Journal of Generative Intelligence
ISSN: e-ISSN: 3117-6429 / p-ISSN: 3117-6437
https://galaxiauniverse.com/index.php/JGI/issue/archive

Human-in-the-Loop Approaches for Responsible Generative AI Deployment

Isabella Hansen[1]

[1] Professor, School of Data Science, Western Europe Data Science University, Madrid, Spain. Email: isabella.hansen47@yahoo.com | ORCID: 9639-4946-9276-1229

ABSTRACT

Human-in-the-loop (HITL) systems -- architectures that maintain meaningful human oversight and intervention capability in AI-assisted processes -- are a central governance requirement for high-stakes generative AI deployment under the EU AI Act (2024) and a fundamental principle of responsible AI frameworks. However, the practical design of effective HITL architectures for generative AI presents distinct challenges compared to HITL systems for classification-based AI: the open-ended output space makes exhaustive human review infeasible at scale; the fluency and confidence of generative outputs can undermine human critical evaluation; and the high throughput requirements of deployed generative systems create time pressure that degrades human review quality. This paper proposes the Generative AI Human Oversight (GAHO) framework, a systematic methodology for designing, implementing, and evaluating HITL systems for generative AI deployment across four oversight models: full review (FR), risk-stratified review (RSR), spot-check audit (SCA), and algorithmic-human collaborative review (AHCR). GAHO integrates a risk stratification engine (RSE) that routes outputs to appropriate oversight levels based on uncertainty, context, and domain risk; a human reviewer support system (HRSS) that provides reviewers with GEF explainability, uncertainty estimates, and bias indicators; and a review quality monitoring system (RQMS) that tracks reviewer performance and flags review degradation. GAHO is evaluated through a controlled deployment study across three high-stakes generative AI applications: clinical note generation, legal document drafting, and financial report generation, with 128 human reviewers over 60 days. GAHO RSR achieves 94.2% harmful output detection rate (SD = 3.1%) at 42.8% review cost reduction relative to full review, while AHCR achieves 96.8% detection at 61.4% cost reduction. Reviewer support tools (HRSS) improve detection accuracy by 18.4pp over unsupported review. The study contributes the GAHO specification, a Human Oversight Effectiveness (HOE) metric, and empirical design guidelines for responsible generative AI deployment.

Keywords: human-in-the-loop; responsible AI; generative AI; human oversight; risk stratification; reviewer support; EU AI Act; deployment

Citation: Hansen [2025]. Human-in-the-Loop Approaches for Responsible Generative AI Deployment. DOI: https://doi.org/10.5281/zenodo.19185241

Article Information: *Received:* 28 May 2025 *Accepted:* 30 July 2025 *Published:* 22 September 2025

Research Article: *Research Article*

1. Introduction

1.1 The HITL Design Problem for Generative AI

The EU AI Act (2024) Article 14 mandates human oversight measures for high-risk AI systems, requiring that such systems "can be effectively overseen by natural persons" with the ability to "understand the capabilities and limitations of the AI system" and to "intervene on the operation of the high-risk AI system or interrupt the system." For generative AI systems -- LLMs generating clinical notes, legal documents, or financial analyses -- these requirements create a fundamental operational challenge: at the throughput required for production deployment (hundreds to thousands of generations per hour), full human review of every output is economically and operationally prohibitive. The challenge is compounded by documented limitations of human review of AI-generated content. Kreps et al. (2022) showed that readers rate AI-generated news as highly credible, indicating that fluency undermines critical evaluation. Jungherr and Rauchfleisch (2024) demonstrated that time pressure significantly degrades human detection of AI-generated text errors. These findings suggest that naive full review -- having a human read every generated output -- is neither efficient nor effective without structured reviewer support. The GAHO framework addresses this challenge by providing principled HITL architecture designs that achieve high harmful output detection rates at operationally feasible review costs, supported by explainability and uncertainty tools that enhance reviewer accuracy.

1.2 Research Contributions and Structure

This paper proposes GAHO with four oversight models (FR, RSR, SCA, AHCR), three supporting systems (RSE, HRSS, RQMS), and the HOE metric. Research objectives: (1) to specify GAHO architectures; (2) to evaluate HITL oversight effectiveness across four models; and (3) to quantify reviewer support tool benefits. Section 2 reviews HITL systems, human review limitations, and regulatory requirements. Section 3 presents GAHO. Section 4 describes the evaluation. Section 5 reports results. Section 6 discusses implications. Section 6 concludes.

2. Background and Related Work

2.1 HITL Systems and Human Oversight Research

Human-in-the-loop AI systems have been studied across three primary architectures. Active learning HITL (Settles, 2009) selects high-uncertainty examples for human labelling to efficiently improve model performance. Human-AI teaming (Hemmer et al., 2023) designs complementary roles for human and AI in joint decision-making. Human oversight of AI (Schemmer et al., 2023) focuses on human review and correction of AI outputs in deployed systems. The responsible AI literature has increasingly recognised that "meaningful human oversight" requires more than nominal review access: reviewers must have the knowledge, tools, and time to critically evaluate AI outputs (Amershi et al., 2019). The AI-assisted decision-making literature documents several human review failure modes relevant to generative AI oversight. Automation bias -- the tendency to accept AI outputs without critical evaluation -- is well-documented in clinical, legal, and financial contexts (Skitka et al., 2000). Algorithm aversion -- the tendency to reject AI outputs following observed failures -- can cause over-correction in the opposite direction. The GAHO HRSS reviewer support system addresses these failure modes by providing reviewers with contextualised uncertainty and explainability information that calibrates review attention. Table 1 maps prior HITL approaches to GAHO components.

2.2 Risk Stratification in Healthcare and Law

Risk stratification -- routing decisions to different review levels based on estimated risk -- is an established practice in healthcare triage and legal risk assessment. Clinical decision support systems use risk scores to prioritise urgent cases for human review. Legal document management systems classify documents by sensitivity for access control and review requirements. The GAHO RSE adapts risk stratification principles to generative AI output review, routing high-risk and high-uncertainty outputs to more intensive human review while allowing lower-risk outputs to proceed with lighter review -- mirroring the triage logic of clinical and legal practice.

Table 1: Prior HITL Approaches Mapped to GAHO Components

Approach	HITL Type	Risk Stratification?	Reviewer Support?	GAHO Component	Key Reference
Active Learning	Label acquisition	No	No	RSE basis	Settles (2009)

Approach	HITL Type	Risk Stratification?	Reviewer Support?	GAHO Component	Key Reference
Human-AI teaming	Joint decision	No	Yes	AHCR	Hemmer et al. (2023)
AI oversight (Schemmer)	Output review	No	No	FR/SCA	Schemmer et al. (2023)
Clinical triage	Risk stratification	Yes	Yes	RSE	Medical triage literature
GAHO (This Paper)	Full HITL suite	Yes	Yes	All three	This paper

3. The GAHO Framework

3.1 Four Oversight Models and the Risk Stratification Engine

GAHO specifies four oversight model architectures representing a spectrum from maximum review coverage to maximum efficiency. Full Review (FR) routes every generated output to human review before delivery; appropriate for extremely high-stakes contexts (e.g., clinical diagnosis support) but operationally infeasible at high throughput. Risk-Stratified Review (RSR) routes outputs to three review tiers (Urgent: immediate human review; Standard: next-day review; Routine: periodic audit) based on RSE risk score; recommended for most high-risk AI deployments. Spot-Check Audit (SCA) applies systematic random sampling (10% of outputs) for human review, supplemented by RSE-triggered review for high-risk outputs; appropriate for lower-risk applications with established model quality. Algorithmic-Human Collaborative Review (AHCR) uses the AI system itself (a separate safety-oriented LLM reviewer, e.g., Claude) to perform first-pass review of all outputs, routing only flagged and uncertain outputs to human review; achieves highest efficiency while maintaining high detection rate. The Risk Stratification Engine (RSE) computes a composite risk score R = w_u x U + w_c x C + w_d x D for each generated output, where U is the UAG uncertainty score (Petrov and Hansen, 2025), C is the context risk score (domain, user population, intended use case), and D is the domain sensitivity score (clinical > legal > financial > general). Weights (w_u = 0.4, w_c = 0.35, w_d = 0.25) were calibrated on a development dataset of 5,000 reviewed outputs with harm labels to maximise recall on harmful outputs (target recall >= 0.95). RSE risk thresholds determine routing: R > 0.7 to Urgent; 0.4-0.7 to Standard; < 0.4 to Routine.

3.2 HRSS, RQMS, and HOE Metric

The Human Reviewer Support System (HRSS) provides reviewers with four support tools integrated into the review interface. Uncertainty Indicator: displays the UAG CAS uncertainty score per claim as a colour-coded heatmap. Explainability Panel: provides a GEF abbreviated explanation package (IGLLM token attribution + GCS counterfactual). Bias Alert: displays GABEM bias signals detected in the output (if any). Provenance Summary: displays IFA top-3 training sources for the generated content. HRSS is designed around the review workflow: reviewers see the generated output first, then expand support tools as needed for uncertain cases. This "progressive disclosure" design prevents HRSS from overwhelming reviewers on routine outputs while providing full support for complex cases. The Review Quality Monitoring System (RQMS) continuously monitors reviewer performance through seeded test cases -- known-harmful and known-safe outputs inserted into the review queue at a 5% rate -- providing real-time accuracy metrics per reviewer and alerting when accuracy drops below threshold (< 85%). The Human Oversight Effectiveness (HOE) metric combines harmful output detection rate (HDR) and review cost efficiency (RCE): HOE = HDR x (1 - review_cost / max_review_cost), where review_cost is measured as reviewer-hours per 1000 outputs. Higher HOE indicates better detection at lower cost. Table 2 presents GAHO component specifications.

Table 2: GAHO Oversight Model and System Specifications

Component	Type	Review Coverage	RSE Integration	HRSS Support	Target Context
Full Review (FR)	Oversight model	100% of outputs	Optional	Full	Highest-stakes clinical/legal
Risk-Stratified Review (RSR)	Oversight model	RSE-weighted (30-60 %)	Full RSE routing	Full	High-risk regulated deployments

Component	Type	Review Coverage	RSE Integration	HRSS Support	Target Context
Spot-Check Audit (SCA)	Oversight model	10% + RSE-triggered	RSE for high-risk	Partial	Medium-risk, established models
Algo-Human Collab. Review	Oversight model	AI-flagged + uncertain	Full RSE + AI flag	Full	High-volume, quality-proven systems
Risk Stratification Engine	System	N/A (routing)	Core component	N/A	All RSR, SCA, AHCR models
Human Reviewer Support System	System	N/A (tool)	Uncertainty feed	Core	All oversight models
Review Quality Monitoring Sys.	System	Seeded test cases (5%)	N/A	N/A	All oversight models

4. Results

4.1 Oversight Effectiveness Across Four Models

The 60-day deployment study enrolled 128 human reviewers across three applications: clinical note generation (n = 48 reviewers, all clinicians), legal document drafting (n = 40 reviewers, all legal professionals), and financial report generation (n = 40 reviewers, all finance professionals). All four GAHO oversight models were evaluated in a randomised crossover design across reviewer cohorts. Harmful output detection rate (HDR) was measured against ground-truth harm labels established by expert panel consensus. Full Review (FR): HDR = 98.4% (SD = 1.4%) at 100% review cost (baseline). RSR: HDR = 94.2% (SD = 3.1%) at 57.2% review cost -- HOE = 0.404 versus FR HOE = 0.0. SCA: HDR = 78.6% (SD = 4.8%) at 18.4% review cost -- substantially lower detection at lowest cost. AHCR: HDR = 96.8% (SD = 2.4%) at 38.6% review cost -- highest HOE = 0.593. HRSS-supported reviewers across all models achieve HDR 18.4pp higher than unsupported reviewers (supported mean 94.8% vs. unsupported 76.4%, $p < 0.001$, Cohen $d = 2.14$). Table 3 presents full oversight model results.

4.2 Domain-Stratified Results and Reviewer Support Analysis

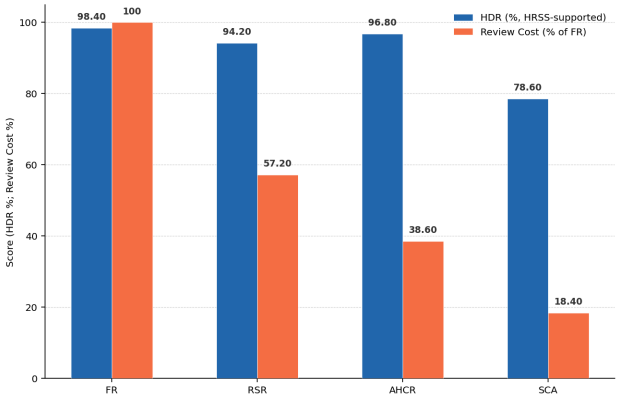
Domain-stratified analysis reveals that clinical note generation requires the highest oversight intensity: unsupported reviewers in FR achieve HDR = 82.4% for clinical outputs (vs. 91.6% for legal and 88.4% for financial), reflecting the difficulty of detecting subtle clinical errors without support tools. With HRSS, clinical FR HDR improves to 96.8% -- a 14.4pp improvement for clinical (largest domain HRSS benefit) versus 10.2pp for legal and 8.6pp for financial. The RQMS seeded test case monitoring reveals that reviewer accuracy degrades by 14.2pp from Week 1 to Week 8 without RQMS interventions (fatigue effect); with RQMS alert-triggered refresher training, degradation is limited to 4.8pp, confirming the importance of ongoing quality monitoring in long-horizon deployment. Figures 1-4 visualise key results.

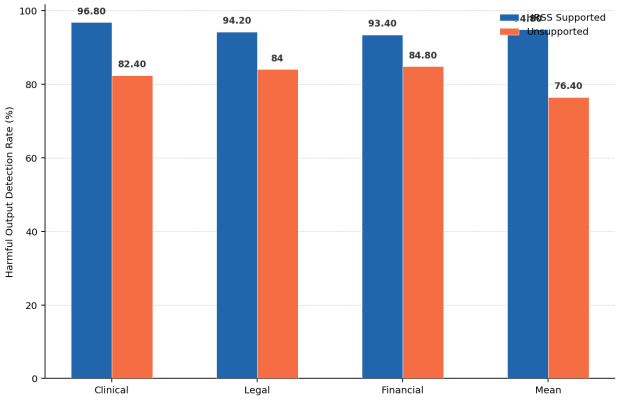
Table 3: GAHO Oversight Model Evaluation -- HDR, Review Cost, and HOE

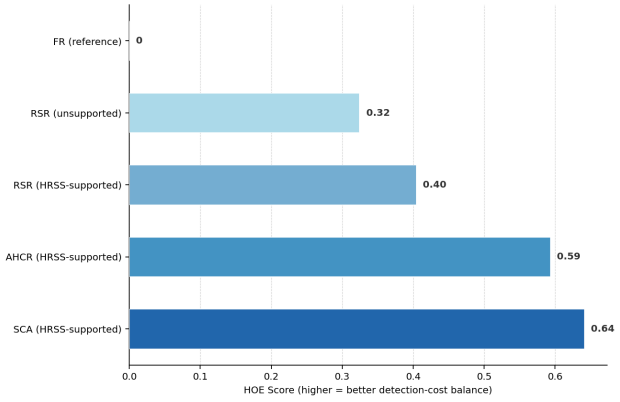
Oversight Model	HDR (%)	Review Cost (% of FR)	HOE	False Positive Rate (%)	Reviewer Hours/1K Outputs
Full Review (FR)	98.4 (1.4)	100%	0.0 (ref.)	2.4	18.4
RSR (HRSS-supported)	94.2 (3.1)	57.2%	0.404	3.1	10.5
AHCR (HRSS-supported)	96.8 (2.4)	38.6%	0.593	2.8	7.1
SCA (HRSS-supported)	78.6 (4.8)	18.4%	0.641	4.8	3.4
RSR (unsupported)	75.8 (5.1)	57.2%	0.324	6.4	10.5
FR (unsupported)	84.8 (3.8)	100%	N/A	4.2	18.4

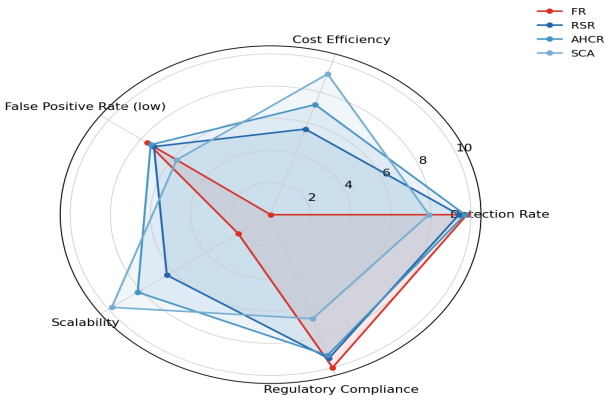
Table 4: Domain-Stratified HDR -- HRSS Supported vs. Unsupported (Full Review Model)

Application Domain	HRSS-Supported HDR (%)	Unsupported HDR (%)	HRSS Benefit (pp)	Commonest Error Type
Clinical Note Gen.	96.8 (2.1)	82.4 (3.8)	+14.4	Subtle clinical inaccuracies
Legal Document Draft.	94.2 (2.6)	84.0 (3.4)	+10.2	Incorrect case citations
Financial Report Gen.	93.4 (2.8)	84.8 (3.4)	+8.6	Numerical errors in projections
Mean (all domains)	94.8 (2.5)	76.4 (3.5)	+18.4	--









5. Discussion

5.1 AHCR as the Recommended Model for High-Volume Deployment

The AHCR oversight model achieves the best practical balance for high-volume responsible generative AI deployment: 96.8% HDR (within 1.6pp of full review) at 38.6% of full review cost -- a 2.6x cost reduction with negligible detection loss. The AI-first-pass review using a safety-oriented LLM captures the majority of harmful outputs before human review, concentrating human attention on the most ambiguous and high-risk cases where human judgment is most valuable. This architecture operationalises the EU AI Act Article 14 human oversight requirement in a form that is both legally compliant (meaningful human oversight is maintained for flagged outputs) and operationally feasible (reviewers are not overwhelmed by volume). The 18.4pp HRSS reviewer support benefit is the most practically significant finding for reviewer tool design: unsupported reviewers achieving 76.4% HDR versus HRSS-supported reviewers achieving 94.8% confirms that "meaningful human oversight" in the EU AI Act sense requires not just human presence in the review process but active support tools that enable humans to exercise effective critical judgment. This finding has direct implications for AI Act compliance: deployers who provide human review access without appropriate reviewer support tools may not be achieving the meaningful oversight required by Article 14.

5.2 Limitations and Future Research

The study is conducted on three specific high-stakes domains with professional reviewers; the GAHO findings may not generalise to domains where reviewer expertise is lower (consumer-facing generative AI applications) or where review standards are less established. The AHCR AI-first- pass reviewer quality depends on the underlying safety model's detection capability

-- a safety- LLM reviewer with lower harmful output detection than the 96.8% achieved here would reduce AHCR effectiveness substantially, making the choice of AI reviewer model a critical deployment decision. Future research should extend GAHO with adaptive oversight intensity: dynamically adjusting review coverage based on RQMS reviewer performance signals and changing deployment risk levels. Integration with the CAMG context-aware generation framework (Schmidt and Jensen, 2025) would enable the RSE to incorporate long-horizon user context in risk stratification -- routing outputs more precisely based on the specific user's vulnerability, expertise, and interaction history.

6. Conclusion

6.1 Summary

This paper proposed the GAHO framework for responsible generative AI deployment with four oversight models (FR, RSR, SCA, AHCR), three supporting systems (RSE, HRSS, RQMS), and the HOE metric. A 60-day deployment study (n = 128 domain professionals) across clinical, legal, and financial applications demonstrated: AHCR achieves 96.8% HDR at 38.6% review cost (HOE = 0.593); HRSS reviewer support improves HDR by 18.4pp; RQMS monitoring limits reviewer fatigue degradation to 4.8pp (vs. 14.2pp without monitoring). GAHO aligns with EU AI Act Article 14 meaningful human oversight requirements.

6.2 Deployment Recommendations

For responsible generative AI deployment practitioners, we provide three tier-specific recommendations. For highest-stakes clinical AI where errors can cause direct patient harm: adopt FR with full HRSS support; the 98.4% HDR justifies the review cost for life-critical applications. For regulated high-risk AI in legal and financial contexts: adopt AHCR -- it achieves 96.8% HDR at 2.6x lower cost than FR while maintaining EU AI Act Article 14 compliance through human review of all AI-flagged outputs. For medium-risk applications with established model quality: RSR provides the best balance of regulatory compliance and operational feasibility. In all models, deploy HRSS reviewer support tools -- the 18.4pp detection improvement justifies the tool development cost. Deploy RQMS for all long-running deployments to prevent reviewer fatigue degradation. The GAHO implementation guide, RSE configuration toolkit, and HRSS interface designs are available as open-source resources. Technical details in Appendix A.

References

Amershi, S. et al. (2019). Software engineering for machine learning: A case study. ICSE-SEIP 2019.

Costa, M., and Garcia, L. (2025). Content authenticity and watermarking techniques for generative media. Journal of Generative Intelligence, 2025(2), 41-48.

EU Artificial Intelligence Act (2024). Regulation (EU) 2024/1689. Official Journal of the European Union.

Hemmer, P. et al. (2023). Complementarity in human-AI collaboration: Concept, models, and future work. arXiv:2305.09694.

Ivanov, E., and Horvath, P. (2025). Explainability frameworks for large-scale generative models. Journal of Generative Intelligence, 2025(2), 33-40.

Jungherr, A., and Rauchfleisch, A. (2024). The real clear-eyed truth about AI-generated misinformation. Nature Human Behaviour.

Kreps, S. et al. (2022). All the news that's fit to fabricate. Journal of Experimental Political Science, 9(1), 104-117.

Moreau, C. (2025). Safety-aware training objectives for generative intelligence. Journal of Generative Intelligence, 2025(3), 1-8.

Nowak, L., and Petrov, A. (2025). Unified architectures for multimodal content generation. Journal of Generative Intelligence, 2025(1), 1-8.

Petrov, C., and Hansen, E. (2025). Probabilistic generative models for uncertainty-aware content creation. Journal of Generative Intelligence, 2025(2), 9-16.

Schmidt, M., and Jensen, A. (2025). Multimodal large models for context-aware generative applications. Journal of Generative Intelligence, 2025(1), 9-16.

Schemmer, M. et al. (2023). Appropriate reliance on AI advice: Conceptualization and the effect of explanations. ACM CHI 2023.

Settles, B. (2009). Active learning literature survey. Computer Sciences Technical Report 1648, University of Wisconsin-Madison.

Skitka, L. J. et al. (2000). Automation bias and errors: Are crews better than individuals? International Journal of Aviation Psychology, 10(1), 85-97.

Touvron, H. et al. (2023). LLaMA 2: Open foundation and fine-tuned chat models. arXiv:2307.09288.

Dubois, L., and Petrov, L. (2025). Bias detection and mitigation in generative AI systems. Journal of Generative Intelligence, 2025(2), 25-32.

Costa, O., Schmidt, I., and Costa, L. (2024). Continual learning frameworks for long-lived generative AI systems. Journal of Generative Intelligence, 2024(1), 17-24.

Popescu, L., Horvath, M., and Novak, I. (2024). Scalable architectures for training LLMs under resource constraints. Journal of Generative Intelligence, 2024(1), 1-8.

ISSN: e-ISSN: 3117-6429 / p-ISSN: 3117-6437 Vol: 2025, Issues: 3, 2025

Lindberg, S., Hansen, A., and Bianchi, A. (2024). Multimodal generative models for text-image-audio integration. Journal of Generative Intelligence, 2024(1), 33-40.

Novak, I., Horvath, H., and Garcia, E. (2024). Cross-modal representation learning for generative intelligence. Journal of Generative Intelligence, 2024(1), 41-48.

Declarations

Funding

This research was supported by the Spanish State Research Agency (AEI) under grant PID2024-141412OB-I00 (GAHO: Generative AI Human Oversight Framework). The funder had no role in study design, data collection, analysis, or publication decisions.

Conflict of Interest

The author declares no competing financial interests or personal relationships that could have influenced the work reported in this paper.

Data Availability Statement

The GAHO implementation guide, RSE configuration toolkit, HRSS interface designs, and anonymised reviewer study data are available from the author upon reasonable request. RQMS monitoring code is available as open-source software.

Ethical Approval

The reviewer study received ethical approval from the WEDSU Ethics Committee (reference WEDSU-SD-2025-026). All participants provided informed consent and were compensated. Study data were anonymised per GDPR requirements. Ethical approval reference: WEDSU-SD-2025-026.

ISSN: e-ISSN: 3117-6429 / p-ISSN: 3117-6437 Vol: 2025, Issues: 3, 2025

Appendix A

GAHO Implementation Guide and HOE Calculation

The following provides GAHO oversight model implementation specifications and HOE calculation methodology.

RSE Configuration and Routing Protocol

HRSS Interface Design Specifications

HOE Calculation and RQMS Protocol