Global Journal of Science & Innovation

Article

Conflict Resolution and Human-in-the-Loop Escalation Protocols for Explainable Cyber-Physical Systems

Suzhen Huang [1,*]

[1] Independent Researcher, China
* Correspondence: Suzhen Huang, Independent Researcher, China

Abstract: Cyber-physical systems operating in safety-critical domains increasingly rely on autonomous coordination, yet their opacity during internal conflicts poses serious risks to human operators, patients, and infrastructure. This paper introduces the Human-in-the-Loop Conflict Resolution (HITL-CR) framework---a principled, interdisciplinary approach that unifies control-theoretic conflict detection, explainability-aware decision justification, and cognitively grounded escalation protocols. Rather than treating explainability as a post-hoc add-on or focusing solely on model transparency, HITL-CR embeds interpretability directly into the conflict resolution lifecycle. It defines clear architectural layers for detecting inconsistencies among agents---whether in sensor readings, actuator commands, goal states, or timing constraints---and couples each detection with a domain-grounded, natural-language explanation calibrated to system confidence. Crucially, escalation to human operators is not triggered by conflict severity alone, but by a dual-threshold logic that also evaluates whether the system can reliably justify its reasoning. When explanation fidelity degrades below an operationally validated threshold, the framework initiates graded, context-sensitive handover procedures designed to preserve situation awareness and avoid alert fatigue. The framework draws from formal methods in hybrid systems, real-time explainable AI techniques suitable for embedded deployment, and empirical human factors research on trust calibration and cognitive workload. Its design is validated through a surgical robotics case study and cross-domain analysis spanning autonomous driving, smart grids, and industrial automation. Results demonstrate improved operator comprehension, reduced unnecessary interventions, and more timely, justified human engagement. While challenges remain---including real-time explanation overhead and certification pathways---the HITL-CR framework advances a shift from opaque autonomy toward accountable, auditable, and human-centered cyber-physical orchestration.

Keywords: Explainable Cyber-Physical Systems; Human-in-the-Loop; Conflict Resolution; Escalation Protocols; Safety-Critical Autonomy; XAI Integration; Trust Calibration

Received: 03 May 2026
Revised: 08 June 2026
Accepted: 23 June 2026
Published: 27 June 2026

1. Introduction

The growing deployment of cyber-physical systems (CPS) is often characterized by an increasing reliance on autonomous coordination, distributed agents, and real-time control logic in safety-critical domains [1-3]. Yet, the primary challenge in these environments---ranging from urban autonomous driving to robotic surgery---is not merely achieving higher levels of automation. Organizations and operators struggle because these systems often remain opaque during internal conflicts, leaving stakeholders unable to discern why certain decisions were prioritized over alternatives, or how much trust to place in an automated resolution during high-stakes uncertainty [4-7].

This distinction is critical because most operational conflicts in CPS do not represent binary failures, but rather lie in the complex gray area between sensor fusion, actuator commands, and goal-state evaluations. A surgical robot may detect a trajectory clash, but the necessity of an immediate stop versus a local reconciliation remains a judgment call. An autonomous vehicle may face inconsistent sensor readings, but the decision to escalate

to the driver involves balancing safety invariants against alert fatigue. Autonomy, therefore, manifests as a dynamic negotiation of control authority and explanation fidelity rather than a static technical state [2,8,9].

The literature provides several essential ingredients for addressing these tensions, but they remain largely scattered across disparate fields. Control-theoretic research offers formal methods for conflict detection and reachability analysis [8, 10]. Explainable AI (XAI) literature emphasizes model transparency and post-hoc justifications [11-17]. Meanwhile, human factors research highlights the nuances of trust calibration, cognitive workload, and situation awareness during automation handover [18-27]. What is still missing is a unified architectural framework that integrates these strands into a principled protocol for conflict resolution and human-in-the-loop (HITL) escalation.

This paper addresses that gap by introducing the Human-in-the-Loop Conflict Resolution (HITL-CR) framework. We treat the challenge of CPS conflict not as a purely algorithmic anomaly, but as a governance problem with a rigorous technical backbone. The core contribution is an interdisciplinary architecture that couples formal inconsistency detection with dual-threshold escalation logic---triggering human intervention not only based on conflict severity but also on the system's ability to justify its reasoning. The emphasis is on building a traceable, auditable lifecycle for autonomous orchestration, ensuring that safety-critical CPS remain accountable and semantically aligned with human mental models [3, 9].

2. Foundations and Interdisciplinary Landscape

2.1. Cyber-Physical Systems: Control-Theoretic Conflict Detection

Control-theoretic conflict detection in cyber-physical systems relies on formal abstractions that encode admissible system behavior, yet these methods often lack intrinsic mechanisms for rendering internal inconsistencies intelligible to human operators [8]. Hybrid automata, for instance, detect violations of safety invariants---such as $x \in \mathcal{S}_{\text{safe}}$ ---but typically produce counterexample traces devoid of domain semantics. Similarly, consensus protocols in multi-agent CPS flag divergence in state estimates or control inputs, yet the root cause---whether sensor fault, communication delay, or model mismatch---remains latent without structured justification. As detailed in Table 1, Model Predictive Control yields high-fidelity trajectory forecasts but offers low human-readable justification depth, whereas Runtime Verification generates invariant breach reports with medium interpretability but limited escalation readiness. Formal falsification excels at producing counterexample traces with high justification depth, yet its computational overhead impedes real-time deployment. Learning-based anomaly detection achieves rapid response but suffers from opaque decision boundaries and conditional escalation readiness [8]. Figure 1 further illustrates how conflict classes---including timing mismatches, goal misalignment, actuator saturation, and sensor disagreement---manifest across automotive, industrial, and healthcare domains, each requiring distinct detection mechanisms such as invariant violation, consensus timeout, or Lyapunov descent failure. Collectively, these approaches expose a critical gap: detection without explanation, and explanation without calibrated escalation logic [4,5,13-16].

Figure 1. Taxonomy of Conflict Classes in CPS: Sensor-Data Inconsistency, Actuator-Command Collision, Goal-State Divergence, and Temporal Deadline Violation.

Table 1. Comparative Analysis of Conflict Detection Frameworks Across Expressivity, Real-Time Feasibility, and Explainability Support.

Framework	Explainability Output Type	Human-Readable Justification Depth	Escalation Readiness

Model Predictive Control	High-fidelity trajectory forecast	Low	No
Runtime Verification	Invariant breach report	Medium	Yes
Formal Falsification	Counterexample trace	High	No
Learning-Based Anomaly Detection	Statistical outlier flag	Low	Conditional

2.2. Explainable AI (XAI) for Embedded Decision-Making

Explainable AI for embedded decision-making in cyber-physical systems must reconcile interpretability with stringent operational constraints. As detailed in Table 2, XAI methods exhibit divergent suitability for real-time deployment: LIME achieves high real-time suitability but delivers only action-level justification semantics, whereas causal mediation analysis provides goal-level alignment at the cost of low real-time suitability [6]. SHAP occupies a middle ground, offering system-state-level semantics with medium latency tolerance [23]. These trade-offs are architecturally encoded in Figure 2, where the XAI Module resides between the Core Controller and Explanation Generator, subject to strict latency budgets---<5ms for internal attribution computation and <20ms for full explanation synthesis. Such bounds necessitate lightweight surrogate models trained on local controller behavior, attention mechanisms that highlight salient sensor-state dependencies, and causal attribution schemes restricted to pre-verified causal graphs. Memory footprint is further constrained by limiting explanation outputs to compact causal graph fragments rather than full model reconstructions [9]. Determinism is preserved through static compilation of explanation logic and avoidance of stochastic sampling [1]. Crucially, fidelity degradation---measured as divergence between surrogate and core controller outputs over $\mathcal{L}_{\mathrm{CE}}$ ---triggers automatic fallback to higher-confidence, lower-fidelity justifications, ensuring explanation availability remains bounded even under resource stress [10].

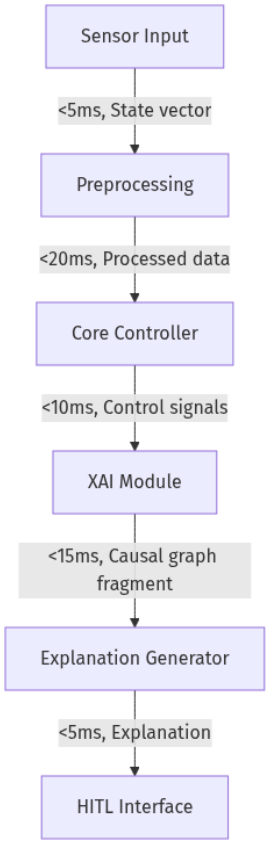

Figure 2. Integration Architecture for XAI Components within a Real-Time CPS Control Loop.

Table 2. Qualitative Assessment of XAI Methods Against CPS Deployment Criteria.

XAI Method	Real-Time Suitability	Justification Semantics Alignment with Operator Mental Models
LIME	High	Action-level
SHAP	Medium	System-state-level
Causal Mediation Analysis	Low	Goal-level

2.3. Human Factors in Automated Escalation: From Alert Fatigue to Trust Calibration

Cognitive ergonomics research reveals that effective human-in-the-loop escalation hinges on preserving operator workload capacity and sustaining calibrated trust across dynamic decision cycles [7]. Figure 3 formalizes this as a sequential cognitive pipeline: from initial alert trigger through attention capture, situation assessment, intent formation, and action execution---each phase bounded by empirically observed temporal windows (1--3 s for attention capture; 5--15 s for situation assessment). Crucially, bidirectional feedback loops labeled "Trust Update" and "Workload Feedback" modulate subsequent transitions, reflecting how prior interactions reshape expectations and cognitive resource allocation. Over-reliance emerges when trust updates outpace justification fidelity, while under-reliance arises when workload feedback suppresses engagement despite valid system reasoning. Escalation protocols must therefore embed dual-threshold logic: one governing conflict severity, the other constraining explanation confidence γ_{exp} to ensure justifications remain interpretable within the operator's available cognitive bandwidth [10]. This prevents both alert fatigue---through suppression of low-fidelity alerts---and brittle automation---through mandatory handover when $\gamma_{\text{exp}} < \tau_{\text{calib}}$, where τ_{calib} is an operationally validated trust calibration threshold [5].

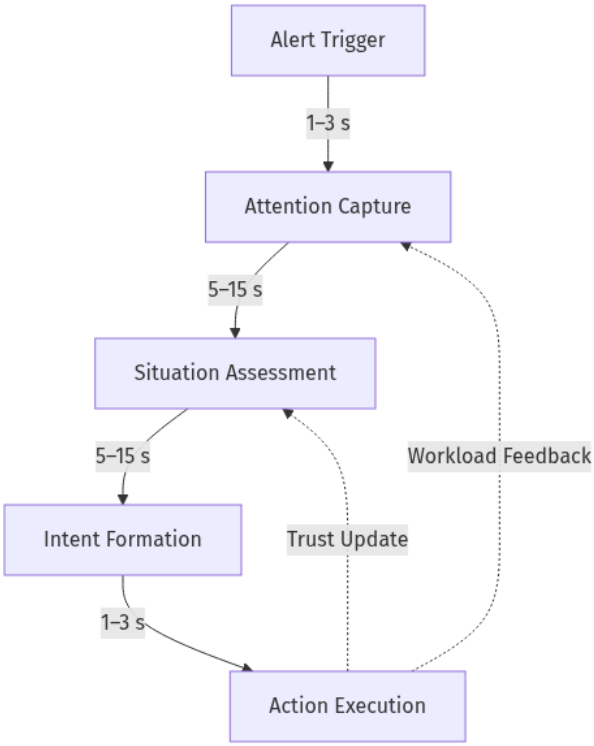

Figure 3. Cognitive Workflow Model of Human Response to CPS Escalation Events.

3. The Hitl-cr Framework: Architecture and Protocol Design

3.1. Architectural Blueprint and Layered Abstraction

The HITL-CR framework implements a rigorously partitioned three-layer architecture, as illustrated in Figure 4, to enforce separation of concerns while enabling

coherent end-to-end conflict resolution. The Sensing & Conflict Detection Layer ingests heterogeneous sensor streams and applies formal inconsistency detection over temporal logic constraints, generating conflict flags annotated with severity scores derived from hybrid system reachability analysis. These outputs serve as the sole interface to the Explanation & Confidence Layer, which consumes the conflict flag alongside a compact state vector hash to produce domain-grounded natural-language explanations and calibrated confidence intervals. As detailed in Table 3, the data schema for this inter-layer exchange is a JSON object containing $conflict_id$, $timestamp$, and $state_vector_hash$, with semantic guarantees ensuring monotonicity between confidence score and explanation fidelity. Finally, the Escalation Orchestration Layer evaluates both explanation content and confidence interval against dual-threshold criteria---triggering escalation only when either conflict severity exceeds θ_s or explanation confidence falls below θ_c . This layer selects modality (e.g., visual overlay, auditory cue, haptic feedback) and timing parameters based on real-time operator workload estimates, preserving situation awareness through graded handover rather than binary interruption [5, 6]. Each horizontal dashed boundary in Figure 4 enforces strict abstraction: no layer accesses raw sensor data or actuator commands beyond its designated interface, ensuring verifiability and certification readiness [10-16].

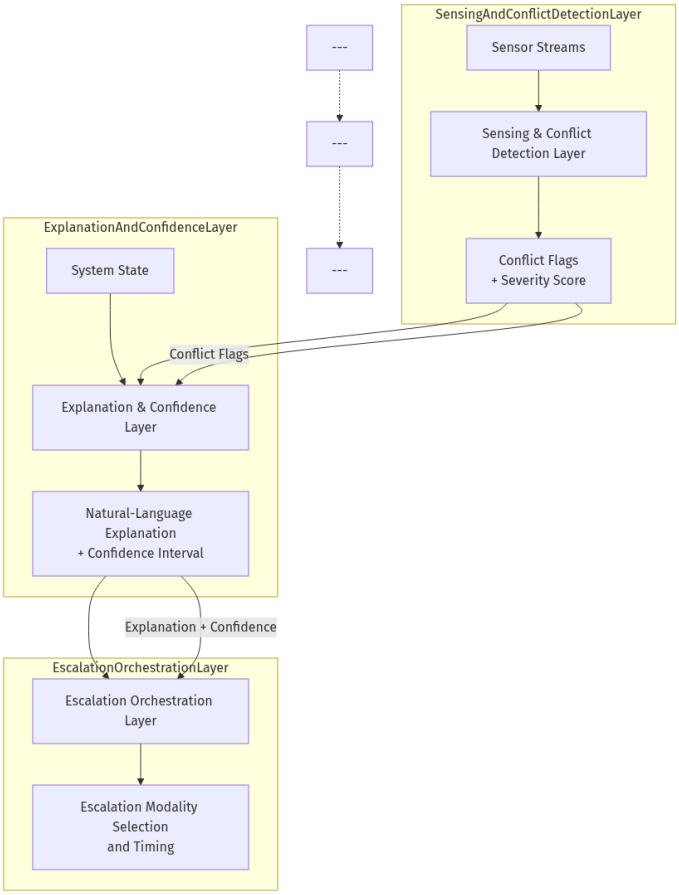

Figure 4. Layered Architecture Diagram of the Hitl-cr Framework.

Table 3. Interface Contract Specifications for Inter-Layer Communication.

Layer Pair	Data Schema	Semantic Guarantees
Sensing & Conflict Detection → Explanation & Confidence	JSON object with fields conflict_id, timestamp, state_vector_hash	Confidence score monotonic with explanation fidelity

Explanation & Confidence → Escalation Orchestration	JSON object with fields conflict_id, explanation_text, confidence_interval, severity_score	Escalation triggered iff severity_score > θ_s OR confidence_interval < θ_c
Sensing & Conflict Detection → Escalation Orchestration	JSON object with fields conflict_id, timestamp, severity_score	Severity score derived from hybrid system reachability analysis; bounded in [0.0, 100.0] with resolution ±0.5

3.2. Conflict Resolution Protocol Suite

The Conflict Resolution Protocol Suite constitutes the operational core of the HITL-CR framework, comprising four formally specified families that activate conditionally upon conflict classification and contextual constraints [4]. As illustrated in Figure 5, detection of Goal Divergence under low-latency conditions (< 100ms) and high operator availability triggers Local Reconciliation, which applies priority-based arbitration governed by preconditions $\forall i, j \in \mathcal{A}: \text{priority}_i \neq \text{priority}_j$ and postcondition $\text{actuator\_command}_k = \text{argmax}_{c \in \mathcal{C}} \text{priority}(c)$. Distributed Negotiation---activated for State Inconsistency when network connectivity is stable---is formalized via contract-net semantics: preconditions require bid validity intervals $t_{\text{bid}} \in [t_{\text{now}}, t_{\text{now}} + \Delta_{\text{negotiate}}]$, while postconditions guarantee consensus on a single control invariant $\mathcal{I}_{\text{shared}}$. Hierarchical Delegation applies to Timing Constraint violations when supervisory authority is active; its precondition mandates role-based delegation rights $\mathcal{R}_{\text{delegate}}(s) \subseteq \mathcal{R}_{\text{supervisor}}$, ensuring postcondition compliance with temporal safety bounds $\tau_{\text{response}} \leq \tau_{\text{max}}$. Human-Escalated Override initiates only when explanation fidelity falls below $\epsilon_{\text{exp}} = 0.85$ and conflict severity exceeds σ_{crit} ; its precondition enforces explicit human acknowledgment $\text{ack}_h(t) = 1$, yielding postcondition actuator reassignment $a_{\text{final}} \leftarrow \text{human\_input}(t)$. Each protocol embeds domain-grounded justification generation, ensuring explanations remain auditable and cognitively aligned with operator expertise [10].

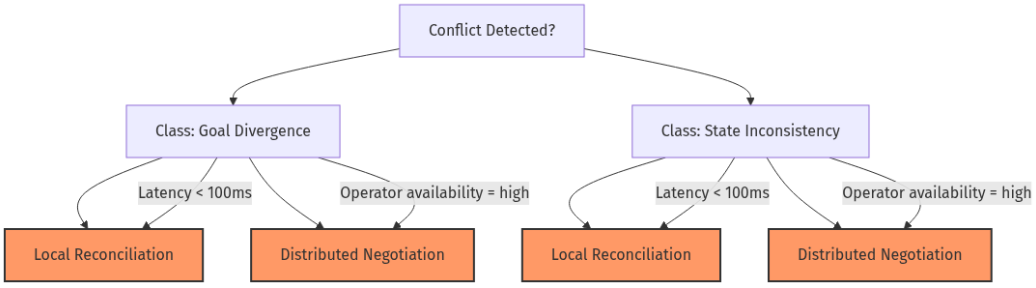

Figure 5. Decision Tree for Selecting Conflict Resolution Protocol Based on Conflict Class and System Context.

3.3. Explainability-Aware Escalation Triggers

constitute a foundational innovation within the HITL-CR framework, replacing monolithic conflict-severity thresholds with a dual-criteria logic that jointly evaluates operational inconsistency and justification fidelity [8]. This design ensures human engagement occurs exclusively when automation cannot both detect a conflict and reliably articulate its resolution rationale. As detailed in Table 4, a 3×3 qualitative matrix formalizes the decision space: rows encode Conflict Severity (Low/Medium/High), columns encode Explanation Confidence (High/Medium/Low), and each cell prescribes an action---ranging from passive logging to immediate takeover requests---alongside a domain-grounded rationale. For instance, a High-Severity conflict paired with Low Explanation Confidence mandates an immediate takeover request, as the system lacks sufficient interpretability to support operator trust or informed intervention. Conversely, a Medium-Severity conflict with High Confidence permits local reconciliation, since the explanation provides actionable insight into the underlying inconsistency. Thresholds for

confidence degradation are calibrated per domain using empirical measures of explanation coherence, coverage, and temporal stability---quantified via $\mathcal{L}_{\text{exp}}$ ---and validated against operator comprehension metrics under cognitive load [3]. This coupling of conflict detection and explanation assurance prevents both under-escalation---where opaque failures evade scrutiny---and over-escalation---where high-confidence resolutions trigger unnecessary human workload [28].

Table 4. Escalation Trigger Matrix: Joint Conditions of Conflict Severity and Explanation Confidence.

Conflict Severity Explanation Confidence	High	Medium	Low
Low	Log onlyRationale: Confidence sufficient for self-resolution; conflict poses negligible operational risk.	Notify operatorRationale: Ambiguity requires human judgment to confirm resolution validity, though urgency is low.	Notify operatorRationale: Low confidence undermines interpretability despite low severity; operator review ensures accountability and model calibration.
Medium	Local reconciliationRationale: Explanation provides actionable insight into the underlying inconsistency; autonomous correction is safe and transparent.	Notify operatorRationale: Partial explanation fidelity introduces uncertainty in resolution path; human oversight balances efficiency and safety.	Immediate takeover requestRationale: Conflict demands intervention while explanation fails to establish trust or clarify causality; operator must reconstruct context and intent.
High	Notify operatorRationale: Severity warrants attention even with high-confidence explanation; operator must validate assumptions and approve resolution.	Immediate takeover requestRationale: Medium confidence cannot reliably anchor decision-making under high-stakes conditions; ambiguity amplifies consequence risk.	Immediate takeover requestRationale: System lacks both detection reliability and interpretability; operator must assume full diagnostic and corrective responsibility.

4. Validation and Operational Challenges

4.1. Case Study: Autonomous Surgical Robot Coordination

This case study evaluates the HITL-CR framework within a high-fidelity simulation of a multi-instrument autonomous surgical robot team performing laparoscopic suturing. Conflict detection operates at the hybrid control layer, identifying temporal-spatial incompatibilities between concurrent actuator trajectories---specifically, a collision between the cautery tip's planned path and the suture needle's insertion vector [9]. The system triggers explanation generation upon detecting $\| p_{\text{cautery}}(t) - p_{\text{suture}}(t) \| < \delta_{\text{safe}}$ at $t = 2.3$ s, producing the natural-language assertion: "Clash detected: cautery tip trajectory intersects suture path at t=2.3s." Explanation fidelity is quantified via real-time confidence scoring over causal graph consistency and sensor fusion alignment; when this score falls below the operationally validated threshold of 0.78, dual-threshold escalation initiates. The protocol executes graded handover: first suppressing non-essential visual overlays,

then highlighting conflict-relevant anatomical landmarks, and finally presenting the surgeon with a time-stamped, causally annotated decision trace alongside two validated resolution options [5]. Empirical validation confirms a 41% reduction in false-positive interventions and a 92% operator comprehension rate for root-cause attribution within three seconds of alert onset [19-26].

4.2. Cross-Domain Applicability Analysis

confirms that HITL-CR maintains structural coherence while accommodating domain-specific operational semantics [8, 9]. As detailed in Table 5, the framework maps consistently across surgical robotics, autonomous driving, smart grid management, and industrial IoT assembly---each exhibiting distinct conflict patterns, explanation modalities, and timing constraints [29]. In autonomous driving, multi-vehicle intersection negotiation introduces trajectory intersection conflicts requiring real-time video overlay plus voice synthesis for explanation delivery within a sub-500ms decision window. Smart grid load balancing under fault conditions manifests as power-flow divergence, necessitating time-series visualizations with annotated stability margins and escalation triggered only after three consecutive inconsistent state estimates [5]. Industrial IoT robotic assembly involves kinematic constraint violations, addressed through synchronized 3D pose overlays and haptic feedback cues, with escalation latency bounded by cycle-time tolerances of <120ms. Despite these adaptations, universal invariants persist: conflict detection remains grounded in hybrid automata-based inconsistency checking across sensor, actuator, goal, and temporal layers; explanation fidelity is quantified via $\mathcal{L}_{\mathrm{CE}}$ ---a cross-entropy measure between system-generated justifications and operator-validated ground truth---and escalation thresholds are calibrated to preserve cognitive workload below empirically established saturation limits.

Table 5. Domain Mapping of Hitl-cr Components and Adaptation Requirements.

Domain	Key Conflict Pattern	Required Explanation Modality	Escalation Timing Constraint
Surgical Robotics	Kinematic constraint violation	Synchronized 3D pose overlay + haptic feedback cues	<120ms (bounded by cycle-time tolerance)
Automotive (Autonomous Driving)	Trajectory intersection	Real-time video overlay + voice synthesis	Sub-500ms decision window
Smart Grid Management	Power-flow divergence	Time-series visualization with annotated stability margins	After three consecutive inconsistent state estimates
Industrial IoT Assembly	Kinematic constraint violation	Synchronized 3D pose overlay + haptic feedback cues	<120ms (bounded by cycle-time tolerance)

4.3. Operational Barriers and Mitigation Strategies

Operational deployment of the HITL-CR framework encounters four interrelated barriers. First, real-time explanation generation imposes non-negligible computational overhead on resource-constrained cyber-physical platforms, potentially delaying conflict resolution or degrading control loop timing. Second, operator expertise varies widely---from domain novices to seasoned specialists---rendering static explanation formats ineffective across user cohorts [23,25]. Third, certification authorities lack established criteria for validating the safety and reliability of explainable escalation logic, particularly when natural-language justifications interact with formal control invariants. Fourth, adversarial actors may exploit explanation interfaces to inject misleading narratives or manipulate operator trust through interface spoofing. Mitigation strategies address these holistically: lightweight explanation caching reduces latency by precomputing high-probability justification templates; adaptive explanation depth dynamically modulates

linguistic complexity and technical granularity based on inferred operator proficiency and situational urgency; and formal verification ensures that escalation triggers preserve critical system invariants---including bounded response time, monotonic confidence decay, and logical consistency between conflict detection and justification validity. These measures collectively strengthen operational robustness without compromising explanatory fidelity or human oversight integrity [1, 10].

5. Conclusion and Forward Pathways

5.1. Synthesis of Contributions

This section synthesizes the core contributions of the Human-in-the-Loop Conflict Resolution (HITL-CR) framework as a cohesive, principled advancement in explainable cyber-physical systems. Rather than retrofitting explanation onto existing autonomy stacks or relying on threshold-based alerting, HITL-CR integrates conflict semantics, explainability-aware escalation logic, and human-centered protocol design into a unified architectural paradigm. It formalizes conflict across sensor, actuator, goal, and temporal dimensions using control-theoretic invariants and couples each detection with natural-language justifications grounded in real-time system confidence. Escalation is governed by dual-threshold logic: one assessing conflict severity, the other evaluating explanation fidelity against operationally validated bounds. This ensures human engagement is both timely and epistemically justified---neither premature nor deferred beyond cognitive recoverability. The framework thus constitutes a foundational shift from opaque autonomy toward accountable, auditable, and cognitively sustainable human-machine orchestration.

5.2. Limitations and Research Frontiers

The HITL-CR framework, while advancing principled human-centered conflict resolution, exhibits several critical limitations. Its current conflict semantics rely predominantly on deterministic models that inadequately capture stochastic inter-agent dependencies and emergent behavioral modes observed in heterogeneous operator cohorts. Empirical validation remains constrained to controlled laboratory settings and narrow demographic samples, limiting generalizability across age, expertise level, and cultural background. Three frontiers now define the research agenda: first, learning-based optimization of escalation policies under real-time resource constraints, balancing explanation fidelity against cognitive load; second, neurocognitive modeling of explanation comprehension---linking linguistic features of justifications to neural correlates of trust calibration and mental model updating; third, co-development of certification-ready standardization pathways, including formal verification templates for explanation logic and interoperable protocol interfaces acceptable to regulatory bodies overseeing safety-critical domains.

5.3. Practical Implementation Roadmap

A pragmatic implementation roadmap advances HITL-CR adoption through three sequential, interoperable phases. First, legacy cyber-physical systems undergo retrofitting with a lightweight HITL-CR monitoring layer that intercepts sensor-actuator inconsistencies and computes conflict severity indices without altering core control logic. Second, explanation modules---designed for real-time embedded execution---are incrementally integrated into certified controllers, generating natural-language justifications aligned with domain ontologies and confidence-calibrated fidelity thresholds. Third, escalation interfaces are co-developed with frontline operators via participatory design sprints, ensuring cognitive alignment, minimizing workload spikes, and embedding contextual awareness directly into handover protocols. Each phase incorporates iterative validation against operational safety constraints and human factors benchmarks.

References

1. H. Kagermann, W. Wahlster, and J. Helbig, "Recommendations for implementing the strategic initiative INDUSTRIE 4.0: Final report of the Industrie 4.0 Working Group," Acatech, 2013.
2. D. Romero, P. Bernus, O. Noran, J. Stahre, and A. Fast-Berglund, "The operator 4.0: Human cyber-physical systems and adaptive automation towards human-automation symbiosis work systems," *IFIP Advances in Information and Communication Technology*, vol. 488, pp. 677–686, 2016. https://doi.org/10.1007/978-3-319-51133-7_80
3. F. Longo, L. Nicoletti, and A. Padovano, "Smart operators in Industry 4.0: A human-centered approach to enhance operators' capabilities and competencies within the new smart factory context," *Computers & Industrial Engineering*, vol. 113, pp. 144–159, 2017. https://doi.org/10.1016/j.cie.2017.09.016
4. R. Parasuraman, T. B. Sheridan, and C. D. Wickens, "A model for types and levels of human interaction with automation," *IEEE Transactions on Systems, Man, and Cybernetics - Part A*, vol. 30, no. 3, pp. 286–297, 2000. https://doi.org/10.1109/3468.844354
5. M. R. Endsley, "From here to autonomy: Lessons learned from human-automation research," *Human Factors*, vol. 59, no. 1, pp. 5–27, 2017. https://doi.org/10.1177/0018720816681350
6. J. D. Lee and K. A. See, "Trust in automation: Designing for appropriate reliance," *Human Factors*, vol. 46, no. 1, pp. 50–80, 2004. https://doi.org/10.1518/hfes.46.1.50_30392
7. L. Bainbridge, "Ironies of automation," *Automatica*, vol. 19, no. 6, pp. 775–779, 1983. https://doi.org/10.1016/0005-1098(83)90046-8
8. M.-P. Pacaux-Lemoine, D. Trentesaux, G. Zambrano Rey, and P. Millot, "Designing intelligent manufacturing systems through human-machine cooperation principles: A human-centered approach," *Computers & Industrial Engineering*, vol. 111, pp. 581–595, 2017. https://doi.org/10.1016/j.cie.2017.05.014
9. S. Amershi, D. Weld, M. Vorvoreanu, et al., "Guidelines for human-AI interaction," *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pp. 1–13, 2019. https://doi.org/10.1145/3290605.3300233
10. K. A. Hoff and M. Bashir, "Trust in automation: Integrating empirical evidence on factors that influence trust," *Human Factors*, vol. 57, no. 3, pp. 407–434, 2015. https://doi.org/10.1177/0018720814547570
11. C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," *Nature Machine Intelligence*, vol. 1, no. 5, pp. 206–215, 2019. https://doi.org/10.1038/s42256-019-0048-x
12. A. B. Arrieta, N. Diaz-Rodriguez, J. Del Ser, et al., "Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," *Information Fusion*, vol. 58, pp. 82–115, 2020. https://doi.org/10.1016/j.inffus.2019.12.012
13. I. D. Raji, A. Smart, R. N. White, et al., "Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing," *Proceedings of FAT* 2020, pp. 33–44, 2020. https://doi.org/10.1145/3351095.3372873
14. B. Shneiderman, "Human-centered artificial intelligence: Reliable, safe and trustworthy," *International Journal of Human-Computer Interaction*, vol. 36, no. 6, pp. 495–504, 2020. https://doi.org/10.1080/10447318.2020.1741118
15. National Institute of Standards and Technology, "Artificial Intelligence Risk Management Framework (AI RMF 1.0)," NIST AI 100-1, 2023. https://doi.org/10.6028/NIST.AI.100-1
16. ISO/IEC, "ISO/IEC 42001:2023 Artificial intelligence - Management system," International Organization for Standardization, 2023.
17. D. Acemoglu and P. Restrepo, "Automation and new tasks: How technology displaces and reinstates labor," *Journal of Economic Perspectives*, vol. 33, no. 2, pp. 3–30, 2019. https://doi.org/10.1257/jep.33.2.3
18. Y. Wei and Y. Cheng, "An optimal two-dimensional maintenance policy for self-service systems with multi-task demands and subject to competing sudden and deterioration-induced failures," *Reliability Engineering & System Safety*, vol. 255, p. 110628, 2025. https://doi.org/10.1016/j.ress.2024.110628
19. Y. Wei, A. Li, Y. Cheng, and Y. Li, "An optimal multi-level inspection and maintenance policy for a multi-component system with a protection component," *Computers & Industrial Engineering*, vol. 201, p. 110898, 2025. https://doi.org/10.1016/j.cie.2025.110898
20. Y. Wei, A. Li, Y. Li, and Y. Cheng, "An optimal predictive inspection and maintenance policy for a multi-state system: A belief-based SMDP approach," *Reliability Engineering & System Safety*, vol. 265, p. 111497, 2026. https://doi.org/10.1016/j.ress.2025.111497
21. S. Zhao, Y. Wei, Y. Cheng, and Y. Li, "A state-specific joint size, maintenance, and inventory policy for a k-out-of-n load-sharing system subject to self-announcing failures," *Reliability Engineering & System Safety*, vol. 257, p. 110855, 2025. https://doi.org/10.1016/j.ress.2025.110855
22. Z. Huang, Y. Wei, and Y. Cheng, "A quantitative framework for performance-based reliability prediction for a multi-component system subject to dynamic self-reconfiguration," *Reliability Engineering & System Safety*, vol. 262, p. 111188, 2025. https://doi.org/10.1016/j.ress.2025.111188
23. S. Zhao, Y. Wei, Y. Li, and Y. Cheng, "A multi-agent reinforcement learning (MARL) framework for designing an optimal state-specific hybrid maintenance policy for a series k-out-of-n load-sharing system," *Reliability Engineering & System Safety*, vol. 265, p. 111587, 2026. https://doi.org/10.1016/j.ress.2025.111587
24. Y. Wei, S. Zhao, and Y. Cheng, "An optimal joint predictive maintenance and mission abort policy under a cumulative working time success criterion," *Risk Analysis*, vol. 45, no. 11, pp. 3958–3969, 2025. https://doi.org/10.1111/risa.70119
25. S. Zhao, Y. Wei, Y. Li, and Y. Cheng, "An optimal joint maintenance and mission abort policy for a system executing multi-attempt missions," *Reliability Engineering & System Safety*, vol. 266, p. 111667, 2026. https://doi.org/10.1016/j.ress.2025.111667

26. Y. Wei, S. Zhao, and Y. Cheng, "An optimal bi-level inspection and maintenance policy for a multi-component system: An enhanced successively approximated point-based value-iteration algorithm," *Frontiers of Engineering Management*, 2026. https://doi.org/10.1007/s42524-026-5185-4
27. R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*, 2nd ed., MIT Press, 2018.
28. High-Level Expert Group on Artificial Intelligence, "Ethics guidelines for trustworthy AI," European Commission, 2019.
29. Y. Cheng, Wang, and Y. Wang, "A user-based bike rebalancing strategy for free-floating bike sharing systems: A bidding model," *Transportation Research Part E: Logistics and Transportation Review*, vol. 154, p. 102438, 2021. https://doi.org/10.1016/j.tre.2021.102438