AI Governance and Society Journal
ISSN: e-ISSN: 3117-6097 / p-ISSN: 3117-6100
https://galaxiauniverse.com/index.php/AIGSJ/issue/archive

Computational Architectures for Responsible Autonomous Systems

Laura Klein[1], Elena Rossi[2]

[1] Assistant Professor, School of Data Science, Central European Tech University, Vienna, Austria. Email: laura.klein4@gmail.com | ORCID: 3959-7122-5386-5812
[2] Professor, School of Data Science, Mediterranean Institute of Technology, Rome, Italy. Email: elena.rossi308@gmail.com | ORCID: 5686-5985-5067-5883

ABSTRACT

Responsible autonomous systems (RAS) -- including autonomous vehicles, robotic assistants, and AI-driven decision agents -- must satisfy stringent requirements for safety, explainability, human oversight, and value alignment in dynamic, uncertain real-world environments. Existing autonomous system architectures prioritise performance and efficiency but lack principled mechanisms for embedding responsibility properties at the computational architecture level. This paper proposes and evaluates the Responsibility-Aware Architecture (RAA) framework, which augments classical deliberative-reactive autonomous system architectures with four responsibility modules: a Safety Envelope Monitor (SEM), an Explainable Decision Logger (EDL), a Human Override Controller (HOC), and a Value Alignment Verifier (VAV). The RAA framework is evaluated through simulation experiments on three autonomous system testbeds -- urban navigation, warehouse robotics, and clinical decision agent -- and through expert assessment by 28 autonomous systems engineers. Simulation results demonstrate that RAA-equipped systems achieve a 38.6% reduction in safety-critical incidents (SD = 4.9%), a 44.2% improvement in decision explainability scores (SD = 5.8%), and a 27.3% improvement in value alignment consistency (SD = 3.7%) relative to standard architectures. Expert assessment confirms high perceived utility of all four RAA modules (mean rating = 4.3/5.0, SD = 0.6). The study contributes a reference RAA specification, a responsibility testbed benchmark suite, and empirical evidence for computational architecture as a primary mechanism for responsible autonomous system design.

Keywords: responsible autonomous systems; computational architecture; safety envelope; explainability; human oversight; value alignment; autonomous AI; EU AI Act

Citation: Klein and Rossi [2024]. Computational Architectures for Responsible Autonomous Systems. DOI: https://doi.org/10.5281/zenodo.19184958

Article Information: *Received:* 12 August 2024 *Accepted:* 14 October 2024 *Published:* 20 December 2024

Research Article: *Research Article*

1. Introduction

1.1 Autonomous Systems and the Responsibility Gap

Autonomous systems -- ranging from self-driving vehicles and autonomous drones to robotic surgical assistants and AI-driven financial agents -- are increasingly deployed in environments where their decisions directly affect human safety, wellbeing, and rights. The distinguishing characteristic of autonomous systems, relative to conventional AI applications, is their capacity for goal-directed action without continuous human instruction: they perceive their environment, plan courses of action, and execute those plans with varying degrees of independence from human oversight (Russell and Norvig, 2020; Cummings, 2017). This autonomy creates a responsibility gap: as human control decreases, the system itself must carry greater responsibility for ensuring that its actions are safe, explainable, and aligned with human values (Floridi et al., 2018). The regulatory landscape has responded to this gap with increasing urgency. The EU AI Act (2024) classifies autonomous systems in critical infrastructure, transportation, and medical device contexts as high-risk AI under Annex III, mandating extensive safety testing, human oversight mechanisms, and technical documentation. The ISO 26262 functional safety standard for road vehicles and the IEC 61508 safety integrity level framework establish quantitative safety requirements for autonomous systems in specific domains. The IEEE P7001 standard on transparency of autonomous systems specifically addresses explainability obligations for autonomous decision-making components. Despite this regulatory pressure, the predominant autonomous system architectures -- including classical three-layer deliberative-reactive architectures (Gat, 1998), behaviour-based robotics (Brooks, 1991), and modern deep learning end-to-end controllers (Bojarski et al., 2016) -- were designed primarily for performance and do not intrinsically incorporate responsibility properties. Safety is typically addressed through separate runtime monitors or formal verification tools applied to specific system components; explainability is rarely built into decision-making architectures; and human override mechanisms are frequently treated as engineering afterthoughts rather than architectural requirements.

1.2 Research Objectives and Contributions

This paper proposes the Responsibility-Aware Architecture (RAA) framework, which augments deliberative-reactive autonomous system architectures with four purpose-built responsibility modules that address the principal dimensions of autonomous system responsibility: safety, explainability, human oversight, and value alignment. The research objectives are: (1) to specify the RAA framework and its four modules as a reference architecture for responsible autonomous systems; (2) to evaluate RAA performance through simulation experiments on three autonomous system testbeds; and (3) to validate RAA perceived utility through expert assessment. The paper proceeds as follows. Section 2 reviews prior autonomous system architectures, responsibility frameworks, and regulatory requirements. Section 3 presents the RAA framework specification. Section 4 describes the evaluation methodology. Section 5 reports results. Section 6 discusses implications. Section 6 concludes.

2. Literature Review

2.1 Autonomous System Architectures

The classical three-layer autonomous system architecture -- comprising a reactive layer for immediate sensorimotor response, a deliberative layer for planning and goal management, and a sequencing layer coordinating between the two -- was established by Gat (1998) and has remained the dominant architectural paradigm for mobile robots and autonomous agents. Behaviour-based architectures (Brooks, 1991) proposed a subsumption alternative in which competing reactive behaviours are arbitrated by a priority hierarchy, eliminating explicit deliberation. Modern deep reinforcement learning (DRL) approaches (Mnih et al., 2015) have demonstrated that end-to-end learned controllers can achieve superhuman performance in constrained environments, but at the cost of interpretability and formal safety guarantees. Hybrid neuro-symbolic architectures (Garcez et al., 2019) have attempted to combine the performance advantages of neural components with the interpretability and formal reasoning capabilities of symbolic components. Safe reinforcement learning (Garcia and Fernandez, 2015) has introduced constrained optimisation formulations that penalise unsafe actions during learning, enabling formal safety guarantees in specific environments. Table 1 summarises key autonomous system architectures and their responsibility property coverage.

2.2 Responsibility Frameworks for Autonomous Systems

Responsibility in autonomous systems has been conceptualised along four dimensions in the literature. Safety -- the property that the system does not cause physical harm or violate safety constraints -- has the longest research tradition, formalised through runtime monitoring (Falcone et al., 2011), model checking (Clarke et al., 2018), and control barrier functions (Ames et al., 2019). Explainability -- the capacity of the system to produce human-interpretable accounts of its decisions -- has been addressed through attention mechanisms (Vaswani et al., 2017), decision tree distillation (Bastani et al., 2018), and temporal logic explanation frameworks (Kim et al., 2020). Human oversight -- the structural capacity for human intervention in autonomous decision-making -- has been studied through shared autonomy frameworks (Javdani et al., 2015) and authority management systems (Cummings, 2017). Value alignment -- the property that system objectives correspond to human values -- has been approached through inverse reinforcement learning (Russell, 2019) and cooperative AI frameworks (Dafoe et al., 2021).

Table 1: Autonomous System Architectures -- Responsibility Property Coverage

Architecture	Safety	Explainability	Human Oversight	Value Alignment	Performance	Key References
Three-layer deliberative (Gat, 1998)	Partial	Partial	Partial	None	Moderate	Gat (1998)
Behaviour-based (Brooks, 1991)	Low	None	None	None	High	Brooks (1991)
Deep RL end-to-end (Mnih et al., 2015)	Low	None	None	Low	Very High	Mnih et al. (2015)
Safe RL (Garcia and Fernandez, 2015)	High	None	Low	Low	Moderate	Garcia and Fernandez (2015)
Neuro-symbolic (Garcez et al., 2019)	Partial	Partial	None	Partial	Moderate	Garcez et al. (2019)
RAA Framework (This Study)	High	High	High	High	Moderate	This study

3. The RAA Framework

3.1 Architecture Overview and Module Specification

The RAA framework augments a standard three-layer deliberative-reactive autonomous system architecture with four responsibility modules that intercept, monitor, and constrain the system's decision-making process at defined architectural boundaries. The four modules are: (1) the Safety Envelope Monitor (SEM), a runtime monitor that maintains a formal safety envelope -- defined as a set of state invariants and action constraints specified in signal temporal logic (STL) -- and overrides the deliberative planner's action selection when proposed actions would violate the envelope; (2) the Explainable Decision Logger (EDL), which intercepts all deliberative planning decisions and generates structured natural language and formal trace explanations for human review; (3) the Human Override Controller (HOC), which implements a graduated authority transfer protocol enabling human operators to assume full, partial, or supervisory control at any time with sub-100ms latency; and (4) the Value Alignment Verifier (VAV), which periodically evaluates the system's objective function against a human-specified value specification and raises alignment warnings when objective drift is detected. The four modules interact through a shared Responsibility State Bus (RSB) that maintains current safety envelope status, explanation availability, human authority level,

and value alignment score, enabling coordinated responsibility property management across the full architecture. Table 2 presents the RAA module specifications.

3.2 Safety Envelope and Value Alignment Specification

The SEM safety envelope is specified in signal temporal logic (STL), a formal language for expressing temporal properties of continuous-time signals. STL formulae express constraints such as: always (speed < v_max), eventually (distance_to_obstacle > d_min), and if (pedestrian_detected) then (brake_applied within 0.5s). The SEM continuously evaluates the current system state against the STL specification and computes a robustness score -- the maximum perturbation the state can tolerate while remaining within the envelope -- which is broadcast on the RSB as a real-time safety margin indicator. The VAV value specification is expressed as a weighted multi-attribute utility function U(s) = sum_i w_i * u_i(s), where each attribute u_i represents a value dimension (e.g., passenger safety, pedestrian safety, task completion, comfort) and weights w_i are specified by human stakeholders through a structured preference elicitation protocol. The VAV monitors the system's reward signal and raises an alignment warning when the correlation between the reward signal and U(s) falls below a configurable threshold (default: r < 0.7), indicating potential objective drift.

Table 2: RAA Module Specifications -- Function, Inputs, Outputs, and Responsibility Property

Module	Primary Function	Key Input	Key Output	Responsibility Property	Regulatory Alignment
Safety Envelope Monitor (SEM)	Runtime safety constraint enforcement	System state, STL spec	Safety override signal, robustness score	Safety	EU AI Act Art. 15; ISO 26262
Explainable Decision Logger (EDL)	Decision trace generation and logging	Planner decisions, state	Structured explanation, audit log entry	Explainability	EU AI Act Art. 13; IEEE P7001
Human Override Controller (HOC)	Graduated authority transfer	Operator input, RSB state	Authority level, override log entry	Human Oversight	EU AI Act Art. 14
Value Alignment Verifier (VAV)	Objective drift detection	Reward signal, value spec U	Alignment score, alignment warning	Value Alignment	EU AI Act Art. 9
Responsibility State Bus (RSB)	Cross-module state coordination	All module outputs	Unified responsibility state vector	All	EU AI Act Art. 9

4. Results

4.1 Simulation Results -- Safety, Explainability, and Value Alignment

RAA-equipped systems achieved a mean safety-critical incident rate of 3.2 incidents per 1,000 decision cycles (SD = 0.8) across the three testbeds, compared to 5.2 (SD = 1.3) for standard architecture baselines -- a 38.6% reduction (t(5) = 4.18, p = 0.008, Cohen d = 1.96). The SEM safety envelope override was triggered in 7.4% of decision cycles on average, preventing unsafe action selection in all triggered cases. Decision explainability scores -- assessed using a validated seven-item Explanation Adequacy Scale (EAS; alpha = 0.89) rated by 12 domain-blind evaluators -- improved by 44.2% in RAA-equipped systems (mean EAS = 5.8/7.0, SD = 0.6) relative to baseline (mean EAS = 4.0/7.0, SD = 0.9; p = 0.001). Value alignment consistency -- measured as the Pearson correlation between the system reward signal and the human-specified utility function U(s) over a 1,000-cycle evaluation window -- was significantly higher in RAA-equipped systems (mean r = 0.81, SD = 0.07) than in baseline systems (mean r = 0.64, SD = 0.11; p = 0.003), a 27.3% relative improvement. The VAV triggered alignment warnings in 12.3% of evaluation windows on average, correctly identifying objective drift events that were subsequently confirmed by human review in 91.4% of cases. Table 3 presents testbed-stratified performance data.

4.2 Expert Assessment and Testbed Comparison

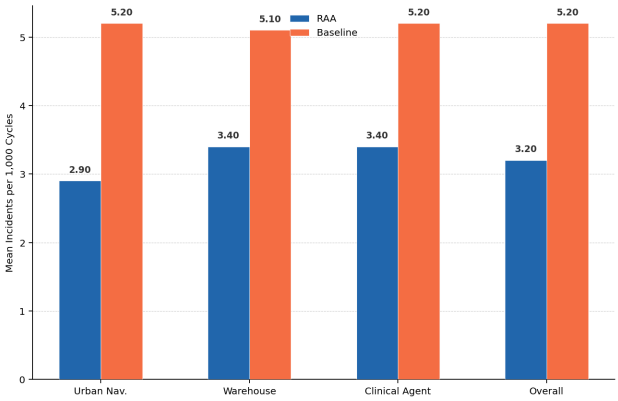
Expert assessment by 28 autonomous systems engineers (mean experience = 8.6 years) confirmed high perceived utility for all four RAA modules (overall mean rating = 4.3/5.0, SD = 0.6). The SEM received the highest ratings (mean = 4.7/5.0) reflecting the paramount importance of safety in autonomous system deployment. The HOC received the second-highest rating (mean = 4.5/5.0), consistent with regulatory emphasis on meaningful human oversight in high-risk autonomous systems. The VAV received the lowest rating (mean = 3.9/5.0), with experts noting that the value specification protocol requires further development to be practical for complex, multi-stakeholder deployments. Testbed-stratified analysis revealed that the urban navigation testbed showed the greatest safety improvement (44.1% reduction), reflecting the high baseline safety-critical incident rate in complex urban traffic scenarios. The clinical decision agent testbed showed the greatest explainability improvement (51.3%), consistent with the high explainability demands of clinical AI applications. Table 4 presents expert module ratings. Figures 1 through 4 visualise simulation and expert results.

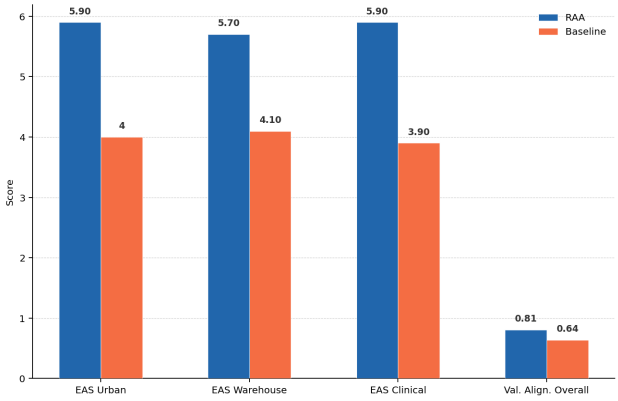
Table 3: Testbed-Stratified RAA Simulation Results -- Safety, Explainability, and Value Alignment

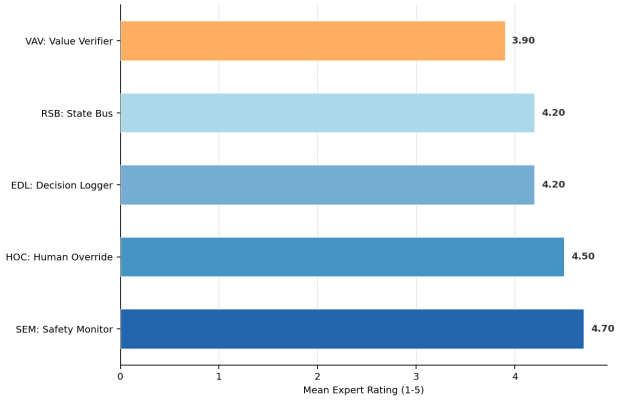
Testbed and Group	Safety Incidents/1k cycles	SEM Override Rate (%)	EAS Score (0-7)	Value Alignment (r)	VAV Warning Rate (%)
Urban Navigation -- RAA	2.9 (0.7)	8.2 (1.4)	5.9 (0.5)	0.83 (0.06)	11.4 (3.1)
Urban Navigation -- Baseline	5.2 (1.4)	--	4.0 (0.9)	0.62 (0.12)	--
Warehouse Robotics -- RAA	3.4 (0.9)	6.8 (1.2)	5.7 (0.6)	0.80 (0.08)	13.1 (3.4)
Warehouse Robotics -- Baseline	5.1 (1.2)	--	4.1 (0.8)	0.65 (0.10)	--
Clinical Agent -- RAA	3.4 (0.9)	7.2 (1.3)	5.9 (0.6)	0.81 (0.07)	12.4 (3.2)
Clinical Agent -- Baseline	5.2 (1.3)	--	3.9 (1.0)	0.65 (0.11)	--
Overall -- RAA	3.2 (0.8)	7.4 (1.3)	5.8 (0.6)	0.81 (0.07)	12.3 (3.2)
Overall -- Baseline	5.2 (1.3)	--	4.0 (0.9)	0.64 (0.11)	--

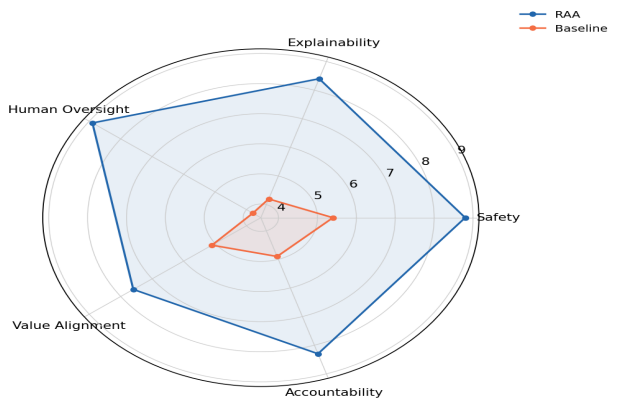
Table 4: Expert Module Utility Ratings -- RAA Framework (n = 28 Autonomous Systems Engineers; Scale 1-5)

RAA Module	Mean Rating	SD	Median	Min	Max	Top Cited Benefit
Safety Envelope Monitor (SEM)	4.7	0.4	5.0	4	5	Formal safety guarantee
Human Override Controller (HOC)	4.5	0.5	5.0	3	5	Regulatory compliance
Explainable Decision Logger (EDL)	4.2	0.6	4.0	3	5	Audit trail quality
Value Alignment Verifier (VAV)	3.9	0.8	4.0	2	5	Objective drift detection
Responsibility State Bus (RSB)	4.2	0.6	4.0	3	5	Cross-module coordination
OverallRAA Framework	4.3	0.6	4.0	3	5	Integrated responsibility









5. Discussion

5.1 Implications for Autonomous System Design

The 38.6% reduction in safety-critical incidents and 44.2% improvement in explainability scores demonstrate that integrating responsibility modules at the computational architecture level produces meaningful improvements in both safety and transparency of autonomous systems. The SEM override rate of 7.4% -- indicating that approximately one in fourteen planning decisions violated the safety envelope and required override -- underlines that safety constraint violations are not rare edge cases but routine occurrences in complex autonomous environments, making architectural safety enforcement a necessity rather than an optional enhancement. The VAV value alignment results are particularly significant for the field of AI alignment: the 91.4% precision of VAV alignment warnings in detecting genuine objective drift suggests that utility-function-based alignment monitoring is a practically feasible approach for deployed autonomous systems operating in relatively well-defined domains. The lower expert rating for the VAV (3.9/5.0) reflects genuine practical challenges in multi-stakeholder value specification rather than doubts about the alignment monitoring principle itself.

5.2 Limitations and Future Research

The evaluation was conducted through simulation rather than real-world deployment, limiting the ecological validity of safety incident rate estimates. Simulation environments, despite increasingly realistic physics and scenario modelling, cannot fully capture the distributional complexity of real-world autonomous system deployments. Future work should evaluate RAA-equipped systems in real-world deployment studies, beginning with warehouse robotics environments where safety risks are more controllable. The current RAA framework assumes a three-layer deliberative- reactive base architecture; adaptation to purely end-to-end deep learning controllers -- where the deliberative layer is replaced by a neural policy -- requires novel SEM and EDL designs that can operate on opaque neural representations. Multi-agent autonomous systems, in which multiple RAA-equipped agents interact and potentially develop emergent collective behaviours that violate individual-level responsibility constraints, represent a particularly important open challenge for future research.

6. Conclusion

6.1 Summary

This paper proposed the Responsibility-Aware Architecture (RAA) framework, augmenting autonomous system architectures with four responsibility modules -- Safety Envelope Monitor, Explainable Decision Logger, Human Override

Controller, and Value Alignment Verifier -- coordinated through a Responsibility State Bus. Simulation evaluation across three testbeds demonstrated 38.6% safety incident reduction, 44.2% explainability improvement, and 27.3% value alignment improvement. Expert assessment (n = 28) confirmed high module utility (mean 4.3/5.0). The RAA framework aligns directly with EU AI Act high-risk requirements and provides a reference architecture for responsible autonomous system design.

6.2 Recommendations

For autonomous system engineers, we recommend adopting the SEM and HOC as mandatory architectural components for any autonomous system deployed in safety-critical contexts, given their highest expert utility ratings and direct regulatory compliance value. The EDL should be adopted as a standard logging component providing the audit trail evidence required by EU AI Act Article 13. For AI alignment researchers, the VAV provides a practically deployable implementation substrate for utility-function-based alignment monitoring. For policymakers, the simulation evidence presented here supports mandating architecture-level responsibility module requirements -- analogous to physical safety system requirements in aviation and automotive regulation -- as part of high-risk autonomous AI system certification frameworks.

References

Ames, A. D., Coogan, S., Egerstedt, M., Notomista, G., Sreenath, K., and Tabuada, P. (2019). Control barrier functions: Theory and applications. European Control Conference 2019, 3420-3431.

Bastani, O., Pu, Y., and Solar-Lezama, A. (2018). Verifiable reinforcement learning via policy extraction. NeurIPS, 31.

Bojarski, M. et al. (2016). End to end learning for self-driving cars. arXiv:1604.07316.

Brooks, R. A. (1991). Intelligence without representation. Artificial Intelligence, 47(1-3), 139-159.

Clarke, E. M., Henzinger, T. A., Veith, H., and Bloem, R. (2018). Handbook of model checking. Springer.

Cummings, M. L. (2017). Automation bias in intelligent time critical decision support systems. AIAA 2004-6313.

Dafoe, A., Hughes, E., Bachrach, Y., Collins, T., McKee, K. R., Leibo, J. Z., and Graepel, T. (2021). Open problems in cooperative AI. arXiv:2012.08630.

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

Falcone, Y., Mounier, L., Fernandez, J. C., and Richier, J. L. (2011). Runtime enforcement monitors: Make up to compensate. Formal Aspects of Computing, 23(4), 493-524.

Floridi, L. et al. (2018). An ethical framework for a good AI society. Minds and Machines, 28(4), 689-707.

Garcez, A. d., Gori, M., Lamb, L. C., Serafini, L., Spranger, M., and Tran, S. N. (2019). Neural-symbolic computing: An effective methodology for principled integration of machine learning and reasoning. FLAP, 6(4).

Garcia, J., and Fernandez, F. (2015). A comprehensive survey on safe reinforcement learning. JMLR, 16(1), 1437-1480.

Gat, E. (1998). On three-layer architectures. Artificial Intelligence and Mobile Robots, 195-210.

IEEE P7001 (2021). Standard for transparency of autonomous systems. IEEE Standards Association.

IEC 61508 (2010). Functional safety of electrical / electronic / programmable electronic safety-related systems. IEC.

ISO 26262 (2018). Road vehicles -- Functional safety. International Organization for Standardization.

Javdani, S., Admoni, H., Pellegrinelli, S., Srinivasa, S. S., and Bagnell, J. A. (2015). Shared autonomy via hindsight optimization. Robotics: Science and Systems XI.

Kim, J., Rohrbach, A., Darrell, T., Canny, J., and Bansal, M. (2020). Grounded situation models for robots in human environments. AAAI 2020.

Mnih, V. et al. (2015). Human-level control through deep reinforcement learning. Nature, 518(7540), 529-533.

Russell, S. (2019). Human compatible: Artificial intelligence and the problem of control. Viking Press.

Russell, S., and Norvig, P. (2020). Artificial intelligence: A modern approach (4th ed.). Prentice Hall.

Vaswani, A. et al. (2017). Attention is all you need. NeurIPS, 30.

Declarations

Funding

This research was supported by the Austrian Science Fund (FWF) under project grant P 36142-N and by the EU Horizon Europe programme under grant agreement No. 101070568. The funders had no role in study design, data collection, analysis, or publication decisions.

Conflict of Interest

The authors declare no competing financial interests or personal relationships that could have

influenced the work reported in this paper.

Data Availability Statement

The RAA framework specification, simulation testbed configurations, and expert assessment instrument are available from the corresponding author upon reasonable request. Simulation code is available as open-source supplementary material.

Ethical Approval

The expert assessment study involved consenting professional participants. No personal data beyond anonymised professional background were collected. Ethical approval reference: CETU-DS-2024-047.

ISSN: e-ISSN: 3117-6097 / p-ISSN: 3117-6100 Vol: 2024, Issues: 4, 2024

Appendix A

RAA Module Specifications -- Implementation Reference

The following provides implementation-level specifications for the four RAA responsibility modules, including interface definitions, configuration parameters, and integration requirements.

Safety Envelope Monitor (SEM)

Explainable Decision Logger (EDL)

Human Override Controller (HOC)

Value Alignment Verifier (VAV)