

A Monotone Redundancy-Aware Decision-Burden Method for Explainable Human Oversight in Human–Robot Collaboration

Seyma Yaman Kayadibi

`seyma.yamankayadibi@live.vu.edu.au`

Victoria University

Research Article

Keywords: trustworthy robotics, explainable autonomy, human–robot collaboration, human oversight, human-in-the-loop decision support, monotonicity, redundancy-aware aggregation, Artificial Age Score

Posted Date: August 4th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10506135/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

A Monotone Redundancy-Aware Decision-Burden Method for Explainable Human Oversight in Human–Robot Collaboration

Seyma Yaman Kayadibi^{1*}

^{1*}Institute for Sustainable Industries and Liveable Cities, Victoria
University, Melbourne, VIC, Australia.

Corresponding author(s). E-mail(s):
seyma.yamankayadibi@live.vu.edu.au;

Abstract

Trustworthy human–robot collaboration requires decision-support methods that expose why human oversight is recommended without allowing an aggregate score to override direct safety conditions. This study reformulates the redundancy-adjusted Artificial Age Score (AAS) as a monotone, event-level decision-burden method. Six normalized inputs describe task hazard severity, human safety risk, environmental uncertainty, task criticality, exposure frequency, and intervention complexity. Logarithmic burden evidence is aggregated using a symmetric pairwise-overlap correction. Unlike a ratio-based redundancy adjustment, the proposed correction is continuous when components become inactive. A sufficient condition on the overlap graph and discount strength guarantees that increasing any burden input cannot decrease the score. An additive attribution allocates each redundancy discount to its participating components, while an independent hard gate preserves mandatory review for direct safety indicators.

The method was evaluated on 65,536 scrambled Sobol profiles in a controlled computational study. Against an independent rule-based oversight policy, the corrected score achieved a held-out ROC AUC of 0.892 (95% bootstrap CI 0.887–0.897) and balanced accuracy of 0.800 (0.793–0.807), compared with 0.880 and 0.788 for an unadjusted logarithmic score. Across 1,179,648 coordinatewise perturbation checks, the corrected method produced no score decreases; the earlier ratio adjustment produced 3,484. Sensitivity, input-noise, weight-perturbation, and scenario analyses further characterize model behaviour. The results establish mathematical and computational validity under controlled profiles, not clinical, occupational-safety, or field effectiveness. Domain-specific elicitation and prospective human-factors validation remain necessary before deployment.

Keywords: trustworthy robotics, explainable autonomy, human–robot collaboration, human oversight, human-in-the-loop decision support, monotonicity, redundancy-aware aggregation, Artificial Age Score

Article type: Methodology

1 Introduction

Collaborative robotic systems distribute perception, planning, execution, monitoring, and intervention across people and machines. The human role is therefore not eliminated when autonomy increases; it is reorganized around supervision, exception handling, and recovery [1, 2]. In a shared workspace, a supervisor may need to interpret several concurrent signals: a task may be hazardous, a person may be exposed, the environment may be uncertain, production may be time-critical, exposure may be frequent, or intervention may be difficult. A useful decision-support layer should integrate these signals while keeping the path from inputs to oversight advice intelligible.

This requirement sits at the intersection of human–robot interaction (HRI), explainable artificial intelligence (XAI), and risk governance. HRI research emphasizes that safety, task allocation, trust, communication, ergonomics, and shared autonomy are coupled rather than isolated concerns [3–6]. Trust is better calibrated when system limits and reasons for intervention are visible [7, 8]. XAI research likewise argues that an explanation should be meaningful to its recipient and relevant to the decision being made [9, 10]. For an oversight recommendation, a total score alone is consequently insufficient: the system should identify contributing dimensions, distinguish direct safety triggers from accumulated operational burden, and behave predictably when an input worsens.

Existing safety standards do not authorize a generic AI score to certify a collaborative operation. ISO 12100 frames machinery safety through hazard identification, risk estimation, and risk reduction [11]; ISO/TS 15066 addresses collaborative robot applications in relation to the robot system, human, workspace, and task [12]. The NIST AI Risk Management Framework similarly emphasizes context, monitoring, documentation, and human accountability [13]. These sources motivate a narrow role for the method developed here: it is an interpretable oversight-support layer, not a replacement for risk assessment, safety-rated control, or stop functions.

The Artificial Age Score (AAS) was introduced as a log-scaled, redundancy-aware measure of structural burden in generative-AI memory systems [14]. A previous extension mapped the same conceptual pipeline to HRC events. Close examination of its redundancy ratio, however, reveals an important mathematical limitation. When no other component is active, the redundancy ratio is defined as zero. Activating an arbitrarily small overlapping component can make that ratio jump to a positive constant. The total score can therefore be discontinuous at an inactive boundary and can decrease when a burden input increases. Such behaviour is undesirable for supervisory decision support because two practically indistinguishable profiles may receive materially different classifications.

This study develops a corrected AAS-derived methodology with four design requirements:

1. *continuity*: a vanishing input must have a vanishing redundancy effect;
2. *monotonicity*: increasing one burden component, with all others fixed, must not reduce total burden;
3. *traceability*: the score must decompose into non-negative component contributions; and
4. *non-override*: aggregate scoring must not suppress mandatory review caused by direct safety indicators.

The contribution is methodological rather than a new robot-control algorithm. First, a pairwise overlap correction replaces the unstable ratio and yields a simple sufficient condition for coordinatewise monotonicity. Second, a component attribution makes the redundancy discount auditable. Third, a two-stage decision architecture separates a hard safety gate from a calibrated supervisory threshold. Fourth, a reproducible computational protocol evaluates the corrected score against linear and unadjusted baselines, audits monotonicity directly, and examines sensitivity to parameters, noise, and weights. The following research questions guide the study:

RQ1: Can a redundancy adjustment be formulated so that event-level burden is continuous, non-negative, and provably monotone in every input?

RQ2: Does redundancy adjustment improve concordance with an independently specified oversight policy relative to simpler aggregation baselines?

RQ3: How sensitive are scores and rankings to model parameters, input noise, and weight perturbations?

The intended unit of analysis is one decision event, not a person, robot, or complete work shift. A profile may be recomputed when observations change, but the present score has no memory and does not infer a latent psychological state. It summarizes declared evidence under a declared configuration. This distinction is important because the term “burden” can otherwise be confused with operator workload, fatigue, or clinical strain. Those constructs require validated measurement instruments and are not estimated here.

The manuscript is organized as follows. The next section relates the method to supervisory control, XAI, AAS, and redundant-evidence aggregation. Section 3 defines the measurement contract and overlap prior. Section 4 derives the corrected score and proves its main properties. Section 5 gives the complete computational validation protocol, followed by results in Section 6. Section 7 discusses interpretation, deployment, and limitations. The appendices provide the numerical specification, additional derivations, detailed scenario decompositions, pseudocode, and a reproducibility checklist.

2 Related work

2.1 Human oversight in collaborative robotics

Human supervisory control concerns the monitoring and intermittent direction of automated processes [1]. Parasuraman et al. distinguish information acquisition, information analysis, decision selection, and action implementation, showing that human

participation may differ across automation functions [2]. Endsley’s account of situation awareness further connects effective intervention to perception of current state, comprehension of its meaning, and projection of future consequences [15]. An oversight-support method should therefore summarize system state without hiding the evidence needed for comprehension.

In HRC, the need for such support is amplified by physical proximity and joint task execution. Surveys identify collision avoidance, motion planning, control, human prediction, communication, and psychological safety as complementary dimensions of safe interaction [6, 16]. Industrial reviews also report that ergonomic design, intuitive interfaces, trust, and work organization influence whether collaboration is effective [5, 17, 18]. A decision-burden score can organize monitoring information, but cannot establish that a robot cell is safe; that claim requires task-specific engineering evidence and validation.

2.2 Explainability and monotonic decision support

Explanations are social and contrastive: people typically seek reasons that distinguish one outcome from plausible alternatives [9]. In intelligent systems, useful explanations therefore depend on both technical transparency and the decision context [10]. For HRI, Hayes and Shah demonstrate the value of generating explanations about a robot’s policy in human-understandable terms [19], while human-AI interaction guidelines recommend making system status, capabilities, and consequences visible [20].

For an additive oversight model, two properties make an explanation easier to defend. First, component attributions should reconcile exactly with the displayed total. Second, worsening an input should not improve the score unless a clearly stated compensating mechanism exists. The present application assumes no such compensation: each input is a burden coordinate, not a benefit. Coordinatewise monotonicity is therefore a semantic requirement rather than merely a numerical preference.

2.3 Aggregation and redundant evidence

A weighted sum is transparent but treats all evidence as independent. When several dimensions partly describe the same underlying concern, additive aggregation can double-count evidence. Multi-criteria decision analysis addresses weights, trade-offs, and sensitivity, but the validity of an aggregation rule remains application dependent [23]. The original AAS addresses overlap through a component-specific redundancy ratio [14]. Its transfer to HRC is attractive because it combines a nonlinear penalty with explicit overlap. Yet a ratio whose denominator depends on which other inputs are active can behave poorly at sparse boundaries. The proposed method retains transparent pairwise overlap while removing this dependence on an activity-normalized denominator.

2.4 From artificial age to event-level decision burden

The phrase “Artificial Age” originates in an evaluation of generative-AI memory persistence rather than robotics [14]. Its underlying construction is broader than

chronological age: an observed deviation is converted into a consistency value, transformed by a regularized logarithmic kernel, and aggregated after accounting for redundant evidence. The resulting quantity is a structural score. In the present application, the observable deviation is not memory loss; it is the normalized burden assigned to a collaboration event.

This transfer requires care. A formula developed for one measurement context does not acquire validity in another simply because its inputs can be normalized. Three layers must be distinguished. The *measurement layer* defines what each u_i means and how evidence becomes a value in $[0, 1]$. The *aggregation layer* maps those values to an event score. The *decision layer* maps the score and external safety conditions to an oversight status. The present study contributes primarily to the second layer and supplies a transparent demonstration of the third. It does not validate the first layer for a particular workplace.

The transfer also changes the interpretation of redundancy. In a memory evaluation, overlap can reflect metamorphic relations or shared failure evidence. In HRC decision support, overlap may reflect common conceptual sources, such as three indicators all describing direct safety concern. The demonstrator represents this relation through a fixed, inspectable graph. Such a graph is a modelling prior: it can prevent naive double counting, but it must not be described as learned correlation, causal dependence, or a universal ontology.

2.5 Trustworthy robotics, human-centred AI, and governance boundaries

Trustworthy robotics requires more than task performance: supervisory methods should make their limits, decision basis, and allocation of authority inspectable. Human-centred AI similarly emphasizes systems that amplify human capability while remaining reliable, safe, and trustworthy [21]. Socially meaningful transparency also depends on organizational and human context, not only on exposing internal computation [22]. For the proposed method, this means that displaying a score and six contributions is necessary but not sufficient. A deployment should also expose the input provenance, current parameter version, missing-data status, gate that fired, and action available to the supervisor.

Governance boundaries are particularly important in robotics because an informational recommendation may be mistaken for a safety function. The present score has no safety integrity level, does not monitor its own sensor chain, and does not command a stop. It is therefore positioned outside safety-rated control. Its output may inform an authorized person or an operational workflow, but the underlying risk assessment and protective measures remain authoritative. This boundary motivates the non-override design requirement and the explicit distinction between mandatory and score-triggered review.

3 Design requirements and event representation

3.1 Normalized event profile

Definition 1 (HRC event profile) An event is represented by

$$\mathbf{u} = (u_1, \dots, u_6) \in [0, 1]^6, \quad (1)$$

where higher values consistently indicate greater burden.

Table 1 defines the six coordinates and gives examples of evidence from which a site-specific normalization could be constructed. The examples are deliberately non-prescriptive. A deployment must document its sensor processing, expert scales, missing-data rules, temporal window, and calibration population.

Table 1 Event dimensions and possible operational evidence. The present study evaluates normalized values directly and does not prescribe a universal measurement instrument.

Input	Dimension	Examples of site-specific evidence
u_1	Task hazard severity	Energy, sharp or hot tooling, payload, speed, or consequence categories from a task risk assessment.
u_2	Human safety risk	Separation, exposure, body region, protective measures, and estimated likelihood of harmful contact.
u_3	Environmental uncertainty	Occlusion, sensor disagreement, localization uncertainty, unexpected objects, or workspace change.
u_4	Task criticality	Consequence of delay or failure, recovery window, production or care dependency.
u_5	Exposure frequency	Repetition rate, duration, or cumulative time during which the condition is present.
u_6	Intervention complexity	Number of recovery steps, access difficulty, time pressure, expertise, or coordination needed.

The first three dimensions are treated as direct safety indicators for the demonstration gate; the final three describe operational monitoring demand. This distinction is architectural, not ontological: operational variables can also influence safety, and each deployment must determine which observations connect to safety-rated functions.

3.2 Measurement contract and design assumptions

Normalization creates comparability only after the endpoints have been defined. For each dimension, the deployment specification should state what evidence corresponds to 0 and 1, whether intermediate values are linear or ordinal, how long observations are aggregated, and what happens when evidence is unavailable. A value of zero means “no burden represented by this input under the declared measurement rule”; it does not mean that the overall operation is safe. Likewise, a value of one represents the upper endpoint of the chosen scale, not a universal physical maximum.

The model assumes directional consistency: every u_i is coded so that a larger value means no less burden. If an operational quantity has the opposite direction, such as separation distance or confidence, it must be transformed before entry. The model also assumes that the six values refer to a sufficiently synchronized event window. Combining a momentary uncertainty estimate with a weekly exposure frequency without a documented temporal rule would make the profile difficult to interpret.

Table 2 translates the conceptual requirements into properties that can be inspected before deployment. The first four are properties of the aggregation architecture; the final three are requirements on the surrounding measurement and governance process.

Table 2 Design and governance requirements for an oversight-support score.

Requirement	Meaning	Verification used here
Continuity	Negligible changes in an input should not create a finite jump in the score.	Analytic construction and boundary counterexample.
Coordinatewise monotonicity	Increasing a burden input while fixing all others should not lower the total.	Sufficient theorem plus 1,179,648 perturbation checks.
Bounded redundancy	Overlap may discount repeated evidence but cannot reverse it or create negative burden.	Bounds on h_κ , S_{AAS} , and c_i .
Exact traceability	Displayed component contributions must sum exactly to the displayed score.	Symmetric allocation in equation (20) .
Measurement provenance	Each normalized value must have a documented source, scale, window, and missing-data rule.	Required for deployment; not empirically evaluated here.
Safety non-override	An aggregate score cannot cancel a direct safety gate.	Two-stage decision architecture in equation (23) .
Version accountability	Weights, overlaps, thresholds, and transformations must be versioned and reviewable.	Complete numerical specification and code supplement.

3.3 Weights and conceptual overlap

Let $\mathbf{w} \in \mathbb{R}_{\geq 0}^6$, $\sum_i w_i = 1$, specify the relative contribution of the six dimensions. The demonstrator uses

$$\mathbf{w} = (0.25, 0.23, 0.16, 0.14, 0.10, 0.12). \quad (2)$$

These values preserve the priorities of the earlier HRC formulation and are treated as elicited design priors, not learned population parameters.

Pairwise conceptual overlap is represented by a symmetric matrix $Q = [q_{ik}]$ with $q_{ik} \in [0, 1]$ and $q_{ii} = 0$. To keep the computational case study auditable, each

dimension is assigned a two-label feature set:

$$\begin{aligned} G_1 &= \{\text{direct safety, task}\}, & G_2 &= \{\text{direct safety, human}\}, \\ G_3 &= \{\text{direct safety, environment}\}, & G_4 &= \{\text{operational burden, task}\}, \\ G_5 &= \{\text{operational burden, repetition}\}, & G_6 &= \{\text{operational burden, intervention}\}. \end{aligned}$$

For $i \neq k$, $q_{ik} = |G_i \cap G_k|/|G_i \cup G_k|$. Thus, pairs sharing one label have $q_{ik} = 1/3$; other pairs have zero overlap. The non-zero edges are $(1, 2)$, $(1, 3)$, $(2, 3)$, $(1, 4)$, $(4, 5)$, $(4, 6)$, and $(5, 6)$. Define the weighted degree $d_i = \sum_{k \neq i} q_{ik}$. Here,

$$\mathbf{d} = (1, 2/3, 2/3, 1, 2/3, 2/3), \quad d_{\max} = 1. \quad (3)$$

The labels make the prior inspectable, but they do not establish empirical correlation or causal dependence.

For completeness, the resulting matrix is

$$Q = \frac{1}{3} \begin{bmatrix} 0 & 1 & 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 1 & 0 \end{bmatrix}. \quad (4)$$

The connection between u_1 and u_4 arises from their shared “task” label; the two three-node groups arise from their shared safety or operational-burden label. Setting a pair to zero means only that the demonstrator applies no conceptual discount to that pair. It is not a claim of statistical independence.

4 Monotone redundancy-aware method

4.1 Consistency and logarithmic burden evidence

For input u_i , the earlier AAS pathway first defines a deviation

$$\Delta_i = \frac{u_i}{1 - u_i + \eta} \quad (5)$$

and consistency $x_i = 1/(1 + \Delta_i)$. Algebraically, this composition simplifies to

$$x_i = \frac{1 - u_i + \eta}{1 + \eta}. \quad (6)$$

The simplified form is numerically direct and makes monotonicity explicit. With stabilizers $\eta > 0$ and $\epsilon > 0$, define the logarithmic penalty and base evidence

$$p_i(u_i) = -\log_2 \left(\frac{x_i + \epsilon}{1 + \epsilon} \right), \quad (7)$$

$$y_i = w_i p_i(u_i). \quad (8)$$

The experiments set $\eta = \epsilon = 10^{-6}$. Each p_i is continuous and strictly increasing on $[0, 1]$:

$$\frac{\partial p_i}{\partial u_i} = \frac{1}{\ln(2)(1+\eta)(x_i + \epsilon)} > 0. \quad (9)$$

Proposition 1 (Range of the transformed evidence) *For $\eta, \epsilon > 0$ and $u_i \in [0, 1]$, p_i is finite and*

$$0 \leq p_i(u_i) \leq -\log_2 \left(\frac{\eta/(1+\eta) + \epsilon}{1 + \epsilon} \right) = p_{\max}. \quad (10)$$

Consequently, $0 \leq y_i \leq w_i p_{\max}$.

Proof Equation (6) maps $[0, 1]$ to $[\eta/(1+\eta), 1]$. The penalty is continuous and increasing in u_i by equation (9); its minimum is $p_i(0) = 0$ and its maximum is attained at $u_i = 1$. Multiplication by $w_i \geq 0$ gives the evidence bound. $\square$

This bound is useful for numerical interpretation. The regularizers make the score finite even at the upper boundary; they do not materially change values away from that boundary. Because the penalty is logarithmic, equal increments of u_i have larger effects as u_i approaches one. The noise analysis later evaluates this intended but practically important curvature.

4.2 Why the activity-normalized ratio is unsuitable

The discontinued formulation assigned component i a redundancy ratio

$$R_i(\mathbf{u}) = \begin{cases} \frac{\sum_{k \neq i} w_k u_k q_{ik}}{\sum_{k \neq i} w_k u_k}, & \sum_{k \neq i} w_k u_k > 0, \\ 0, & \text{otherwise,} \end{cases} \quad (11)$$

and used $\sum_i (1 - R_i) y_i$. The conditional definition bounds R_i , but not its change at the boundary. Consider $u_3 = 0.999$ and every other input zero. Then $R_3 = 0$. If $u_1 = \delta > 0$ is introduced while all other values remain fixed, $R_3 = q_{31} = 1/3$ for every δ , however small. Consequently,

$$\lim_{\delta \downarrow 0} S_{\text{ratio}}(\delta, 0, 0.999, 0, 0, 0) \neq S_{\text{ratio}}(0, 0, 0.999, 0, 0, 0). \quad (12)$$

Numerically, the score falls from 1.5941 at $\delta = 0$ to 1.0627 as $\delta \downarrow 0$. This is both a discontinuity and a monotonicity violation. The issue is structural: activity normalization makes the magnitude of a redundancy ratio independent of the amount of newly activated evidence.

4.3 Pairwise overlap correction

For non-negative base evidence a, b , define a smooth overlap function

$$h_\kappa(a, b) = \frac{ab}{a + b + \kappa}, \quad \kappa > 0. \quad (13)$$

Proposition 2 (Properties of pairwise overlap) *For $a, b \geq 0$ and $\kappa > 0$, h_κ is symmetric, continuous, non-negative, non-decreasing in a and b , and non-increasing in κ . It satisfies*

$$0 \leq h_\kappa(a, b) \leq \min(a, b), \quad h_\kappa(a, 0) = h_\kappa(0, b) = 0. \quad (14)$$

Proof All claims except the bounds follow directly from [equation \(13\)](#) and its partial derivatives. If $a, b > 0$, then

$$\frac{h_\kappa(a, b)}{a} = \frac{b}{a + b + \kappa} < 1, \quad \frac{h_\kappa(a, b)}{b} = \frac{a}{a + b + \kappa} < 1.$$

The cases with a zero argument hold by substitution. Finally, $\partial h_\kappa / \partial \kappa = -ab / (a + b + \kappa)^2 \leq 0$. $\square$

The parameter κ controls how quickly overlap becomes substantial when two evidence values are small. It is not a probability or an extra redundancy coefficient. Pair specificity is supplied by q_{ik} , while γ controls the overall discount strength. The monotone redundancy-aware AAS score is

$$S_{\text{AAS}}(\mathbf{u}) = \sum_{i=1}^6 y_i - \gamma \sum_{1 \leq i < k \leq 6} q_{ik} h_\kappa(y_i, y_k) \quad (15)$$

with discount strength $\gamma \geq 0$. The main analysis uses $\gamma = 0.80$ and $\kappa = 0.05$.

Theorem 1 (Sufficient condition for coordinatewise monotonicity) *If $\gamma d_{\max} \leq 1$, then S_{AAS} is continuous on $[0, 1]^6$ and non-decreasing in every coordinate u_i .*

Proof Continuity follows from [equations \(6\), \(7\) and \(13\)](#) because $\eta, \epsilon, \kappa > 0$. For $a, b \geq 0$,

$$\frac{\partial h_\kappa(a, b)}{\partial a} = \frac{b(b + \kappa)}{(a + b + \kappa)^2} \in [0, 1]. \quad (16)$$

Therefore,

$$\frac{\partial S_{\text{AAS}}}{\partial y_i} = 1 - \gamma \sum_{k \neq i} q_{ik} \frac{\partial h_\kappa(y_i, y_k)}{\partial y_i} \geq 1 - \gamma d_i \geq 0. \quad (17)$$

Because y_i depends only on u_i and $\partial y_i / \partial u_i \geq 0$ by [equation \(9\)](#), the chain rule gives $\partial S_{\text{AAS}} / \partial u_i \geq 0$. $\square$

For the demonstrator, $\gamma d_{\max} = 0.8$, leaving a monotonicity margin of 0.2. Equality at one is permitted by the theorem, but a positive margin gives more tolerance to later overlap changes.

Proposition 3 (Marginal bounds and global continuity) *Under $\gamma d_{\max} \leq 1$,*

$$1 - \gamma d_i \leq \frac{\partial S_{\text{AAS}}}{\partial y_i} \leq 1. \quad (18)$$

Let $x_{\min} = \eta/(1 + \eta)$. Then

$$0 \leq \frac{\partial S_{\text{AAS}}}{\partial u_i} \leq \frac{w_i}{\ln(2)(1 + \eta)(x_{\min} + \epsilon)} = L_i. \quad (19)$$

Thus, S_{AAS} is globally Lipschitz on $[0, 1]^6$ with respect to the weighted ℓ_1 bound $|S_{\text{AAS}}(\mathbf{u}) - S_{\text{AAS}}(\mathbf{v})| \leq \sum_i L_i |u_i - v_i|$. If $w_i > 0$ and $\gamma d_i < 1$, the score is strictly increasing in u_i .

Proof The derivative of h_κ with respect to either evidence argument lies in $[0, 1]$, giving [equation \(18\)](#). Combining it with [equation \(9\)](#) gives the non-negative derivative in [equation \(19\)](#). Since $x_i \geq x_{\min}$, replacing x_i by $x_{\min}$ gives the finite upper bound. The mean value theorem applied along coordinate segments yields the Lipschitz inequality. Strictness follows when both factors in the chain rule are strictly positive. $\square$

The main specification has $w_i > 0$ and $\gamma d_i \leq 0.8$ for every dimension. It is therefore strictly, not merely weakly, increasing in all six inputs. The global Lipschitz constants are conservative because they use the maximum logarithmic slope at $u_i = 1$; observed perturbation behaviour is reported separately.

4.4 Exact component attribution

For explanation, allocate half of every pairwise discount to each participating component:

$$c_i = y_i - \frac{\gamma}{2} \sum_{k \neq i} q_{ik} h_\kappa(y_i, y_k). \quad (20)$$

Corollary 1 (Reconciliation and non-negativity) *If $\gamma d_{\max} \leq 1$, then*

$$\sum_i c_i = S_{\text{AAS}}, \quad c_i \geq \frac{1}{2} y_i \geq 0, \quad \frac{1}{2} \sum_i y_i \leq S_{\text{AAS}} \leq \sum_i y_i. \quad (21)$$

Proof Every pairwise term appears as two half-discounts, so contributions reconcile with [equation \(15\)](#). Since $h_\kappa(y_i, y_k) \leq y_i$,

$$c_i \geq y_i \left(1 - \frac{\gamma d_i}{2}\right) \geq \frac{1}{2} y_i.$$

The score bounds follow by summing over components and noting that every discount is non-negative. $\square$

The output can therefore report the total score, all six reconciled contributions, the largest contributor, and the relevant gate state without negative attributions.

4.5 Two-stage oversight architecture

Aggregation is deliberately separated from direct safety gating. Let

$$M(\mathbf{u}) = \mathbb{I}\{u_1 \geq 0.75 \vee u_2 \geq 0.70 \vee u_3 \geq 0.80\} \quad (22)$$

denote mandatory human review in the computational demonstrator. If $M = 1$, review is mandatory regardless of the aggregate score. If $M = 0$, a calibrated threshold τ maps the score to a supervisory recommendation:

$$D(\mathbf{u}) = \begin{cases} \text{mandatory human review,} & M(\mathbf{u}) = 1, \\ \text{supervisory review,} & M(\mathbf{u}) = 0 \text{ and } S_{\text{AAS}} \geq \tau, \\ \text{ordinary operation,} & \text{otherwise.} \end{cases} \quad (23)$$

These thresholds are experimental values, not safety limits. In an operational system, a safety-rated protective function must remain external to this informational layer.

4.6 Decision record and computational complexity

A complete output should contain more than the decision label. The proposed decision record is

$$\mathcal{O} = (\mathbf{u}, S_{\text{AAS}}, \mathbf{c}, i^*, M, \mathbb{I}\{S_{\text{AAS}} \geq \tau\}, D, \mathcal{V}), \quad (24)$$

where $i^* = \arg \max_i c_i$ is the dominant contribution and $\mathcal{V}$ identifies the version of the input transformation, weights, overlap matrix, gate thresholds, and supervisory threshold. Ties in i^* should be retained rather than resolved silently. Input timestamps, provenance, and missingness indicators should accompany this record in a deployed system.

For m dimensions and overlap edge set $E = \{(i, k) : i < k, q_{ik} > 0\}$, evaluation requires $O(m + |E|)$ arithmetic operations and $O(m)$ working memory when Q is stored sparsely. The six-dimensional demonstrator has seven non-zero edges. The operational sequence is:

1. validate bounds, dimensions, weights, symmetry, and the monotonicity condition;
2. compute x_i , p_i , and y_i for each component;
3. initialize $c_i = y_i$ and visit each non-zero edge once;
4. subtract half of its pairwise discount from each endpoint;
5. sum $\mathbf{c}$ to obtain S_{AAS} and verify numerical reconciliation;
6. evaluate the hard gate before the learned supervisory threshold; and
7. return the score, contributions, dominant evidence, gate provenance, decision, and configuration version.

The full language-neutral pseudocode appears in Appendix D. Validation should fail closed when a required input or configuration item is missing; imputation is a separate measurement decision and is not performed implicitly by the score.

5 Computational validation

5.1 Evaluation design

The evaluation tests internal mathematical behaviour and concordance with a controlled policy. It does not use observations from workers, robots, incidents, or workplaces. A scrambled Sobol sequence [24] generated $2^{16} = 65,536$ profiles in $[0, 1]^6$ using seed 20,260,726. Low-discrepancy sampling was chosen to cover the bounded space more evenly than an equally sized independent random sample.

The evaluation comprises five parts:

1. a boundary counterexample comparing the ratio adjustment with [equation \(15\)](#);
2. a coordinatewise monotonicity audit at perturbations 10^{-4} , 0.01, and 0.05;
3. held-out concordance with an independently defined rule policy;
4. parameter, input-noise, and weight sensitivity analyses; and
5. four worked HRC profiles for interpretability.

The protocol was fixed in code before the reported summaries were generated. [Table 3](#) separates quantities selected by the design from quantities estimated using the training split. This matters because the supervisory threshold is fitted, whereas the weights, overlap graph, hard gate, and reference policy are not.

Table 3 Controlled computational validation protocol.

Element	Specification	Role
Profile design	65,536 six-dimensional scrambled Sobol points; seed 20,260,726.	Deterministic space-filling evaluation set.
Model constants	$\eta = \epsilon = 10^{-6}$, $\gamma = 0.8$, $\kappa = 0.05$, weights in equation (2) , Q in equation (4) .	Fixed main specification.
Hard gate	$u_1 \geq 0.75$, $u_2 \geq 0.70$, or $u_3 \geq 0.80$.	Fixed external priority rule.
Reference labels	Non-additive rules in equations (25) to (27) .	Controlled construct benchmark.
Split	27,524 non-mandatory profiles, shuffled once and divided 1:1.	13,762 training and 13,762 test profiles.
Threshold	Maximum Youden J on training ROC for each scoring method.	Only fitted scalar per method.
Uncertainty	500 percentile bootstrap resamples of the held-out set.	Resampling stability, not population inference.
Monotonicity audit	Six positive coordinate perturbations at each of three step sizes.	Direct numerical property check.
Robustness	24 (γ, κ) settings; 60 noise runs; 250 weight draws.	Parameter and input sensitivity.

5.2 Independent reference policy

An independent rule policy supplies controlled labels for RQ2. It is not derived from any of the compared scores. Mandatory profiles first satisfy [equation \(22\)](#). Among remaining profiles, reference supervisory review is positive when at least one of the following holds:

$$u_1 \geq 0.70 \vee u_2 \geq 0.65 \vee u_3 \geq 0.75, \quad (25)$$

$$u_4 \geq 0.90 \vee u_5 \geq 0.90 \vee u_6 \geq 0.82, \quad (26)$$

$$\sum_{i=4}^6 \mathbb{I}(u_i \geq 0.72) \geq 2. \quad (27)$$

The policy encodes near-threshold safety accumulation and either single or paired operational demand. It was specified as a deliberately non-additive benchmark to test whether the scores rank profiles in a way that agrees with a transparent supervisory logic. It is not empirical ground truth and does not establish real-world predictive accuracy.

The Sobol set contained 38,012 mandatory profiles (58.00%), 13,847 non-mandatory supervisory profiles (21.13%), and 13,677 ordinary profiles (20.87%). The 27,524 non-mandatory profiles were shuffled with the fixed seed and split equally into training and test sets ($n = 13,762$ each). Only non-mandatory profiles entered threshold fitting and concordance metrics because mandatory outcomes are determined independently of the score.

5.3 Comparators and metrics

Three scoring methods were compared:

$$S_{\text{linear}}(\mathbf{u}) = \sum_i w_i u_i, \quad (28)$$

$$S_{\text{log}}(\mathbf{u}) = \sum_i y_i, \quad (29)$$

$$S_{\text{AAS}}(\mathbf{u}) \text{ as defined in } \a href="#">\text{equation (15)}. \quad (30)$$

For each method, a threshold was selected on the training set by maximizing Youden’s J (sensitivity plus specificity minus one) on the receiver operating characteristic (ROC) curve [\[25\]](#). The fixed threshold was then applied once to the held-out test set. Reported metrics are ROC area under the curve (AUC), balanced accuracy, F1, precision, sensitivity, and specificity. Percentile 95% confidence intervals for AUC and balanced accuracy use 500 bootstrap resamples of the test set [\[26\]](#). Because the labels are algorithmically generated rather than sampled clinical or workplace outcomes, these intervals quantify resampling stability within the controlled test distribution.

Let TP, TN, FP, and FN denote agreement counts with the controlled reference policy. The thresholded metrics were calculated as

$$\text{sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \quad \text{specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}}, \quad (31)$$

$$\text{precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \quad \text{balanced accuracy} = \frac{\text{sensitivity} + \text{specificity}}{2}, \quad (32)$$

$$\text{F1} = \frac{2 \text{precision sensitivity}}{\text{precision} + \text{sensitivity}}. \quad (33)$$

ROC AUC was computed from continuous scores and is therefore independent of the selected threshold. No profile from the test split was used for threshold selection.

5.4 Monotonicity audit and numerical tolerance

For each Sobol profile $\mathbf{u}$, coordinate i , and step $\delta \in \{10^{-4}, 0.01, 0.05\}$, the perturbed profile was

$$\mathbf{u}^{(i,\delta)} = (u_1, \dots, \min(1, u_i + \delta), \dots, u_6). \quad (34)$$

A violation was recorded when $S(\mathbf{u}^{(i,\delta)}) - S(\mathbf{u}) < -10^{-12}$. The tolerance prevents floating-point round-off at a theoretically zero difference from being counted as a decrease. Profiles already at the upper endpoint remain valid comparisons with a difference of zero. Both the discontinued ratio score and the corrected score were evaluated on exactly the same profile-perturbation pairs.

This audit is intentionally stronger than evaluating a small number of hand-selected trajectories, but it is not a proof over a continuum. The analytic guarantee in [theorem 1](#) provides that proof for admissible parameters; the audit verifies that the implementation conforms to it on a large deterministic design.

5.5 Sensitivity and robustness analyses

Parameter sensitivity crossed $\gamma \in \{0, 0.2, 0.4, 0.6, 0.8, 1.0\}$ with $\kappa \in \{0.01, 0.05, 0.10, 0.25\}$. Every combination satisfies $\gamma d_{\max} \leq 1$. A new training threshold was selected for each setting before test evaluation.

Input-noise robustness used 4,096 randomly selected profiles. Independent Gaussian noise with standard deviation $\sigma \in \{0.01, 0.03, 0.05\}$ was added to each coordinate and clipped to $[0, 1]$; 20 replicates were run per level. Mean absolute score change, its 95th percentile, and Spearman rank correlation were recorded.

Weight robustness used 2,048 profiles and 250 draws from a Dirichlet distribution with concentration $\boldsymbol{\alpha} = 200\mathbf{w}$. Each perturbed vector remained non-negative and summed to one. Spearman correlation between nominal and perturbed rankings summarizes stability while respecting that the illustrative weights are uncertain.

All analyses were implemented in Python 3 using NumPy, SciPy, pandas, scikit-learn, and Matplotlib. The script and full-precision outputs supplied as Online Resource 1 recreate every table, figure, and numerical result.

5.6 Reproducibility controls and evidential scope

Randomness is localized and seeded: Sobol scrambling, the non-mandatory split, bootstrap resampling, noise generation, profile subsampling, and Dirichlet draws each use the declared seed or a fixed offset. Intermediate outputs are written as plain CSV or JSON files. Figures are generated from those outputs by the same script. The validation code rejects out-of-range profiles, non-normalized weights, asymmetric overlap matrices, invalid overlap values, non-positive κ , and parameter combinations that violate the sufficient monotonicity condition.

The phrase “held-out” describes separation inside this computational design; it does not imply external validation. Likewise, confidence intervals do not convert synthetic rule labels into empirical outcomes. The reference policy is useful because it is independent of the three scoring formulas and contains coupled conditions that a simple weighted sum may not rank well. Its purpose is to test whether the aggregation behaves as intended under a transparent construct, not to claim that the policy itself is optimal.

6 Results

6.1 Boundary behaviour and monotonicity

Figure 1 shows the sparse boundary profile. With $u_3 = 0.999$, the ratio-adjusted score changes from 1.5941 at $u_1 = 0$ to 1.0627 for $u_1 = 10^{-12}$. The corrected score remains continuous (1.5941 at both displayed precision levels) and then increases with u_1 . The dashed line at 1.50 illustrates how the ratio discontinuity could reverse a high-burden classification solely because a negligible second input becomes active.

Table 4 gives the numerical trace. The corrected score changes by less than 3×10^{-13} between $u_1 = 0$ and 10^{-12} , consistent with continuity. The ratio score changes by approximately -0.5314 over the same perturbation.

Table 4 Numerical boundary counterexample with $u_3 = 0.999$ and all unlisted inputs equal to zero.

u_1	Legacy ratio score	Corrected score
0	1.594065	1.594065
10^{-12}	1.062710	1.594065
10^{-9}	1.062710	1.594065
10^{-6}	1.062710	1.594065
10^{-3}	1.062950	1.594332
10^{-2}	1.065126	1.596754
5×10^{-2}	1.075043	1.607835
10^{-1}	1.088044	1.622462

The exhaustive perturbation audit evaluated $65,536 \times 6 = 393,216$ coordinate changes at each step size. The ratio method produced 1,292, 1,202, and 990 decreases

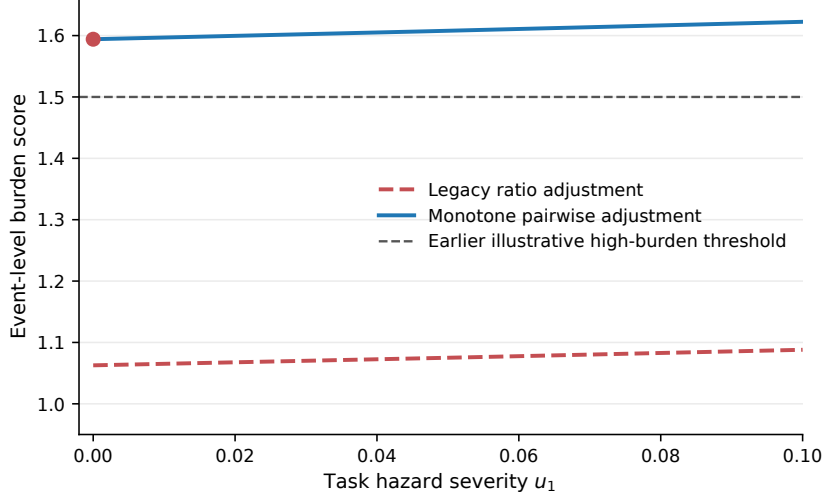


Fig. 1 Boundary counterexample. The activity-normalized ratio (red dashed curve) has an isolated value at $u_1 = 0$ and drops as an arbitrarily small overlapping input becomes active. The pairwise method (blue solid curve) is continuous and monotone. The horizontal dashed threshold is illustrative and is not the learned threshold used later

at steps 10^{-4} , 0.01, and 0.05, respectively—3,484 across 1,179,648 comparisons. The corrected method produced none (figure 2). The computational result is consistent with theorem 1; it supplements rather than replaces the proof.

Table 5 Detailed coordinatewise monotonicity audit. Decrease rate is relative to 393,216 comparisons per row.

Method	Step	Decreases	Rate (%)	Minimum ΔS
Legacy ratio	10^{-4}	1,292	0.3286	-1.8581×10^{-4}
Corrected pairwise	10^{-4}	0	0	5.3736×10^{-6}
Legacy ratio	0.01	1,202	0.3057	-0.017464
Corrected pairwise	0.01	0	0	0.000545
Legacy ratio	0.05	990	0.2518	-0.069621
Corrected pairwise	0.05	0	0	0.002878

6.2 Held-out policy concordance

Table 6 reports held-out performance. The corrected score reached ROC AUC 0.892 (95% CI 0.887–0.897), compared with 0.880 (0.874–0.885) for unadjusted logarithmic burden and 0.757 (0.749–0.765) for the linear score. Balanced accuracy was 0.800 (0.793–0.807), 0.788 (0.781–0.793), and 0.682 (0.674–0.690), respectively. The corrected method improved AUC by 0.012 and balanced accuracy by 0.013 over

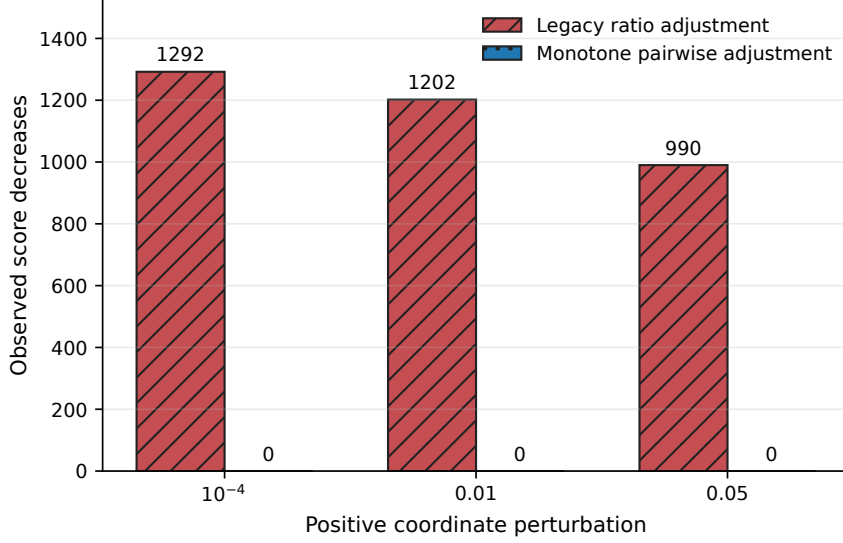


Fig. 2 Coordinatewise monotonicity audit. Patterned bars show the number of score decreases after increasing one burden input while holding the other five fixed. Each method and step size contains 393,216 comparisons

the unadjusted nonlinear baseline. This is a modest, distribution-specific improvement; the stronger contribution is that redundancy is incorporated without sacrificing monotonicity.

Table 6 Held-out concordance with the controlled reference policy ($n = 13,762$ non-mandatory profiles). Parentheses give 95% bootstrap confidence intervals. Thresholds were selected on a separate training set.

Method	Thresh.	AUC	Bal. acc.	F1	Prec.	Sens.	Spec.
Linear weighted sum	0.408	0.757 (0.749–0.765)	0.682 (0.674–0.690)	0.698	0.674	0.725	0.638
Unadjusted log burden	1.001	0.880 (0.874–0.885)	0.788 (0.781–0.793)	0.783	0.811	0.757	0.819
Monotone redundancy-aware AAS	0.885	0.892 (0.887–0.897)	0.800 (0.793–0.807)	0.801	0.809	0.793	0.807

At the learned threshold $\tau = 0.8847$, the corrected method produced 5,473 true negatives, 1,306 false positives, 1,444 false negatives, and 5,539 true positives. Precision was 0.809. These categories denote agreement or disagreement with the controlled rule policy only.

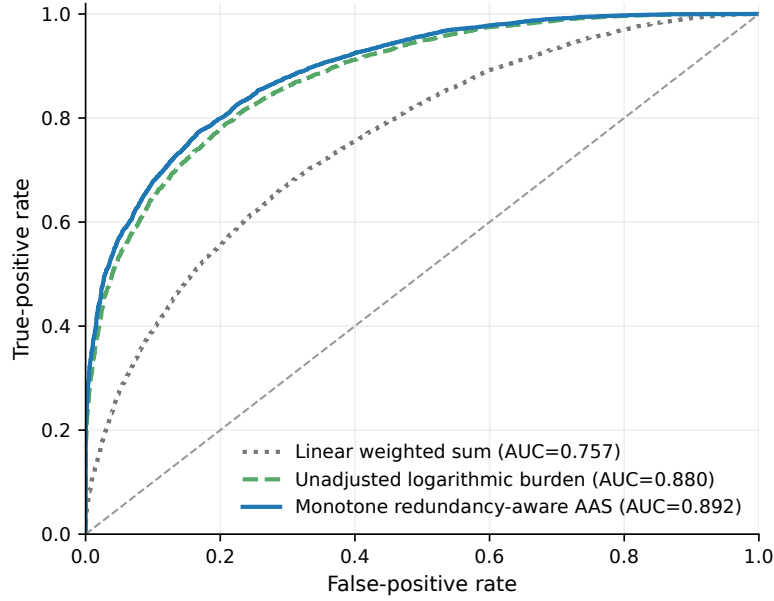


Fig. 3 ROC curves for the three scores on the held-out, non-mandatory test set. Line styles distinguish the methods independently of colour. Labels come from the independent rule policy in equations (25) to (27), not from observed workplace outcomes

Table 7 Held-out agreement counts at the training-selected thresholds. Positive denotes reference supervisory review among non-mandatory profiles.

Method	TN	FP	FN	TP
Linear weighted sum	4,325	2,454	1,919	5,064
Unadjusted logarithmic burden	5,550	1,229	1,699	5,284
Monotone redundancy-aware AAS	5,473	1,306	1,444	5,539

Relative to the unadjusted logarithmic score, the corrected method identified 255 additional positive reference profiles while producing 77 additional false positives. This trade-off raised sensitivity from 0.757 to 0.793 and lowered specificity from 0.819 to 0.807. The threshold was not chosen to privilege either error type; maximizing Youden’s J gives them equal weight. A real deployment would instead define error costs with stakeholders and could select a different operating point on the ROC curve.

6.3 Parameter sensitivity

Across the 24 parameter combinations, test ROC AUC ranged from 0.880 to 0.896 and balanced accuracy from 0.788 to 0.804. The highest AUC occurred at $\gamma = 1.0$, $\kappa = 0.01$, which lies on the zero-margin boundary of the sufficient monotonicity condition. The

main setting, $\gamma = 0.8, \kappa = 0.05$, retained a positive margin of 0.2 while achieving AUC 0.892 and balanced accuracy 0.800. This conservative choice trades a small amount of benchmark concordance for tolerance to later revisions of the overlap graph.

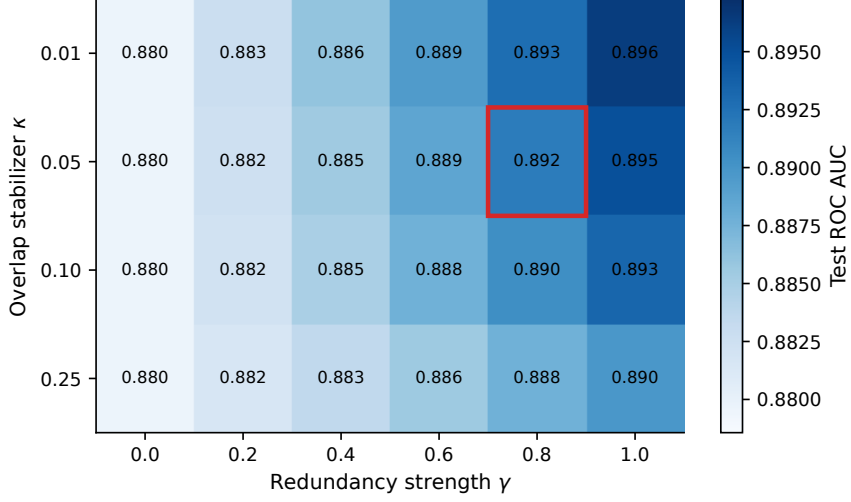


Fig. 4 Held-out ROC AUC across redundancy strength γ and smoothing parameter κ . Every cell satisfies the sufficient monotonicity condition; the red-outlined cell is the main specification

Table 8 provides the $\kappa = 0.05$ slice. Increasing γ progressively strengthened the pairwise discount and improved concordance within the tested benchmark. This pattern does not justify selecting $\gamma = 1$ automatically: at that endpoint the sufficient monotonicity margin is zero, so any later increase in graph degree would violate the condition unless γ were recalibrated.

Table 8 Sensitivity to redundancy strength at $\kappa = 0.05$. Each threshold was selected independently on the training split.

γ	Margin $1 - \gamma d_{\max}$	Threshold	Test AUC	Balanced acc.
0.0	1.0	1.0007	0.8796	0.7877
0.2	0.8	0.9700	0.8825	0.7917
0.4	0.6	0.9409	0.8855	0.7945
0.6	0.4	0.9131	0.8886	0.7981
0.8	0.2	0.8847	0.8917	0.8003
1.0	0.0	0.8668	0.8949	0.8038

6.4 Noise and weight robustness

Input perturbation preserved most rank order (figure 5). Mean Spearman correlation was 0.993 at $\sigma = 0.01$, 0.963 at $\sigma = 0.03$, and 0.926 at $\sigma = 0.05$; the minimum across 20 replicates was 0.991, 0.958, and 0.920, respectively. Mean absolute score change increased from 0.087 to 0.237 and 0.375. The comparatively large upper-tail changes at higher noise levels reflect the steep logarithmic penalty near $u_i = 1$. This feature appropriately signals extreme burden but also makes input calibration and uncertainty reporting important.

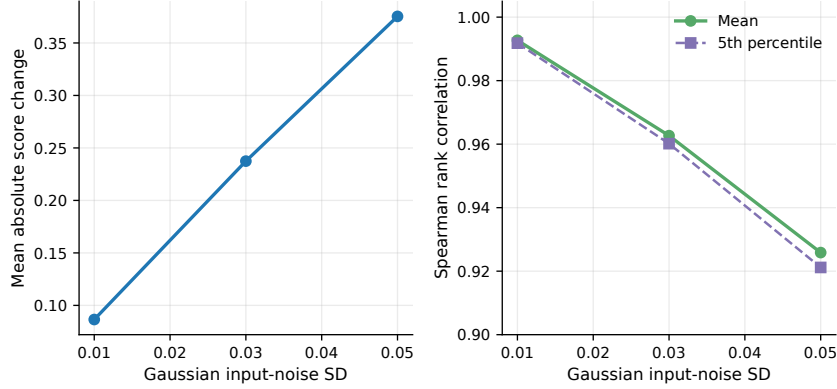


Fig. 5 Robustness to independent clipped Gaussian input noise. Markers show means across 20 replicates at each standard deviation; the dashed series in the right panel shows the fifth percentile of Spearman rank correlation

Table 9 Input-noise robustness summaries across 20 replicates per level. The 95th-percentile change is averaged across replicates.

σ	Mean $ \Delta S $	Mean 95th pct. $ \Delta S $	Mean ρ	Minimum ρ
0.01	0.0865	0.2634	0.9927	0.9909
0.03	0.2374	1.5755	0.9627	0.9580
0.05	0.3753	2.1748	0.9259	0.9203

For the 250 Dirichlet weight draws, the median Spearman correlation with the nominal ranking was 0.989, the 5th percentile was 0.971, and the minimum was 0.948. Rankings were therefore stable under moderate perturbations around the illustrative weights, although this test does not substitute for stakeholder elicitation.

The noise and weight analyses address different uncertainties. Noise changes event measurements profile by profile, while weight perturbation changes the relative importance assigned to every profile. High rank correlation under weight perturbation

indicates that the broad ordering is not an artefact of a single precise weight vector. It does not guarantee stability of individual decisions close to τ ; profiles near the threshold should be accompanied by uncertainty or review bands in future empirical work.

6.5 Worked event profiles

Table 10 illustrates how total burden, hard gating, and component attribution remain distinct. The routine profile stays below the learned threshold. High environmental uncertainty triggers mandatory review even though the other direct safety inputs are low. The operationally demanding profile triggers supervisory review through the score but not the hard gate. The near-threshold safety profile shows accumulated evidence below every mandatory threshold yet above the score threshold.

Table 10 Worked profiles using the main specification and learned supervisory threshold $\tau = 0.8847$. Inputs are listed in the order $(u_1, \dots, u_6)$.

Scenario	Profile	Score	Mandatory	Decision	Dominant contribution
Routine operation	(0.22, 0.12, 0.15, 0.34, 0.56, 0.10)	0.351	No	Ordinary operation	Exposure frequency (0.112)
High environmental uncertainty	(0.234, 0.11, 0.999, 0.56, 0.345, 0.20)	1.921	Yes	Mandatory human review	Env. uncertainty (1.577)
Operationally demanding event	(0.30, 0.25, 0.30, 0.92, 0.90, 0.85)	1.278	No	Supervisory review	Task criticality (0.447)
Near-threshold safety accumulation	(0.70, 0.65, 0.75, 0.30, 0.30, 0.30)	1.120	No	Supervisory review	Task hazard severity (0.379)

6.5.1 Routine operation

For $\mathbf{u} = (0.22, 0.12, 0.15, 0.34, 0.56, 0.10)$, the adjusted contribution vector was

$$\mathbf{c} = (0.0798, 0.0380, 0.0334, 0.0728, 0.1116, 0.0154), \quad (35)$$

which reconciles to $S_{\text{AAS}} = 0.3510$. Exposure frequency was the largest component, but the total remained below τ and all direct-safety values remained below their gates. The result is ordinary operation under the demonstrator. This output does not mean that exposure frequency is unimportant; it means that its transformed evidence, together with the other five components, did not reach the calibrated supervisory operating point.

6.5.2 High environmental uncertainty

For $\mathbf{u} = (0.234, 0.11, 0.999, 0.56, 0.345, 0.20)$, the contribution vector was

$$\mathbf{c} = (0.0749, 0.0311, 1.5774, 0.1508, 0.0541, 0.0332), \quad (36)$$

and $S_{\text{AAS}} = 1.9215$. Environmental uncertainty contributed approximately 82.1% of the total. More importantly, $u_3 = 0.999$ exceeded the illustrative mandatory threshold 0.80. The decision is therefore mandatory human review by gate provenance, not merely a score-based supervisory recommendation. The architecture retains this distinction even though the score is also above τ .

6.5.3 Operationally demanding event

For $\mathbf{u} = (0.30, 0.25, 0.30, 0.92, 0.90, 0.85)$, no direct-safety gate fired. The contributions were

$$\mathbf{c} = (0.1046, 0.0849, 0.0723, 0.4470, 0.2864, 0.2828), \quad (37)$$

giving $S_{\text{AAS}} = 1.2779$. Task criticality was dominant, but exposure frequency and intervention complexity were also substantial. This is the type of profile for which an aggregate monitoring layer is useful: no single direct-safety coordinate requires mandatory review, yet the joint operational evidence exceeds the learned supervisory threshold.

6.5.4 Near-threshold safety accumulation

The profile $\mathbf{u} = (0.70, 0.65, 0.75, 0.30, 0.30, 0.30)$ deliberately places the first three components below their mandatory thresholds. Its contribution vector was

$$\mathbf{c} = (0.3795, 0.3034, 0.2763, 0.0585, 0.0460, 0.0559), \quad (38)$$

and $S_{\text{AAS}} = 1.1196$. The event therefore receives supervisory review without a mandatory-gate claim. The example illustrates a central reason for retaining both mechanisms: thresholding each input independently can miss accumulated non-mandatory evidence, while aggregation alone should not replace direct-safety priority rules.

Appendix C gives u_i , x_i , p_i , y_i , allocated redundancy discount, and c_i for all 24 scenario-component combinations. These decompositions make clear that the dominant component is derived from adjusted evidence rather than from the raw input alone.

7 Discussion

7.1 Answers to the research questions

For RQ1, [theorem 1](#) and [corollary 1](#) establish that the proposed pairwise adjustment is continuous, coordinatewise non-decreasing, non-negative, and exactly attributable

when $\gamma d_{\max} \leq 1$. The boundary example shows why these guarantees matter: bounding a redundancy ratio is insufficient if the ratio changes discontinuously as another component becomes active. By making the discount scale with evidence on both sides of a pair, the new correction vanishes smoothly when either component vanishes.

For RQ2, the corrected score showed greater concordance with the controlled rule policy than both baselines. The improvement over the unadjusted logarithmic score was modest but consistent across AUC, balanced accuracy, F1, and sensitivity. This result should be interpreted as construct-level evidence: the pairwise term better approximated a benchmark containing coupled operational conditions. It is not evidence that the score predicts accidents, workload, appropriate intervention, or robot performance.

For RQ3, performance varied little across the tested parameter grid, and rankings were stable under moderate input and weight perturbation. Nevertheless, logarithmic scaling makes scores sensitive near the upper boundary. That sensitivity is transparent and monotone, but it raises a practical requirement: inputs close to one must have defensible measurement semantics and uncertainty bounds.

7.2 Comparison with alternative aggregation designs

The linear weighted sum is the easiest comparator to explain, but it imposes constant marginal effects and has no explicit representation of overlap. In the controlled policy benchmark, its lower AUC suggests that linear ordering did not capture the near-threshold and paired operational conditions well. The unadjusted logarithmic score adds curvature: high normalized burden receives disproportionately greater evidence. Its large improvement over the linear model indicates that the transformation itself accounts for most of the benchmark gain.

The corrected score adds a smaller second improvement by discounting declared overlap. This pattern is substantively reasonable. Redundancy adjustment should refine the evidence accounting, not dominate the underlying burden transformation. A very large performance gain caused only by a hand-specified overlap graph would raise concern that the benchmark and graph had been co-designed too closely.

The discontinued ratio method also represents overlap but normalizes it by the activity of all other inputs. Its boundedness initially appears attractive; however, the boundary analysis shows that bounding the ratio does not bound its sensitivity. The corrected method moves the normalization inside each pair, causing the overlap to scale with evidence at both endpoints.

A flexible supervised model could probably reproduce the reference policy more closely if trained directly on sufficiently many generated labels. That would answer a different question. The present objective is to formulate a small, auditable aggregation rule whose critical behaviour can be proved and whose components correspond to declared constructs. Predictive flexibility becomes valuable when empirical labels exist, but a learned model would still require monotonic constraints, calibration, explanation, and an independent safety gate.

7.3 Explainability and oversight implications

The method provides three layers of explanation. First, the six normalized inputs expose the event description. Second, component contributions in [equation \(20\)](#) reconcile exactly with the displayed total. Third, the decision output identifies whether review arose from the hard gate or the aggregate threshold. This supports a contrastive question a supervisor is likely to ask: “Was review required by a direct safety condition, or recommended because several non-mandatory burdens accumulated?”

The hard gate is essential. An aggregate model can discount redundant evidence, and the operational threshold can be tuned, but neither should cancel a direct safety condition. Keeping the gate external makes that design constraint explicit. It also avoids presenting AUC or balanced accuracy as evidence about the adequacy of protective measures.

An interface should avoid showing the dominant contribution alone. A single highlighted cause can conceal a close second contributor or a distributed accumulation. A minimum explanation panel would show all six input values, all six adjusted contributions, the total and threshold, the gate state, and the current configuration version. It should distinguish raw evidence y_i from adjusted contribution c_i so that users can see the effect of overlap. For mandatory review, the triggering input and its threshold should be presented before the aggregate score.

Explanations should also state uncertainty. A sensor-derived input could be displayed with a confidence interval or data-quality flag; an expert-assigned value could identify its assessor and time. The current method propagates only point values. Future work could evaluate interval-valued inputs by computing score bounds, which is feasible because coordinatewise monotonicity makes the lower and upper corners of an input box immediately interpretable.

7.4 Calibration and deployment pathway

The demonstrator’s dimensions, weights, overlap labels, and thresholds are hypotheses to be calibrated. A responsible application would proceed in stages:

1. define the operational decision and ensure that legal and safety responsibilities remain with authorized processes;
2. operationalize each u_i using documented evidence, uncertainty handling, and missing-data rules;
3. elicit weights and pairwise overlaps from multiple stakeholders, then test inter-rater agreement and sensitivity;
4. specify hard-gate conditions from the applicable risk assessment and safety architecture;
5. fit the supervisory threshold on representative historical or simulated cases with expert-adjudicated labels;
6. conduct prospective human-in-the-loop evaluation of comprehension, workload, reliance, response time, and failure recovery; and
7. monitor distribution shift, overrides, false alarms, missed cases, and parameter changes after deployment.

Where empirical dependence is available, Q could be estimated from expert judgments, co-occurrence, mutual information, or a causal model. Any data-driven matrix must still be symmetrized, bounded, documented, and checked against $\gamma d_{\max} \leq 1$. If the overlap graph changes, γ can be reduced to restore the guarantee.

7.5 Failure modes and governance controls

Several foreseeable failure modes arise outside the score formula. A stale input may produce a mathematically valid but operationally irrelevant result. A missing value silently encoded as zero may falsely reduce burden. Different teams may assign incompatible meanings to the same normalized value. A threshold optimized on a historical operating regime may drift when tasks, robots, staffing, or sensors change. Frequent false alarms may lead to habitual overrides, while an attractive dashboard may encourage inappropriate reliance.

Controls should therefore surround the calculation. Inputs require freshness limits and explicit missingness; configuration changes require review and versioning; every recommendation and override should be logged; and supervisors need a defined escalation action. Monitoring should stratify false alarms and missed cases by task, site, shift, and relevant human factors rather than reporting only a global accuracy value. A rollback mechanism is needed if a new weight or overlap configuration behaves unexpectedly.

The monotonicity guarantee itself should be tested as a configuration invariant. Before a revised Q or γ is activated, software can compute $d_{\max}$ and reject the configuration if $\gamma d_{\max} > 1$. This is a practical benefit of a sufficient condition that is simple to evaluate. It converts a mathematical property into a pre-deployment control.

7.6 Transferability and future extensions

Although the examples use warehouse language, the aggregation method is not tied to mobile robots. A medical, rehabilitation, laboratory, field, or manufacturing application could define different dimensions and an overlap graph suited to its supervisory decision. Transfer requires a new measurement model and validation study; replacing labels while retaining the numerical parameters would not constitute validation.

Temporal modelling is a natural extension. Repeated moderate events may require a persistence rule even when each event score is below τ . Possible designs include a moving maximum, an exponentially weighted burden history, or a separate alarm-fatigue state. Such dynamics should not be inserted into the present score without new monotonicity and stability analysis. Another extension is interval propagation. Because S_{AAS} is coordinatewise monotone, uncertain inputs $u_i \in [\underline{u}_i, \bar{u}_i]$ imply the exact enclosure

$$S_{\text{AAS}}(\underline{\mathbf{u}}) \leq S_{\text{AAS}}(\mathbf{u}) \leq S_{\text{AAS}}(\bar{\mathbf{u}}). \quad (39)$$

This could support a “review if the upper bound exceeds threshold” policy while keeping measurement uncertainty visible.

7.7 Limitations

The study has six principal limitations. First, all profiles and labels are synthetic; there are no worker, robot, incident, or field data. Second, the reference policy is a designed benchmark, so agreement can only support internal construct validity. Third, the six dimensions may omit important factors such as human workload, robot confidence, organizational context, fatigue, proximity dynamics, and recovery history. Fourth, the weights and feature-set overlaps are illustrative. Robustness around those values does not show that the values are correct. Fifth, Gaussian input noise does not represent every sensor error, calibration bias, missingness pattern, or adversarial condition. Sixth, the model operates on one event vector and does not yet represent temporal persistence, repeated alarms, adaptation, or team-level dependencies.

The score must not be interpreted as a probability of injury, a safety-integrity measure, an automated authorization to continue, or a substitute for standards-based risk reduction. The hard-gate thresholds in this paper are computational examples and have no normative status. Future work should test content validity with domain experts, estimate measurement reliability, compare alternative overlap estimators, perform ablation and external validation, and conduct human-factors experiments before any workplace trial.

8 Conclusion

This paper reformulated an AAS-derived decision-burden score for explainable human oversight in HRC. Replacing an activity-normalized redundancy ratio with smooth pairwise overlap removes a sparse-boundary discontinuity and yields a simple, auditable condition for coordinatewise monotonicity. The corrected score decomposes into non-negative component contributions and operates behind an independent mandatory-review gate. In a controlled Sobol study, it showed no decreases across more than one million coordinate perturbations and modestly improved held-out concordance with an independent, non-additive oversight policy.

The methodological advance is a predictable redundancy correction, not a claim of deployment readiness. The next evidential step is domain-specific measurement design and prospective human-in-the-loop validation. Until then, the method should be treated as a reproducible framework for research on transparent oversight support rather than a safety system.

Supplementary information

Online Resource 1 contains the deterministic validation script, exact environment specification, full-precision computational outputs, and all figure-generation files.

Acknowledgements

The author acknowledges receipt of an Australian Government Research Training Program (RTP) Scholarship.

Declarations

Funding

The author is supported by an Australian Government Research Training Program (RTP) Scholarship. No additional external project funding was received.

Competing interests

The author declares no competing interests.

Ethics approval and consent to participate

Not applicable. The study used only algorithmically generated computational profiles and did not involve human participants, animals, personal data, or workplace records.

Consent for publication

Not applicable.

Data availability

No external dataset was used. The computational data supporting the findings, parameter metadata, and full-precision result tables are provided as Online Resource 1. The supplied script regenerates the scrambled Sobol profiles from the reported seed.

Code availability

The deterministic validation and figure-generation code, exact tested dependencies, and reproduction instructions are provided as Online Resource 1.

Author contributions

Seyma Yaman Kayadibi: conceptualization, methodology, software, validation, formal analysis, visualization, writing—original draft, and writing—review and editing.

Appendix A Complete numerical specification

[Table A11](#) collects every constant needed to reproduce the main analysis. The distinction among fixed design quantities, fitted quantities, and generated summaries prevents a learned threshold from being confused with a safety limit.

Table A11: Complete numerical specification of the controlled study.

Item	Value	Interpretation
Event dimension	$m = 6$	Number of normalized burden inputs.
Weights	$(0.25, 0.23, 0.16, 0.14, 0.10, 0.12)$	Fixed illustrative priority vector; sums to one.
Consistency stabilizer	$\eta = 10^{-6}$	Keeps x_i positive at $u_i = 1$.
Penalty stabilizer	$\epsilon = 10^{-6}$	Regularizes the logarithmic kernel.
Overlap graph	Q in equation (4)	Fixed conceptual Jaccard prior.
Discount strength	$\gamma = 0.80$	Main pairwise redundancy strength.
Overlap stabilizer	$\kappa = 0.05$	Main pairwise smoothing value.
Maximum degree	$d_{\max} = 1$	Largest weighted overlap-graph degree.
Monotonicity margin	$1 - \gamma d_{\max} = 0.20$	Positive margin in the main specification.
Mandatory thresholds	$(0.75, 0.70, 0.80)$ for (u_1, u_2, u_3)	Demonstration gate; not a safety standard.
Supervisory threshold	$\tau = 0.8847153897$	Selected on training data for the corrected score.
Reference near-safety	$(0.70, 0.65, 0.75)$	Controlled benchmark condition.
Reference single operational	$(0.90, 0.90, 0.82)$	Controlled benchmark condition for (u_4, u_5, u_6) .
Reference paired operational	At least two of $u_4, u_5, u_6 \geq 0.72$	Controlled coupled benchmark condition.
Sampling	Scrambled Sobol, $2^{16} = 65,536$	Deterministic low-discrepancy profile design.
Primary seed	20,260,726	Sobol generation and split seed.
Train/test sizes	13,762/13,762	Non-mandatory profiles only.
Bootstrap	500 test-set resamples	Percentile intervals for AUC and balanced accuracy.
Audit steps	$10^{-4}, 0.01, 0.05$	Positive coordinate perturbations.
Audit tolerance	-10^{-12}	Difference below this value counts as a decrease.

Item	Value	Interpretation
Noise design	$\sigma = 0.01, 0.03, 0.05; 20$ replicates each	Independent Gaussian input perturbation with clipping.
Weight design	250 Dirichlet draws, $\alpha = 200\mathbf{w}$	Moderate normalized weight perturbation.

Appendix B General boundary result for the discontinued ratio

The counterexample in the main text is an instance of a general property. It does not depend on the particular values 0.999 or $1/3$.

Proposition 4 (Generic inactive-boundary discontinuity) *Consider the ratio-adjusted score*

$$S_{\text{ratio}}(\mathbf{u}) = \sum_{j=1}^m (1 - R_j(\mathbf{u})) y_j(u_j),$$

with R_j defined by [equation \(11\)](#). Suppose $q_{ik} > 0$, $w_i, w_k > 0$, and choose a profile for which $u_i > 0$ and every $u_j = 0$ for $j \neq i$. If $u_k = \delta > 0$ is introduced, then

$$\lim_{\delta \downarrow 0} S_{\text{ratio}}(\mathbf{u} + \delta \mathbf{e}_k) = S_{\text{ratio}}(\mathbf{u}) - q_{ik} y_i(u_i) \neq S_{\text{ratio}}(\mathbf{u}).$$

Proof At the original profile, the denominator defining R_i is zero, so $R_i = 0$. The only non-zero base evidence is $y_i(u_i)$ and therefore $S_{\text{ratio}}(\mathbf{u}) = y_i(u_i)$. For every $\delta > 0$, component k is the only active component other than i , giving

$$R_i(\mathbf{u} + \delta \mathbf{e}_k) = \frac{w_k \delta q_{ik}}{w_k \delta} = q_{ik}.$$

The contribution from k tends to zero because $y_k(\delta) \rightarrow y_k(0) = 0$. Every other contribution is zero. Hence the displayed limit equals $(1 - q_{ik})y_i(u_i)$, which differs from $y_i(u_i)$ because both factors are positive. $\square$

Corollary 2 (Possible threshold inversion) *For any threshold t strictly between $(1 - q_{ik})y_i(u_i)$ and $y_i(u_i)$, the profile at $\delta = 0$ and profiles with arbitrarily small $\delta > 0$ receive opposite threshold classifications.*

Proof The claim follows directly from [proposition 4](#) and the definition of a strict interval. $\square$

The pairwise correction avoids this result because $h_\kappa(y_i, y_k(\delta)) \rightarrow h_\kappa(y_i, 0) = 0$. It therefore preserves the limit of the unperturbed profile. This boundary behaviour is the reason the ratio formula is retained only as a comparator in the revised study.

Appendix C Detailed scenario decompositions

The following tables report every quantity in the explanation path. “Discount” denotes the half-allocated redundancy amount $y_i - c_i$. Rounding may cause the displayed contributions to differ from the reported total in the final decimal place.

Table C12 Routine-operation component decomposition; $S_{\text{AAS}} = 0.351013$.

Component	u_i	x_i	p_i	y_i	Discount	c_i
Task hazard severity	0.220	0.780000	0.358453	0.089613	0.009801	0.079813
Human safety risk	0.120	0.880000	0.184424	0.042418	0.004417	0.038000
Environmental uncertainty	0.150	0.850000	0.234465	0.037514	0.004164	0.033351
Task criticality	0.340	0.660000	0.599461	0.083924	0.011079	0.072846
Exposure frequency	0.560	0.440001	1.184421	0.118442	0.006795	0.111647
Intervention complexity	0.100	0.900000	0.152003	0.018240	0.002884	0.015356

Table C13 High-environmental-uncertainty component decomposition; $S_{\text{AAS}} = 1.921462$.

Component	u_i	x_i	p_i	y_i	Discount	c_i
Task hazard severity	0.234	0.766000	0.384583	0.096146	0.021239	0.074907
Human safety risk	0.110	0.890000	0.168122	0.038668	0.007566	0.031102
Environmental uncertainty	0.999	0.001001	9.962905	1.594065	0.016627	1.577438
Task criticality	0.560	0.440001	1.184421	0.165819	0.015045	0.150774
Exposure frequency	0.345	0.655000	0.610432	0.061043	0.006975	0.054068
Intervention complexity	0.200	0.800000	0.321927	0.038631	0.005457	0.033174

Table C14 Operationally demanding event decomposition; $S_{\text{AAS}} = 1.277949$.

Component	u_i	x_i	p_i	y_i	Discount	c_i
Task hazard severity	0.300	0.700000	0.514572	0.128643	0.024088	0.104555
Human safety risk	0.250	0.750000	0.415037	0.095458	0.010574	0.084885
Environmental uncertainty	0.300	0.700000	0.514572	0.082332	0.010011	0.072320
Task criticality	0.920	0.080001	3.643823	0.510135	0.063166	0.446969
Exposure frequency	0.900	0.100001	3.321902	0.332190	0.045792	0.286398
Intervention complexity	0.850	0.150001	2.736949	0.328434	0.045612	0.282822

Appendix D Language-neutral evaluation algorithm

Listing 1 Pseudocode for one monotone redundancy-aware HRC event evaluation.

INPUT:

Table C15 Near-threshold safety accumulation decomposition; $S_{AAS} = 1.119591$.

Component	u_i	x_i	p_i	y_i	Discount	c_i
Task hazard severity	0.700	0.300001	1.736959	0.434240	0.054760	0.379480
Human safety risk	0.650	0.350001	1.514568	0.348351	0.044915	0.303436
Environmental uncertainty	0.750	0.250001	1.999991	0.319999	0.043728	0.276271
Task criticality	0.300	0.700000	0.514572	0.072040	0.013574	0.058466
Exposure frequency	0.300	0.700000	0.514572	0.051457	0.005445	0.046013
Intervention complexity	0.300	0.700000	0.514572	0.061749	0.005823	0.055926

```

u[1..m]           normalized event profile
w[1..m]           non-negative weights summing to one
Q[1..m,1..m]      symmetric pairwise-overlap matrix
eta, epsilon      positive penalty stabilizers
gamma, kappa       redundancy parameters
hard_thresholds    external mandatory-review rules
tau               calibrated supervisory threshold
configuration_id   immutable parameter-version identifier

```

VALIDATE:

```

require length(u) = length(w) = m
require every u[i], Q[i,k] in [0,1]
require every w[i] >= 0 and sum(w) = 1
require Q = transpose(Q), diagonal(Q) = 0
require eta > 0, epsilon > 0, kappa > 0, gamma >= 0
degree[i] <- sum over k != i of Q[i,k]
require gamma * max(degree) <= 1

```

TRANSFORM:

```

for i in 1..m:
  x[i] <- (1 - u[i] + eta) / (1 + eta)
  p[i] <- -log2((x[i] + epsilon) / (1 + epsilon))
  y[i] <- w[i] * p[i]
  c[i] <- y[i]

```

PAIRWISE REDUNDANCY:

```

for every pair (i,k) with i < k and Q[i,k] > 0:
  overlap <- y[i] * y[k] / (y[i] + y[k] + kappa)
  discount <- gamma * Q[i,k] * overlap
  c[i] <- c[i] - discount / 2
  c[k] <- c[k] - discount / 2

```

SCORE AND EXPLANATION:

```

score <- sum(c)
require every c[i] >= -numerical_tolerance
require abs(score - sum(c)) <= numerical_tolerance
dominant <- all indices attaining max(c)

```

DECISION:

```

mandatory <- evaluate_hard_gate(u, hard_thresholds)
if mandatory:
  decision <- "mandatory human review"
  provenance <- "hard gate"
else if score >= tau:
  decision <- "supervisory review"
  provenance <- "aggregate threshold"
else:

```

```

decision <- "ordinary operation"
provenance <- "neither gate nor threshold"

RETURN:
u, x, p, y, c, score, dominant,
mandatory, score >= tau, decision, provenance,
configuration_id

```

For batch evaluation, the transformation can be vectorized across profiles while the seven non-zero demonstrator edges are processed sequentially or as array operations. The same validation conditions must be applied before a batch begins. The reference implementation additionally preserves the unrounded values in CSV outputs so that the manuscript’s rounded tables can be audited.

Appendix E Reproducibility checklist and file map

[Table E16](#) records the software versions used for the reported run. Later compatible versions may reproduce the results, but the versions below define the tested environment.

Table E16 Software environment used for the controlled validation.

Software	Version
Python	3.12.13
NumPy	2.3.5
SciPy	1.17.0
pandas	2.2.3
scikit-learn	1.8.0
Matplotlib	3.10.8

Online Resource 1 contains:

README.md reproduction instructions and file map;
requirements.txt exact tested Python-package versions;
validate_framework.py model, validation, tables, and figure generation;
run_metadata.json seed, sample size, model parameters, degrees, and split sizes;
policy_counts.csv generated mandatory, supervisory, and ordinary profile counts;
score_summary.csv full-sample score distribution summaries;
baseline_performance.csv thresholds, held-out metrics, intervals, and agreement counts;
boundary_counterexample.csv sparse-boundary numerical trace;
monotonicity_audit.csv all perturbation counts and score-change extrema;
sensitivity_grid.csv 24 parameter combinations;
noise_robustness.csv 60 noise-analysis records;

weight_robustness.csv 250 Dirichlet perturbation records;
scenario_outputs.csv event-level scenario decisions;
scenario_component_details.csv all 24 component decompositions; and
figures/ the five figures in vector and raster formats.

An independent reproduction should:

1. create a clean environment and record dependency versions;
2. run the script without editing its constants;
3. confirm 65,536 generated profiles and the three policy-class counts;
4. confirm equal non-mandatory training and test sizes of 13,762;
5. confirm the corrected threshold 0.8847153897;
6. confirm corrected test AUC 0.8917169438 and balanced accuracy 0.8002791514;
7. confirm zero corrected decreases and 3,484 legacy decreases across all audit steps;
8. compare scenario scores and decisions with [table 10](#);
9. regenerate all five figures; and
10. retain full-precision output files rather than comparing rounded manuscript values only.

References

- [1] Sheridan TB. *Telerobotics, Automation, and Human Supervisory Control*. Cambridge: MIT Press; 1992.
- [2] Parasuraman R, Sheridan TB, Wickens CD. A model for types and levels of human interaction with automation. *IEEE Trans Syst Man Cybern A*. 2000;30(3):286–297. <https://doi.org/10.1109/3468.844354>
- [3] Goodrich MA, Schultz AC. Human–robot interaction: a survey. *Found Trends Hum Comput Interact*. 2007;1(3):203–275. <https://doi.org/10.1561/1100000005>
- [4] Ajoudani A, Zanchettin AM, Ivaldi S, Albu-Schäffer A, Kosuge K, Khatib O. Progress and prospects of the human–robot collaboration. *Auton Robots*. 2018;42(5):957–975. <https://doi.org/10.1007/s10514-017-9677-2>
- [5] Villani V, Pini F, Leali F, Secchi C. Survey on human–robot collaboration in industrial settings: safety, intuitive interfaces and applications. *Mechatronics*. 2018;55:248–266. <https://doi.org/10.1016/j.mechatronics.2018.02.009>
- [6] Lasota PA, Fong T, Shah JA. A survey of methods for safe human–robot interaction. *Found Trends Robot*. 2017;5(4):261–349. <https://doi.org/10.1561/23000000052>
- [7] Lee JD, See KA. Trust in automation: designing for appropriate reliance. *Hum Factors*. 2004;46(1):50–80. <https://doi.org/10.1518/hfes.46.1.50.30392>
- [8] Hancock PA, Billings DR, Schaefer KE, Chen JYC, de Visser EJ, Parasuraman R. A meta-analysis of factors affecting trust in human–robot interaction. *Hum Factors*. 2011;53(5):517–527. <https://doi.org/10.1177/0018720811417254>

- [9] Miller T. Explanation in artificial intelligence: insights from the social sciences. *Artif Intell.* 2019;267:1–38. <https://doi.org/10.1016/j.artint.2018.07.007>
- [10] Adadi A, Berrada M. Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). *IEEE Access.* 2018;6:52138–52160. <https://doi.org/10.1109/ACCESS.2018.2870052>
- [11] International Organization for Standardization. ISO 12100:2010: Safety of machinery—general principles for design—risk assessment and risk reduction. Geneva: ISO; 2010. <https://www.iso.org/standard/51528.html>
- [12] International Organization for Standardization. ISO/TS 15066:2016: Robots and robotic devices—collaborative robots. Geneva: ISO; 2016. <https://www.iso.org/standard/62996.html>
- [13] National Institute of Standards and Technology. Artificial Intelligence Risk Management Framework (AI RMF 1.0). NIST AI 100-1. Gaithersburg: NIST; 2023. <https://doi.org/10.6028/NIST.AI.100-1>
- [14] Kayadibi SY. Redundancy-as-masking: formalizing the Artificial Age Score (AAS) to model memory aging in generative AI. *Front Artif Intell.* 2026;9:1732691. <https://doi.org/10.3389/frai.2026.1732691>
- [15] Endsley MR. Toward a theory of situation awareness in dynamic systems. *Hum Factors.* 1995;37(1):32–64. <https://doi.org/10.1518/001872095779049543>
- [16] De Santis A, Siciliano B, De Luca A, Bicchi A. An atlas of physical human–robot interaction. *Mech Mach Theory.* 2008;43(3):253–270. <https://doi.org/10.1016/j.mechmachtheory.2007.03.003>
- [17] Hopko S, Wang J, Mehta R. Human factors considerations and metrics in shared space human–robot collaboration: a systematic review. *Front Robot AI.* 2022;9:799522. <https://doi.org/10.3389/frobt.2022.799522>
- [18] Giallanza A, La Scalia G, Micale R, La Fata CM. Occupational health and safety issues in human–robot collaboration: state of the art and open challenges. *Saf Sci.* 2024;169:106313. <https://doi.org/10.1016/j.ssci.2023.106313>
- [19] Hayes B, Shah JA. Improving robot controller transparency through autonomous policy explanation. In: *Proceedings of the 2017 ACM/IEEE International Conference on Human–Robot Interaction*. New York: ACM; 2017. p. 303–312. <https://doi.org/10.1145/2909824.3020233>
- [20] Amershi S, Weld D, Vorvoreanu M, Fournay A, Nushi B, Collisson P, et al. Guidelines for human–AI interaction. In: *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. New York: ACM; 2019. p. 1–13. <https://doi.org/10.1145/3290605.3300233>

- [21] Shneiderman B. Human-centered artificial intelligence: reliable, safe & trustworthy. *Int J Hum Comput Interact*. 2020;36(6):495–504. <https://doi.org/10.1080/10447318.2020.1741118>
- [22] Ehsan U, Liao QV, Muller M, Riedl MO, Weisz JD. Expanding explainability: towards social transparency in AI systems. In: *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. New York: ACM; 2021. Article 82. <https://doi.org/10.1145/3411764.3445188>
- [23] Belton V, Stewart TJ. *Multiple Criteria Decision Analysis: An Integrated Approach*. Boston: Kluwer Academic Publishers; 2002.
- [24] Sobol IM. On the distribution of points in a cube and the approximate evaluation of integrals. *USSR Comput Math Math Phys*. 1967;7(4):86–112. [https://doi.org/10.1016/0041-5553\(67\)90144-9](https://doi.org/10.1016/0041-5553(67)90144-9)
- [25] Fawcett T. An introduction to ROC analysis. *Pattern Recognit Lett*. 2006;27(8):861–874. <https://doi.org/10.1016/j.patrec.2005.10.010>
- [26] Efron B, Tibshirani RJ. *An Introduction to the Bootstrap*. New York: Chapman & Hall; 1993.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [DiscoverRoboticsESM1.zip](#)