Contestable Reasoning Structures for Human-AI Teaming in SOC Investigations

Despoina Giarimpampa
SnT, University of Luxembourg

Roland Meier
Cyber-Defence Campus, armasuisse

Tegawendé F. Bissyandé
SnT, University of Luxembourg

Vincent Lenders
SnT, University of Luxembourg

Jacques Klein
SnT, University of Luxembourg

Abstract

SOCs are deploying LLMs faster than they are deciding how analysts should work with them. The conversational interface that dominates current deployments returns an answer while withholding the investigative state behind it, so analysts can neither contest individual inferences nor reweight evidence nor hand an investigation across shifts. We argue that human-AI teaming in the SOC should be designed around the model's reasoning structure, a schema-bound representation of its working hypotheses, evidence mappings, unresolved uncertainties, and pruned alternatives, rather than around its final answer. We sketch a minimal schema for this structure, instantiate it on an authentication-anomaly scenario, and identify design trade-offs concerning handover compression, trace persistence, confidence rendering, tier placement, and agentic deployment. Throughout, the model proposes investigative structure while the analyst retains authority over how that structure is interpreted and acted upon. Reasoning transparency may improve trust calibration and operational effectiveness, and we frame reasoning structures as objects for co-design and evaluation so that the community can establish the conditions under which that potential is realised.

1 Introduction

Large Language Models (LLMs) have entered Security Operations Centers (SOCs). Recent empirical studies [11, 13] report that analysts adopt them primarily for cognitive support, including contextualisation of alerts, translation between artifact formats, and sense-making during investigations, while keeping decisions in their own hands. As capabilities grow and agentic deployments expand, many organisations still debate whether LLMs belong in the SOC at all. Wherever such systems are already in place, a second question becomes urgent: how should analysts and models collaborate?

Current deployments inherit the conversational interface of consumer chatbots, and that interface fits SOC investigations poorly. An investigation is an extended episode of hypothesis maintenance under uncertainty. Artifacts arrive asynchronously, relevance shifts as evidence accumulates, and the defensibility of an escalation depends on the path taken as much as on the verdict [13]. A single question followed by a single answer captures little of this work. When the model exposes only its answer, the analyst inherits a conclusion stripped of the reasoning that produced it. Two failure modes follow. Cautious analysts discount inferences they are unable to inspect, and overworked analysts accept narratives they lack the time to verify [1, 3, 12, 14].

Position. We argue that human-AI teaming in SOCs should be designed around the model's reasoning structure rather than its answer. By reasoning structure we mean a schema-bound representation of the model's case-relevant commitments, comprising hypotheses, evidence mappings, unresolved uncertainties, escalation conditions, and pruned alternatives. This differs both from a post-hoc explanation appended to a final answer and from a transcript of the model's private reasoning. The structure serves as a teaming primitive, the object analysts read, contest, amend, and hand off. Our answer to the question of who gets to decide is that the model may propose structure while the analyst retains authority over how that structure is interpreted and acted upon.

Scope. We develop the argument inside Tier-2 SOC investigations, where the cost of opacity is sharpest and where embedded fieldwork suggests LLM tools succeed when they augment existing workflows rather than displace them [6]. We focus on SOCs for concreteness, and we return to the broader reach of the argument in the conclusion.

Workshop on Designing Human–AI Teaming for Security (HATS), co-located with the Symposium on Usable Privacy and Security (SOUPS) 2026.
August 23, 2026, Hannover, Germany.

2 The Teaming Problem in SOC Investigations

What analysts do. Analysts maintain several competing explanations for a set of observed artifacts and update their confidence in each as new artifacts arrive. A benign anomaly and an early-stage intrusion often share the same initial footprint, so the discriminating evidence accumulates over hours and across data sources. The value of the work therefore lies in the defensible path from artifacts to escalation, a path that peers, auditors, and the next shift must be able to retrace, rather than in the escalation verdict alone [12].

Why current LLM interaction breaks teaming. A conversational LLM compresses this contingent process into a fluent narrative. The analyst sees the conclusion while the hypothesis space, the evidence weighting, and the discarded alternatives stay hidden inside the prose. This reproduces, in conversational form, the same opacity that has long limited the uptake of black-box detection systems [5]. Chain-of-thought prompting [15] elicits reasoning text yet falls short of addressable structure, because a hypothesis embedded in a paragraph is an entity the shared investigative state can neither reference nor track. The Human-Computer Interaction (HCI) literature on contestable AI converges on the same conclusion from a different direction: systems become contestable when the output itself is structured so that intermediate commitments can be challenged individually [2]. What SOC teaming needs, in short, is for the model's reasoning to become an object in the workflow rather than a story about it.

3 Reasoning Structure

If reasoning is to be addressable, it must be represented as more than a paragraph of explanation. Figure 1 sketches a minimal reasoning structure. Evidence maps to hypotheses through a typed role, marking each artifact as supporting, conflicting, or ambiguous with respect to a given hypothesis. Pruned alternatives remain linked to the active hypotheses so they can be revived when new evidence warrants. An escalation decision is tied to the specific hypotheses and evidence that justify it. The schema is minimal in a precise sense: each element enables an analyst action that a narrative answer fails to support. Structurally, the schema parallels the Analysis of Competing Hypotheses (ACH) [8], a structured analytic technique with a long history in intelligence work. Our move is to treat ACH-style structure as a model output schema rather than an analyst worksheet, so that the model's commitments become objects the analyst can amend.

The schema is the interface itself, and any narrative summary is a rendering of the structured reasoning rather than a replacement for it. This inverts the usual arrangement in which an explanation is appended to an answer. Because the structure decomposes the model's reasoning into separately addressable commitments, the analyst can intervene locally

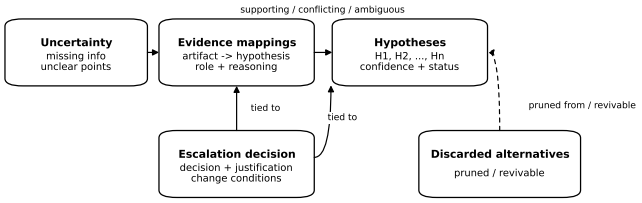

Figure 1: The reasoning structure. The model output is a set of objects that analysts can amend, contest, hand off, or audit, rather than a final answer with an appended explanation.

instead of accepting or rejecting a whole response. She can contest an evidence mapping, amend a confidence band, preserve an unresolved uncertainty, or revive a pruned alternative, each as an isolated edit.

Three properties follow from this design. The structure is *addressable*, since analysts can point to specific hypotheses, evidence roles, or escalation conditions. It is *persistent*, since it can be handed off across shifts, attached to a ticket, or consumed by an after-action report. It is *contestable*, since discarded alternatives remain visible and analysts can inspect why an explanation was ruled out rather than inheriting the conclusion [2]. The third property is the one that distinguishes a teammate from a tool.

4 Example Scenario

We instantiate the schema on a synthetic authentication-anomaly scenario whose features reflect patterns reported in public incident advisories, in which a low-severity alert becomes meaningful only after correlated discovery activity arrives. The interest of the scenario lies in the reasoning structure an analyst should see at each point, rather than in the incident itself.

At **02:14**, the SIEM raises a low-severity alert reporting a successful authentication for user j.doe from an out-of-region IP, preceded by three failed attempts. Given these artifacts, the model populates the hypothesis field of the reasoning structure by matching the observed pattern against common authentication-anomaly explanations. It assigns low initial confidence to three candidates: benign travel or VPN variance (H1), credential stuffing (H2), and targeted credential compromise (H3). These hypotheses serve as initial candidates to be revised as additional artifacts arrive rather than as conclusions. The failures-then-success pattern weakly supports the two malicious explanations, and the absence of a travel notification weakly conflicts with the benign one.

At **02:41**, a Kerberos TGS request to a never-accessed file server and an LDAP enumeration burst arrive from the same session. The compromise hypothesis is promoted, the benign and credential-stuffing hypotheses are demoted, and a misconfigured-audit-policy explanation is pruned yet re-

Table 1: Evolution of the reasoning structure in the example scenario.

Time	New artifact	Reasoning update
02:14	New-region login after failed attempts	H1, H2, H3 remain low confidence; endpoint telemetry missing
02:41	Kerberos request to unseen file server; LDAP enumeration burst	H3 promoted; H1/H2 demoted; escalation condition partially met
Pending	Endpoint telemetry	Confirms, revises, or reopens the compromise hypothesis

tained in the structure. The model recommends Tier-2 escalation with a stated rationale, while the analyst remains free to accept, reject, or hold pending endpoint telemetry. Table 1 summarises the reasoning update.

What the example illustrates is where the analyst intervenes. At **02:43**, the on-shift analyst notices that the LDAP burst overlaps with a scheduled vulnerability scan recorded in the change-management logs. She reclassifies the burst as ambiguous, revives the audit-policy hypothesis, and holds the escalation pending endpoint telemetry. Each intervention is a local edit to a specific field, namely an evidence role, a pruned hypothesis, and an escalation condition. In a paragraph response these commitments stay embedded in prose; in the reasoning structure they are separate objects that can be edited, tracked, and handed over to the next shift.

5 One Possible Implementation Path

A first implementation can proceed without exposing free-form reasoning and without depending on unconstrained prompting. It would require the model to emit a constrained schema whose fields correspond to hypotheses, evidence roles, uncertainties, escalation conditions, and pruned alternatives, with each field validated against the artifacts available in the case. The resulting structure would then be integrated into existing ticketing and handover workflows, where analysts edit fields in place rather than respond to a paragraph. If training or adaptation is used to improve schema quality, it should be evaluated against evidence grounding and contestability rather than schema completeness alone, since a model can fill every field while anchoring few of them in the case at hand. Whether the resulting structure actually helps the team is an empirical question, settled only by evaluation under realistic tier, time-pressure, and override-norm conditions.

6 Decision Trade-offs

We frame the following trade-offs as open questions for co-design rather than positions we defend. Each asks where decision authority should sit between analyst, model, interface, and organisation, and each remains unresolved. The relevant evaluation metrics depend on which trade-off dominates a given deployment (Table 2). Handover compression points to handover quality, confidence rendering to over- and under-reliance, and tier placement to cognitive load. We state our current lean on each question to make the disagreement concrete, while leaving the questions open for the workshop to take up.

Table 2: Evaluation categories for reasoning structures in SOC investigations.

Metric	What it tests
Evidence-grounding accuracy	Whether hypotheses are correctly tied to artifacts via supporting, conflicting, or ambiguous roles.
Contestability	Whether analysts can locate and revise specific model commitments rather than accept or reject the whole output.
Appropriate, over-, and under-reliance [4, 10]	Whether analysts accept useful inferences and reject unsupported ones.
Handover quality	Whether an incoming analyst can reconstruct the investigation state from the structure without a free-form narrative.
Time-to-defensible-decision	Whether analysts reach escalation, rejection, or hold decisions efficiently while preserving rationale.
Cognitive load (NASA-TLX [7])	Whether the structure aids reasoning under pressure or imposes interface burden.

Q1. Handover compression. At shift change, should the structure be (a) a compressed projection, showing active hypotheses with confidence bands, the artifacts driving the top hypothesis, and unresolved uncertainty, with the full schema available on demand, or (b) the full schema, with analyst-driven collapse? **Authority at stake:** the incoming analyst, who inherits whatever the outgoing analyst's tool chose to expose. **Our lean:** (a), because the first task at handover is orientation, and audit comes later.

Q2. Persistence boundary. Should the trace (a) survive into post-incident review and audit, or (b) remain ephemeral and be discarded at ticket close? **Authority at stake:** the organisation, set against the analyst's interest in being able to think out loud without it being held against them. **Our lean:** (a), with the caveat that persistent traces raise real discoverability risks in later litigation, regulatory review, and HR proceedings.

Q3. Confidence rendering. Should the interface (a) display confidence bands directly, or (b) split the display into "model is confident" and "model has evidence," showing both? **Authority at stake:** the interface, which can manufacture or defuse false precision before the analyst sees it. **Our lean:** (b). A band rendered as "moderate to high" looks more authoritative than the same claim hedged in prose, and under time pressure that gap is the trust-calibration failure mode worth designing against. This is the trade-off most directly about automation bias.

Q4. Tier placement. Should the structure enter (a) at Tier-2, where analysts have the room to push back, or (b) at Tier-1, where volume creates the largest efficiency gain? **Authority at stake:** the analyst, whose ability to contest scales inversely with throughput pressure. **Our lean:** (a). The cost of error

is highest at Tier-2, and Tier-2 analysts are best equipped to contest. Pushing the same structure into Tier-1 prematurely risks automation bias dressed up as transparency.

Q5. Agentic deployment. When the model can act as well as advise, should the structure (a) record actions taken, or (b) gate destructive actions such as containment and account disablement while letting reversible ones such as enrichment and correlation proceed? **Authority at stake:** the analyst, whose review window collapses to zero when the system acts first. **Our lean:** (b). Gate the irreversible, trace the reversible.

7 Risks to the Position

The position above can fail in several ways, and we list those we take most seriously.

R1. Transparency without trust calibration. Structured reasoning may look authoritative and thereby increase over-reliance under time pressure, the opposite of the intended effect [1, 12, 14]. Visible structure signals diligence whether or not the underlying inferences deserve it.

R2. Schema mismatch with analyst practice. Analysts vary in how they prune hypotheses and express uncertainty. A schema convenient for the model may prove procrustean for the analyst, and a schema tuned to one SOC may fail to transfer across maturity levels.

R3. Structure compliance without grounding. A model may learn to populate the required fields while linking hypotheses to evidence incorrectly, producing a complete-looking schema on a weak evidential basis. Evidence-grounding accuracy therefore belongs at the centre of evaluation, ahead of schema completeness.

R4. Integration cost. A reasoning structure that fits poorly into existing ticketing and handover flows gets bypassed. Analysts under pressure default to the tools that already work.

R5. Data-governance limits on evaluation. In-boundary deployment constraints limit cross-organisational benchmarks [9], which in turn limits how far findings from one SOC generalise to others.

8 Conclusion

We argue that exposing reasoning structure can make LLMs better SOC teammates by turning model output into a shared object that analysts inspect, contest, amend, and hand off. The claim rests on more than transparency alone; it rests on whether the structure preserves analyst authority while improving evidence grounding, reliance calibration, and handover quality. An implementation path should therefore prioritise constrained, schema-based outputs integrated into existing ticketing and handover workflows over unconstrained conversational explanations. The open question, and the one we bring to this workshop, is under which interface, organisational, and deployment conditions this design improves SOC work. The same question recurs wherever security work involves contestable judgment under uncertainty, from review of AI-generated code to incident-response triage: what does the AI expose, and what does the human get to amend?

References

[1] Daehwan Ahn, Abdullah Almaatouq, Monisha Gulabani, and Kartik Hosanagar. Will we trust what we don't understand? impact of model interpretability and outcome feedback on trust in ai. *arXiv preprint*, 2021.

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

[3] Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel Weld. Does the whole exceed its parts? the effect of ai explanations on complementary team performance. In *Proceedings of the 2021 CHI conference on human factors in computing systems*, pages 1–16, 2021.

[4] Zana Buçinca, Maja Barbara Malaya, and 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):1–21, 2021.

[5] Despoina Giarimpampa, Roland Meier, Tegawendé F. Bissyandé, Vincent Lenders, and Jacques Klein. Exploring the role of artificial intelligence in enhancing security operations: A systematic review. *ACM Computing Surveys*, 58(3):1–38, 2025.

[6] Francis Hahn, Mohd Mamoon, Alexandru G. Bardas, Michael Collins, Jaclyn Lauren Dudek, Daniel Lende, Xinming Ou, and S. Raj Rajagopalan. Non-disruptive disruption: An empirical experience of introducing LLMs in the SOC. In *Workshop on Security Operations Center Operations and Construction (WOSOC)*, 2026.

[7] Sandra G. Hart and Lowell E. Staveland. Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In *Human Mental Workload*, volume 52 of *Advances in Psychology*, pages 139–183. Elsevier, 1988.

[8] Richards J Heuer. *Psychology of intelligence analysis*. Center for the Study of Intelligence, 1999.

[9] Inna Kouper. Data sharing and use in cybersecurity research. *Data Science Journal*, 23, 2024.

[10] John D. Lee and Katrina A. See. Trust in automation: Designing for appropriate reliance. *Human Factors*, 46(1):50–80, 2004.

[11] Martin Lochner and Ksenia Keplinger. Collaborative intelligence: Topic modelling of large language model use in live cybersecurity operations. *arXiv preprint*, 2025.

[12] Iqbal H. Sarker, Helge Janicke, Ahmad Mohsin, Asif Gill, and Leandros Maglaras. Explainable ai for cybersecurity automation, intelligence and trustworthiness: Methods, taxonomy, and challenges. *ICT Express*, 10:935–958, 2024.

[13] Ronal Singh, Shahroz Tariq, Fatemeh Jalalvand, Mohan Baruwal Chhetri, Surya Nepal, Cecile Paris, and Martin Lochner. Llms in the soc: An empirical study of human-ai collaboration in security operations centres. *arXiv preprint arXiv:2508.18947*, 2025.

[14] Emily Wall. Trust junk and evil knobs: Calibrating trust in ai. In *IEEE Pacific Visualization Conference (PacificVis)*, 2024.

[15] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.