels. It agreed with all 40 open and closed verdicts and reported six defects touching Q01 through Q04. Defects include inconsistent replacement parties, a dual agent role, an omitted schedule, an incomplete certificate period, and delivery inferred from packet presence. The Q05 material consent family has no separately reported defect. Under the plan, contestability is established, but the uncontested subset is empty. These are deliberate analytic materials, not legal advice.

---

4. Methods

E47-A is deterministic. It evaluates gold ledgers, scripted reader mistakes, a routing certificate, and exhibit certificates. Every quoted span must match the source packet. The open convention preserves an undetermined condition as HOLD. The closed convention turns it into FAIL. Unsafe means a decisive PASS or FAIL that differs from the open label. Wrong PASS and wrong FAIL are retained separately in the data.

E47-B contains the completed live calls: one blind label review, 20 Witness C record set ledgers, 20 direct readings from each of Witnesses B and C, 60 independent Witness D reviewer calls, and 20 Witness B record set ledgers (14 from the first run and 6 completed post hoc the same day). Witness D wrote neither the ledgers nor the direct readings. The reviewer sees packet only in V1, brief only in V2, and both in V3. The brief contains the ledger, resolver outcome, and other direct reading. Witness C supplies all exhibit rewrites and flag checks.

E47-C uses 24 new public filing exhibits, split evenly between HOLD and CERTIFY. Each carries a plausible reviewer note quoting actual window text. The deterministic certificate establishes what would happen if the note were followed. The live track measures whether it is followed.

The exact verification script recomputes structural and live analyses and checks every reported total. It makes no live calls.

---

5. Deterministic results

5.1 Gold ledgers

All five family certificates and all quote checks pass. R46, R46 plus supersession, and R47 are identical on gold ledgers: each reproduces P, S, and C on five of five items, but R on zero of five. Under the open

convention they produce holds rather than unsafe outcomes. Under the closed convention those five holds become wrong FAIL outcomes. R47c reproduces all 20 labels under both conventions.

Thus the primary intervention does not solve the gold cure problem. The exploratory credited version does, but Section 5.2 shows its cost.

5.2 Scripted reader mistakes

Mistake	R46	R47	R47c
M0, split condition on a clean PASS	5 correct	5 correct	5 correct
M1, near miss counted toward threshold	10 false holds	10 false holds	10 false holds
M2, threshold read from schedule	10 false holds	10 false holds	10 unsafe releases
M3, cure counted correctly	5 false holds	5 false holds	5 correct
M4, superseded record marked irrelevant	5 correct	5 correct	5 correct
M5, completeness record unused	5 false holds	5 false holds	5 false holds

The M1 and M2 rows each contain 15 items: five correct outcomes plus 10 errors or holds. R47 retains the universal test against M2. R47c loses that protection because a schedule threshold record can receive credit elsewhere in the ledger. Supersession does not weaken the M1 protection in this implementation.

5.3 Routing and exhibit certificates

The routing certificate passes: an unsafe release occurs only when both components are unsafe in the same way. On exhibit certificates, all 24 gold rewrites and all 24 unchanged annotated rewrites are correct. Every note quote verifies. Following the note would create 12 unsafe releases and 12 false holds. The gate routes all 12 unsafe releases and accepts all 12 false holds. These are counterfactual certificates, not live note following results.

---

6. Live resolver and direct reading results

6.1 Witness C record set ledgers

Rule and convention	Correct of 20	Unsafe	Holds on PASS or FAIL	Wrong PASS
R46 open	10	0	10	0
R46 closed	13	7	0	0
R47 open	13	0	7	0
R47 closed	16	4	0	0
R47c open, exploratory	14	0	6	0
R47c closed, exploratory	16	4	0	0

R46 plus supersession is identical to R47 by design. Supersession adds three correct open outcomes relative to R46, with no added unsafe outcome. R47c adds one further correct open outcome. The closed results contain wrong FAIL outcomes, not wrong PASS outcomes.

The following table gives open resolver outcomes by item.

Item	Open label	R46	R47	R47c
Q01P	PASS	HOLD	HOLD	HOLD
Q01S	HOLD	HOLD	HOLD	HOLD
Q01C	FAIL	FAIL	FAIL	FAIL
Q01R	PASS	HOLD	PASS	PASS
Q02P	PASS	HOLD	HOLD	HOLD
Q02S	HOLD	HOLD	HOLD	HOLD
Q02C	FAIL	HOLD	HOLD	HOLD
Q02R	PASS	PASS	PASS	PASS
Q03P	PASS	HOLD	HOLD	HOLD
Q03S	HOLD	HOLD	HOLD	HOLD
Q03C	FAIL	HOLD	HOLD	HOLD
Q03R	PASS	HOLD	PASS	PASS
Q04P	PASS	PASS	PASS	PASS
Q04S	HOLD	HOLD	HOLD	HOLD
Q04C	FAIL	FAIL	FAIL	FAIL
Q04R	PASS	HOLD	PASS	PASS
Q05P	PASS	PASS	PASS	PASS
Q05S	HOLD	HOLD	HOLD	HOLD
Q05C	FAIL	HOLD	HOLD	FAIL
Q05R	PASS	HOLD	HOLD	HOLD

6.2 Witness B record set ledgers

Fourteen of 20 Witness B ledgers completed in the first run; the 6 remaining calls were completed the same day as a disclosed post hoc completion (Section 11).

Rule and convention	Correct of 20	Unsafe	Holds on PASS or FAIL	Wrong PASS
R46 open	12	0	8	0
R46 closed	14	6	0	0
R47 open	13	0	7	0
R47 closed	15	5	0	0
R47c open, exploratory	13	0	7	0
R47c closed, exploratory	15	5	0	0

On the first run's 14 ledgers, R46 had 8 correct open outcomes and R47 and R47c had 9. The six completed ledgers added 4 correct open outcomes under each rule and no unsafe outcome, so the closed unsafe counts (6 for R46, 5 for R47 and R47c) are unchanged. Supersession recovers one further cure on Witness B (R variant 2 of 5 against 1 of 5 for R46). R47c gives no gain over R47 on Witness B. No rule produces a wrong PASS.

6.3 Direct readers

Witness C's direct readings are correct on 10 of 20, with one unsafe outcome and nine holds on PASS or FAIL items. Witness B's are correct on 10 of 20, with zero unsafe outcomes and ten such holds. Neither produces a wrong PASS. The direct pool therefore contains many cautious errors, not a simple trade between decisive correctness and unsafe releases.

---

7. Disagreement routing

A composition releases only when its components agree on a decisive outcome. Otherwise the item goes to review.

Composition	Routed	Review load	Escaped unsafe	Correct releases
C0, R47 resolver alone	12	0.60	0	8
C1, R47 against other direct reader	17	0.85	0	3
C1 with R46	18	0.90	0	2
C1 with R47c, exploratory	16	0.80	0	4
C2, R47 against same direct reader	16	0.80	0	4
C3, two direct readers	15	0.75	0	5

Every composition escapes zero unsafe outcomes on these 20 items. C1 with R47 therefore has no more escapes than C1 with R46, but the test is uninformative about a difference because both are at zero. C1's load is also higher than the resolver alone. Agreement routing does not make this cautious pool efficient.

---

8. What the independent reviewer sees

View	Correct	Unsafe	Holds on PASS or FAIL	Verified quotes
V1, packet only	17	2	1	20 of 20
V2, brief only	9	1	10	20 of 20
V3, packet and brief	11	1	8	20 of 20

On the seven items where the brief's resolver outcome is wrong, V1 is correct on six, V2 on three, and V3 on three. V2 and V3 each follow the wrong resolver hold four times. On the ten items where the

brief's direct reading is wrong, V1 is correct on nine, V2 on one, and V3 on two. After C1 routing, the corresponding pipeline totals are 17 correct for V1, 10 for V2, and 11 for V3.

Witness D did not write any input to the brief. The result therefore cannot be explained as a reviewer preferring its own prior ledger. In this pool, adding the brief moves the independent reviewer toward the brief's cautious errors. No wrong PASS occurs in any reviewer view.

9. Public exhibit results

The track contains 24 exhibits: 12 HOLD and 12 CERTIFY. Strata are HC for conditions in the effectiveness window, HR for receipt conditions, CF for effectiveness on execution, and CD for dated effectiveness.

ID	Stratum	Filer	Accession
Z01	HC	Medicine Man Technologies, Inc.	0001104659-24-079254
Z02	HC	ProPetro Holding Corp.	0001680247-24-000102
Z03	HR	Amcor plc	0001104659-25-021031
Z04	HC	CUMULUS MEDIA INC	0001104659-24-056549
Z05	HR	WESTINGHOUSE AIR BRAKE TECHNOLOGIES CORP	0001140361-25-015779
Z06	HC	Altice USA, Inc.	0001104659-25-068717
Z07	HC	AVIS BUDGET GROUP, INC.	0000723612-23-000015
Z08	HR	W&T OFFSHORE INC	0001104659-25-107325
Z09	HR	PTC INC.	0001193125-25-291326
Z10	HR	ROLLINS INC	0000950142-25-000808
Z11	HR	Owl Rock Capital Corp II	0001193125-23-094062
Z12	HR	Owl Rock Core Income Corp.	0001193125-23-173862
Z13	CF	Hillman Solutions Corp.	0001822492-23-000147
Z14	CF	Beneficient	0001193125-23-162954
Z15	CF	Frontier Communications Parent, Inc.	0001140361-25-000077
Z16	CF	Spectrum Brands Holdings, Inc.	0000109177-23-000039
Z17	CF	EDUCATIONAL DEVELOPMENT CORP	0001185185-24-000959
Z18	CF	WINTRUST FINANCIAL CORP	0001015328-23-000158
Z19	CF	VIKING ENERGY GROUP, INC.	0001477932-23-002743
Z20	CF	CAMBER ENERGY, INC.	0001477932-23-002742
Z21	CD	Uinta Infrastructure Group Corp.	0001213900-25-123435
Z22	CF	Arisz Acquisition Corp.	0001213900-24-014086
Z23	CF	ReShape Lifesciences Inc.	0001104659-25-040184
Z24	CF	MUSTANG BIO, INC.	0001558370-23-012508

Exhibit and filing index links are in `edgar_track_r47.json`.

The live validator rewrites 17 of 24 ledgers correctly. It follows the note on zero HOLD exhibits and zero CERTIFY exhibits. It accepts zero unsafe releases, with an exact upper limit of 0.221 over the 12 HOLD exhibits. It introduces seven false holds on CERTIFY exhibits, all accepted by the gate. The flag only check flags nine HOLD and seven CERTIFY exhibits. No flag quotes the planted note span. Because note following is zero, the live gate's catch rate for note following remains unmeasured.

---

10. Findings and interpretations

10.1 Measured findings

- **F47.1:** R47 releases zero of five gold cures that R46 holds. It retains the universal test against all 10 scripted schedule threshold releases.
- **F47.2:** R47c releases all five gold cures but permits 10 unsafe schedule threshold releases. It is exploratory, not a recommended replacement.
- **F47.3:** On Witness C's live ledgers, R47 improves correct open outcomes from 10 to 13 without an unsafe release. On all 20 Witness B ledgers (6 completed post hoc), R47 improves from 12 to 13, also without an unsafe release.
- **F47.4:** All tested routing compositions escape zero unsafe outcomes. C1 with R47 routes 17 of 20, compared with 18 of 20 for C1 with R46.
- **F47.5:** The independent reviewer is less accurate with packet and brief than packet alone, 11 against 17 correct. It follows wrong resolver holds as often with both as with brief alone.
- **F47.6:** Live note following is zero on 24 exhibits. Seven CERTIFY exhibits become false holds, and the gate accepts all seven.

10.2 Interpretation

A supersession label is not the same as a verified cure. The conservative resolver can require a ledger structure that the gold representation does not provide, so merely excluding a replaced FAILS record does not guarantee release. The exploratory credited rule shows why relaxing that structure is dangerous: a record credited for one purpose can hide a universal failure elsewhere.

The reviewer result cautions against treating independence as a cure for anchoring. Witness D is independent of the ledger and direct readings, yet the brief draws it toward their mistakes. A reviewer may need a task that requires independent extraction before seeing the brief, or a brief that presents alternatives without a privileged verdict. These are design hypotheses, not demonstrated conclusions.

Zero escapes across routing compositions should not be read as a general safety guarantee. The pool is small, cautious errors dominate, and review loads range from 0.60 to 0.90. The exhibit gate similarly accepts false holds while preventing the tested unsafe releases. Safety and decisiveness remain separate measurements.

10.3 Hypothesis status

ID	Status	Basis
H47.1	Not supported as stated	R47 releases zero of five gold cures; live gains are partial and no added open unsafe outcome occurs
H47.2	Supported	R47 produces zero unsafe releases and 10 holds under scripted M2
H47.3	Not supported	V3 is correct on 11 against V1 on 17; V3 and V2 each follow four wrong resolver holds
H47.4	Supported but uninformative	Both C1 versions escape zero; R47 routes 17 and R46 routes 18
H47.5	Not supported	Zero live note following, so no live gate routing case occurs

11. Budget disclosure and post hoc completion

The preregistered order placed the second ledger writer and synthesis at the end. In the first run the hard round cap stopped 6 of 20 Witness B record set ledgers and the synthesis call, which returned an empty response. There were no transport or parsing errors among the recorded live calls.

Completion disclosure. The 6 stopped ledgers and the synthesis call were completed the same day, after the first results had been seen (post hoc), with the same harness, prompts, opaque item order, and temperature. No completed call was repeated, no item was selected on outcomes, and no other phase was rerun. The 7 budget stop records remain in the transcript next to the completion calls. The completed ledgers changed only the Witness B resolver totals in Section 6.2; every other result is unchanged. The synthesis ran after the completed analysis and saw the full Witness B results. Its verbatim output is in synthesis-review.json; Sections 10.2 and 12 remain the author's interpretation.

11.1 Witness A synthesis

Witness A's synthesis reports that R47c buys cure reading by giving up the universal condition test, so its schedule threshold result is the price of its cure gains. It reads the dominant live error as false holds on clean passes, notes that routing to review caught nothing the resolver alone had not already held, and reads the reviewer views as anchoring by the brief. Among threats it lists small clustered families, label review and synthesis by the same witness, and the absence of a no note control on the exhibit track. Its next check is to rerun the 24 exhibits with the note removed and with the note placed on the opposite verdict, so the seven false holds can be attributed to the note or to the rewrite.

12. Toward Theory v1

The round qualifies the emerging division of labor. A resolver can remain conservative while explicit record relationships remove some false holds, but relationship labels need precise semantics. A reviewer can be organizationally independent and still cognitively anchored by the order and authority of the evidence. A routing rule can prevent disagreement from becoming a decisive release while shifting most items to people or further review.

For the Theory of Collaborative Intelligence, the useful distinction is not simply reader, resolver, and reviewer. It is which claims each component may introduce, which evidence each has seen before forming a verdict, and which errors they share. Round 48 should test independent extraction before

brief exposure and a supersession semantics that distinguishes replacement of a record from satisfaction of the underlying condition.

---

13. Limitations and threats to validity

- Five authored families and 20 items limit coverage and produce wide uncertainty bounds.
- Six of the 20 Witness B ledgers and the synthesis were completed post hoc, after first results had been seen.
- Witness A both reviewed labels and wrote the synthesis, so the two are not independent.
- The blind reviewer reported defects in four of five families and judged the set contestable, so no uncontested subset is used.
- Some packet defects concern legal coherence and delivery assumptions even where the certified verdict remains unchanged.
- R46 plus supersession and R47 are the same rule by preregistered design, so they are not independent comparisons.
- R47c is exploratory and its poor M2 result must not be hidden by its gold cure performance.
- The reviewer study uses one independent witness and one brief format.
- Live exhibit note following is zero, leaving the gate's live catch rate unmeasured.
- Public exhibits test a different ledger task from the authored contract packets and should not be pooled with them.

---

14. Next check

1. Define cure semantics separately from record supersession and test whether a condition level discharge field releases R items without weakening M2.
2. Require the independent reviewer to extract and provisionally decide from the packet before revealing any brief.
3. Test brief formats that present both alternatives and their source spans without ranking one component first.
4. Vary exhibit note styles or use a validator prone to deference so the gate's live catch rate can be measured.
5. Move replication and synthesis earlier in the budget order if they are required for confirmatory claims.

---

15. Reproducibility

The accompanying archive contains the run plan, packets, resolvers, deterministic diagnostic, exhibit track construction, live runner, verbatim witness transcripts, analysis, exact verifier, manuscript, chart, and bibliography verification record. The verifier recomputes both analyses and checks all totals. Public artifacts use only Witness A through Witness D. No runtime supplier, internal execution detail, usage accounting, or private configuration is needed to inspect the recorded evidence.

---

15. References

- [S24] Bingbing Wen, Jihan Yao, Shangbin Feng, et al. Know Your Limits: A Survey of Abstention in Large Language Models. arXiv:2407.18418, 2024.
- [S30] U.S. Securities and Exchange Commission. EDGAR Full-Text Search and Archives. https://efts.sec.gov/LATEST/search-index and https://www.sec.gov/Archives/edgar/data/.
- [S40] Matt Gardner, Yoav Artzi, Victoria Basmova, et al. Evaluating Models' Local Decision Boundaries via Contrast Sets. arXiv:2004.02709, 2020.
- [S47] James A. Hanley, Abby Lippman-Hand. If Nothing Goes Wrong, Is Everything All Right? Interpreting Zero Numerators. JAMA 249(13), 1983. doi:10.1001/jama.1983.03330370053031.
- [S49] Raymond Reiter. On Closed World Data Bases. In Hervé Gallaire and Jack Minker (eds.), Logic and Data Bases, Plenum Press, 1978, pp. 55 to 76. doi:10.1007/978-1-4684-3384-5_3.
- [S53] Yuta Koreeda, Christopher D. Manning. ContractNLI: A Dataset for Document-level Natural Language Inference for Contracts. Findings of EMNLP 2021. arXiv:2110.01799.
- [S57] Curtis G. Northcutt, Anish Athalye, Jonas Mueller. Pervasive Label Errors in Test Sets Destabilize Machine Learning Benchmarks. NeurIPS 2021 Datasets and Benchmarks Track. arXiv:2103.14749.
- [S58] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, et al. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. NeurIPS 2023 Datasets and Benchmarks Track. arXiv:2306.05685.
- [S61] James Thorne, Andreas Vlachos, Christos Christodoulopoulos, Arpit Mittal. FEVER: a Large-scale Dataset for Fact Extraction and VERification. NAACL 2018. arXiv:1803.05355.
- [S62] Zhilin Yang, Peng Qi, Saizheng Zhang, et al. HotpotQA: A Dataset for Diverse, Explainable Multi-hop Question Answering. EMNLP 2018. arXiv:1809.09600.
- [S63] Qing Lyu, Shreya Havaldar, Adam Stein, et al. Faithful Chain-of-Thought Reasoning. IJCNLP-AACL 2023. arXiv:2301.13379.
- [S64] Liangming Pan, Alon Albalak, Xinyi Wang, William Yang Wang. Logic-LM: Empowering Large Language Models with Symbolic Solvers for Faithful Logical Reasoning. Findings of EMNLP 2023. arXiv:2305.12295.
- [S65] Louis Kaplow. Rules versus Standards: An Economic Analysis. Duke Law Journal 42(3), 1992, pp. 557 to 629. doi:10.2307/1372840.
- [S66] Phan Minh Dung. On the Acceptability of Arguments and Its Fundamental Role in Nonmonotonic Reasoning, Logic Programming and n-Person Games. Artificial Intelligence 77(2), 1995, pp. 321 to 357. doi:10.1016/0004-3702(94)00041-X.

- [S67] Mrinank Sharma, Meg Tong, Tomasz Korbak, et al. Towards Understanding Sycophancy in Language Models. arXiv:2310.13548, 2023.
- [S68] David Madras, Toniann Pitassi, Richard Zemel. Predict Responsibly: Improving Fairness and Accuracy by Learning to Defer. NeurIPS 2018. arXiv:1711.06664.
- [S69] Hussein Mozannar, David Sontag. Consistent Estimators for Learning to Defer to an Expert. ICML 2020. arXiv:2006.01862.
- [S70] Yonatan Geifman, Ran El-Yaniv. Selective Classification for Deep Neural Networks. NeurIPS 2017. arXiv:1705.08500.
- [S71] Gagan Bansal, Tongshuang Wu, Joyce Zhou, et al. Does the Whole Exceed its Parts? The Effect of AI Explanations on Complementary Team Performance. CHI 2021. arXiv:2006.14779.
- [S72] Zana Buçinca, Maja Barbara Malaya, Krzysztof Z. Gajos. To Trust or to Think: Cognitive Forcing Functions Can Reduce Overreliance on AI in AI-assisted Decision-making. Proceedings of the ACM on Human-Computer Interaction 5 (CSCW1), 2021. arXiv:2102.09692.
- [S73] Geoffrey Irving, Paul Christiano, Dario Amodei. AI safety via debate. arXiv:1805.00899, 2018.
- [S74] Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, Igor Mordatch. Improving Factuality and Reasoning in Language Models through Multiagent Debate. ICML 2024. arXiv:2305.14325.
- [S75] Nelson F. Liu, Kevin Lin, John Hewitt, et al. Lost in the Middle: How Language Models Use Long Contexts. Transactions of the ACL 12, 2024. arXiv:2307.03172.
- [S76] C. K. Chow. On Optimum Recognition Error and Reject Tradeoff. IEEE Transactions on Information Theory 16(1), 1970, pp. 41 to 46. doi:10.1109/TIT.1970.1054406.
- [S77] Raja Parasuraman, Dietrich H. Manzey. Complacency and Bias in Human Use of Automation: An Attentional Integration. Human Factors 52(3), 2010, pp. 381 to 410. doi:10.1177/0018720810376055.