Diagnostic Sciences

A Formal Theory of Perception Extension Inside MAVS-GC

Prepared as a MAVS Research Program addition

Version 0.1 - July 2026

Core Thesis
Diagnostic Sciences is the formal study of diagnostic functions as perception-extending sensors inside MAVS-GC. It does not replace MAVS. It explains, evaluates, composes, and optimizes the diagnostic layer G so that MAVS can detect richer failure-relevant structure and govern specialist outputs more intelligently.

Additive Relationship to MAVS-GC
MAVS-GC already defines G, z, A, Theta, W, R, and Pi. Diagnostic Sciences adds a theory for what makes G good: intended scope, marginal diagnostic value, perception extension, influence pathway, breadth/specificity control, set composition, and objective-relative optimization.

Table of Contents

1. Executive Summary

Diagnostic Sciences is the missing theory of the MAVS-GC diagnostic layer. MAVS-GC already separates specialist prediction from governance: specialists produce evidence, diagnostics produce warning signals, severity is aggregated, weights and mitigation adjust the governed consensus, and a final policy accepts or rejects. Diagnostic Sciences asks what makes those diagnostics scientifically good.

The key move is to treat diagnostics not as arbitrary flags, but as sensors. A diagnostic function g_j transforms latent failure evidence into an observable governance signal z_j. The full diagnostic system G is therefore a sensor array. The quality of that array is not measured by how many warnings it emits, but by how much nonredundant, objective-aligned perception it adds to MAVS.

This addition is intentionally compatible with MAVS-GC. It keeps the original tuple M = (X, Phi, F, G, A, W, P, Theta, Pi), keeps the governance decision rule Pi, and concentrates on the internal science of G: how diagnostics are defined, scoped, measured, combined, tuned, and optimized under objectives such as accuracy, robustness, safety, recall, precision, stability, and calibration.

Central Doctrine
Diagnostic quality is not signal magnitude. Diagnostic quality is objective-aligned perception extension: a diagnostic is high quality when it detects intended hidden structure and beneficially changes MAVS behavior for a chosen objective x while minimizing unintended damage.

The strongest existing MAVS-GC evidence layer, Chapter 10B, matters here because it shows why this theory is needed. Under specialist failure and high corruption, MAVS-GC can fail more safely than aggregation baselines by suppressing unsafe acceptance. Diagnostic Sciences formalizes the design question behind that behavior: which diagnostics cause that safer failure mode, when do they overreject, and how should G be optimized for a target objective?

2. MAVS-GC Substrate

Diagnostic Sciences begins from the existing MAVS-GC formal system. The system tuple is:

$$M = (X, Phi, F, G, A, W, P, Theta, Pi)$$

The core pipeline is:

$$x \to Phi(x)=phi \to \{f\_i(phi)\} \to r\_i \to z \to a \to w\_i,m \to theta \to R \to Pi$$

Symbol	Meaning in MAVS-GC
X	Input space.
Phi	Shared feature or representation map.
F={f_i}	Always-on specialist / predictive layer.
G={g_j}	Diagnostic system / perceptual layer.
A	Severity aggregator over diagnostic signals.
W	Contextual specialist rebalancing function.
P	Bounded mitigation evidence functions.
Theta	Governed threshold policy.
Pi	Final acceptance decision rule.

The MAVS-GC definitions used throughout this document are:

$$s\_i = f\_i(phi) \text{ in } [0,1]$$

$$r\_i = 2s\_i - 1 \text{ in } [-1,1]$$

$$z = (g\_1(phi), ..., g\_k(phi)) \text{ in } R\_\{>=0\}\text{^}k$$

$$a = A(z), \text{ where } A \text{ is monotone}$$

$$w\_i = W(i, phi, z)$$

$$m = max\_l\ p\_l(phi) \text{ in } [0,1]$$

$$theta = Theta(a,m) = theta\_0 + lambda\ a - delta\ m$$

$$R(x) = sum\_i\ w\_i\ r\_i$$

$$Pi(R,theta,a) = 1[a < tau\_hard] * 1[R >= theta]$$

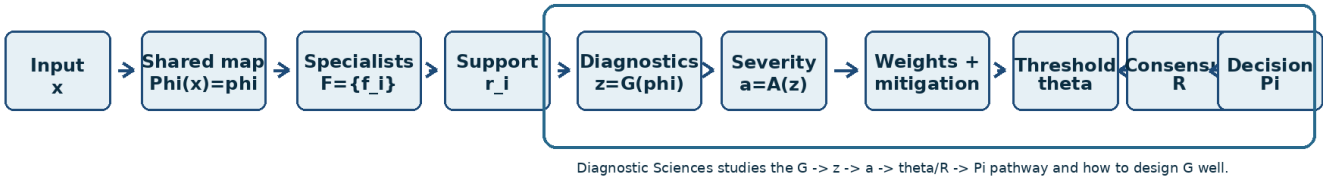
Equivalently, Pi = 0 when a >= tau_hard; otherwise Pi = 1[R >= theta]. The auditable trace is:

$$(r, w, z, a, m, theta, tau\_hard, R, Pi)$$

MAVS-GC Properties Used by Diagnostic Sciences
All-speak evaluation: every specialist evaluates every input. Monotone safety: higher diagnostic severity cannot make acceptance easier. Bounded mitigation: mitigation cannot override the hard veto because the hard veto is inside Pi. Governance separability: A or Theta can change acceptance without retraining specialists.

The MAVS-GC research overview frames the architecture as governance-first: the claim is not universal single-model accuracy, but better system behavior when evidence becomes uncertain, contradictory, corrupted, or unstable. Diagnostic Sciences therefore belongs exactly at the G layer and its downstream governance pathway.

Figure 1. MAVS-GC core pipeline: prediction is separated from governance



3. Why Diagnostic Sciences Is Needed

MAVS-GC already contains diagnostics, but the original formalism does not by itself answer the higher-order design problem: what makes a diagnostic good, how broad should it be, when is it redundant, when does it damage recall, and how should a diagnostic set G be optimized for a target objective?

The Diagnostic Sciences layer answers these questions without changing the MAVS architecture. It treats G as a scientific object rather than merely an implementation detail. A diagnostic function is not judged by whether it looks intuitively useful. It is judged by its intended scope, perception extension, calibration, influence pathway, objective gain, redundancy, and instability.

The Diagnostic Sciences Question
What makes G good? What makes a single diagnostic g_j good? How do we design, compare, compose, and optimize diagnostic functions so MAVS becomes better at some target x, such as accuracy, robustness, safety, recall, precision, stability, or calibration?

This matters because MAVS-GC is better interpreted as a governance framework than as a fixed detector set. The current benchmarks evaluate one concrete configuration, not the entire space of possible diagnostics, severity operators, mitigation strategies, rebalancers, and threshold policies. Diagnostic Sciences is the formal layer that studies that broader design space.

4. Diagnostic Functions as MAVS Sensors

The first principle is that diagnostics are MAVS sensors. In the MAVS tuple, F is the specialist/predictive layer while G is the diagnostic/perceptual layer. The predictive layer produces answer evidence. The diagnostic layer produces condition evidence.

$$F \Rightarrow \text{answer evidence}$$

$$G \Rightarrow \text{condition evidence}$$

The diagnostic layer does not merely predict the answer. It detects the state of the input, model agreement, corruption, uncertainty, contradiction, instability, adversarialness, anomaly, or domain-risk. Therefore, G expands MAVS diagnostic perception over phi.

$$g_j : \text{Phi}(X) \to R_{\geq 0}$$

Each diagnostic function maps a representation to a nonnegative observable signal. It transforms hidden or latent failure evidence into a governance-readable quantity:

$$\text{phi} \to z_j$$

The full diagnostic system is a sensor array:

$$G(\text{phi}) = (g_1(\text{phi}), \ldots, g_k(\text{phi})) = z$$

Severity aggregation and thresholding then turn perception into governance pressure:

$$a = A(z)$$

$$\text{theta} = \text{theta}_0 + \text{lambda}\ a - \text{delta}\ m$$

Good Sensors vs. Bad Sensors
Adding diagnostics always expands dimensionality, but it does not always expand perception. Bad sensors expand z-space without adding useful information. Good sensors add nonredundant information about failure-relevant hidden structure. This distinction is the birth of Diagnostic Sciences.

5. Correlation Failure and Objective-Relative Quality

Correlation failure is the case where all or most specialists share the same flawed understanding of the input. Their supports r_1, r_2, ..., r_N may point in the same wrong direction. A normal ensemble may interpret agreement as confidence, but MAVS can treat agreement as suspicious when diagnostics fire.

$$r_1, r_2, \ldots, r_N \text{ are correlated in the wrong direction}$$

$$\text{agreement} \Rightarrow \text{accept} \qquad [\text{ordinary ensemble behavior}]$$

$$z_{corr} = g_{corr}(\text{phi}, r, F)$$

$$z_{corr} \text{ increases} \to a \text{ increases} \to \text{theta increases}$$

If severity crosses the hard-veto boundary, MAVS shuts down acceptance:

$$a >= tau\_hard \rightarrow Pi = 0$$

This can produce the observed pattern from the diagnostic draft: false positive acceptance can drop to 0 while false rejection can rise to 100. This is not automatically good or bad. It is objective-dependent.

Objective	Interpretation of a correlation-failure shutdown
Safety-critical governance	Potentially excellent, because unsafe acceptance is minimized.
High-recall system	Potentially unacceptable, because too many valid cases are rejected.
Balanced deployment	Requires tuning the diagnostic threshold, hard veto, mitigation, and cost weights.

Objective-Relative Quality
Diagnostic quality is objective-relative. A diagnostic can be excellent for safety and terrible for recall. Therefore Diagnostic Sciences evaluates diagnostics relative to x in {accuracy, robustness, safety, recall, precision, stability, calibration, ...}.

6. Formal Definition of Diagnostic Sciences

Let the Diagnostic Sciences layer be represented as:

$$D\_frak = (D, S, Q, O, B)$$

Component	Role
D	The space of possible diagnostic functions.
S	The scope theory defining where each diagnostic is meant to apply.
Q	The quality theory measuring whether a diagnostic is good.
O	The optimization theory for selecting and tuning diagnostics.
B	The breadth/specificity control theory.

A diagnostic function can be defined in the basic MAVS case as:

$$g : Phi(X) \rightarrow R\_\{>=0\}$$

More generally, the diagnostic can inspect an evidence space E:

$$g : E \rightarrow R\_\{>=0\}$$

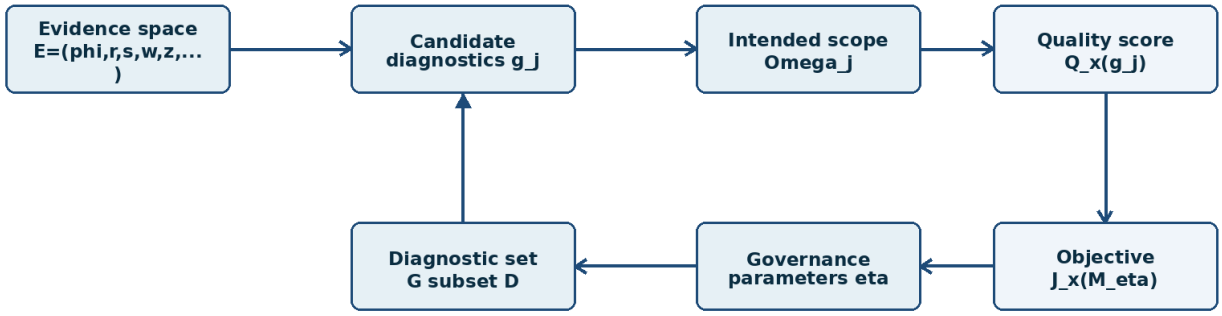
$$E = (phi, r, s, w, z\_not\_j, metadata, trace\ history)$$

The richer evidence space allows diagnostics to inspect the shared representation phi, specialist supports r, raw specialist scores s, specialist weights w, other diagnostics z_not_j, metadata, and trace/history T. The diagnostic output is:

$$g\_j(e) = z\_j$$

Signal magnitude	Interpretation
z_j = 0	No detected issue.
z_j > 0	Some detected diagnostic signal.
z_j >> 0	Strong detected signal.

Figure 2. Diagnostic Sciences as an additive theory over MAVS-GC diagnostics



Core loop: define the objective, define failure modes, generate diagnostics, measure scope-aware quality, compose G, tune governance, evaluate whole MAVS.

7. Intended Scope Theory

A diagnostic function should not be judged in the abstract. It must be judged against its intended scope. This is one of the central additions of Diagnostic Sciences: a diagnostic is only meaningful if the system knows what failure family it is supposed to detect, where it applies, what response it should trigger, and what influence it is intended to have.

$$\text{Omega\_j} = (\text{F\_j}, \text{C\_j}, \text{R\_j}, \text{I\_j})$$

Scope component	Meaning
F_j	The failure-mode family the diagnostic is meant to detect.
C_j	The context or domain where it applies.
R_j	The response behavior it is supposed to trigger.
I_j	The intended influence on MAVS.

Example: Correlation Failure Diagnostic
Diagnostic: g_corr F_corr: specialists agree due to a shared flawed representation. C_corr: high agreement but low independent evidence or diversity. R_corr: raise severity sharply; possibly hard veto. I_corr: reduce unsafe acceptance under correlated failure.

This scope definition immediately gives quality language. A diagnostic is bad if it fires outside its intended scope too often. It is weak if it fails inside its intended scope. It is dangerous if its influence does not match its evidence quality.

8. Diagnostic Quality and Intended Influence

The diagnostic draft proposes a strong definition: a diagnostic is high-quality according to how much intended influence it has on the overall MAVS architecture, meaning how much it extends MAVS perception

to detect more information or signals. Diagnostic Sciences formalizes that idea using marginal system value and unintended damage.

Let M be the MAVS system without diagnostic g_j, and let M^{+g_j} be MAVS with g_j added. The diagnostic quality of g_j for objective x begins as marginal value:

$$Q_x(g_j) = J_x(M^{+g_j}) - J_x(M)$$

For common objectives:

$$Q_{rob}(g_j) = J_{rob}(M^{+g_j}) - J_{rob}(M)$$

$$Q_{safety}(g_j) = J_{safety}(M^{+g_j}) - J_{safety}(M)$$

$$Q_{acc}(g_j) = J_{acc}(M^{+g_j}) - J_{acc}(M)$$

However, decision influence alone is not enough. A diagnostic can change decisions frequently while making the system worse. Define decision influence as:

$$I(g_j;M) = E[\ |\ Pi_{M^{+g_j}}(x) - Pi_M(x)\ |\ ]$$

Influence must be separated into intended influence and unintended damage. Therefore the refined quality score is:

$$Q_x(g_j) = Delta\ J_x(g_j) - gamma\ U(g_j)$$

where Delta J_x(g_j) is the objective gain caused by adding g_j, and U(g_j) is unintended influence or damage.

Unintended influence examples	Description
False rejection inflation	The diagnostic causes too many acceptable cases to be rejected.
Accuracy collapse	The diagnostic improves safety but destroys general correctness beyond the objective budget.
Noise sensitivity	Small irrelevant perturbations cause unstable diagnostic firing.
Redundancy	The diagnostic repeats information already captured by other diagnostics.
Unnecessary severity inflation	The diagnostic raises a without meaningful failure evidence.
Instability under small perturbations	The diagnostic introduces governance volatility.

Refined Quality Definition
Q_x(g_j) = objective gain caused by g_j within its intended scope - damage caused outside its intended scope. This turns Diagnostic Sciences into a system-level theory rather than a local detector-scoring heuristic.

9. Perception Extension

Perception extension measures whether a diagnostic adds relevant signal beyond what MAVS already observes. Let MAVS without g_j observe z_{-j}. With g_j, the diagnostic state becomes z = (z_{-j}, z_j).

$$z = (z_{-j}, z_j)$$

Let Y_{Omega_j} be the latent truth of whether the intended failure mode is present. The diagnostic perception extension of g_j is:

$$PE(g_j) = I(Y_{Omega_j}; z_j \mid z_{-j})$$

This is conditional information gain. If PE(g_j) is high, g_j tells MAVS something useful that the existing diagnostics did not already tell it. If it is approximately zero, then g_j is redundant.

$$I(Y_{Omega_j}; z_j \mid z_{-j}) \approx 0 \rightarrow g_j \text{ is redundant}$$

$$I(Y_{Omega_j}; z_j \mid z_{-j}) > 0 \rightarrow g_j \text{ extends perception}$$

Sensor Interpretation
A sensor is high-quality if it detects relevant latent structure not already captured by the system. This formalizes the intuition that Diagnostic Sciences is the science of MAVS perception extension.

10. Diagnostic Influence Pathway

A diagnostic influences MAVS through two major channels: severity and rebalancing. The severity channel is the direct governance pathway from z_j to a, theta, and Pi.

$$g_j(phi) = z_j \rightarrow A(z) = a \rightarrow Theta(a,m) = theta \rightarrow Pi$$

The rebalancing channel occurs when diagnostics influence specialist weights and therefore consensus:

$$z_j \rightarrow W(i,phi,z) = w_i \rightarrow R \rightarrow Pi$$

Total diagnostic influence can therefore be decomposed into threshold influence and consensus influence:

$$I_{total}(g_j) = I_{theta}(g_j) + I_R(g_j)$$

$$I_{theta}(g_j) = E[\, | theta^{+g_j} - theta^{-g_j} | \,]$$

$$I_R(g_j) = E[\, | R^{+g_j} - R^{-g_j} | \,]$$

The final decision influence is:

$$I_{Pi}(g_j) = P[\, Pi^{+g_j}(x) \neq Pi^{-g_j}(x) \,]$$

A diagnostic is architecturally meaningful if it has nonzero influence on Pi, theta, or R. But architectural meaning is not the same as quality; the influence must be aligned with the diagnostic intended scope and the chosen objective.

High-Quality Diagnostic
High-quality diagnostic = perception extension + correct intended influence. It must add new information and move the MAVS governance process in the right direction under the chosen objective.

11. Breadth, Specificity, and Control Modes

Diagnostic breadth measures how often a diagnostic fires. For a threshold tau:

$$B(g_j; tau) = P[g_j(e) \geq tau]$$

Specificity measures how concentrated the diagnostic is on the intended failure mode:

$$S(g_j) = P[Y_{Omega_j}=1 \mid g_j(e) \geq tau]$$

Sensitivity or recall measures whether it fires when the intended failure mode is present:

$$R_c(g_j) = P[g_j(e) \geq tau \mid Y_{Omega_j}=1]$$

Quantity	Definition	Interpretation
Breadth	P[fires]	How often the diagnostic activates.

Specificity	P[true intended failure | fires]	How pure the firing set is.
Sensitivity	P[fires | true intended failure]	How much intended failure coverage it has.

Diagnostics can be parameterized by rho or by firing threshold tau_j. Lowering tau_j increases breadth but usually increases false alarms. Raising tau_j increases specificity but can miss true failures.

$$\text{tau\_j decreases -> } B(g\_j;\ \text{tau\_j}) \text{ increases}$$

$$\text{tau\_j increases -> } S(g\_j;\ \text{tau\_j}) \text{ increases}$$

Mode	Formal tendency	Use case
Maximum perception mode	maximize B(g_j)	Use when missing danger is worse than false rejection.
Specificity mode	maximize S(g_j)	Use when false rejection is costly.
Balanced mode	maximize F_beta(g_j)	Use when both false alarms and misses matter.

$$\text{F\_beta} = (1+\text{beta}\hat{}2) * (\text{specificity} * \text{sensitivity}) \ / \ (\text{beta}\hat{}2 * \text{specificity} + \text{sensitivity})$$

Breadth-Specificity Tradeoff
tau_j down -> more perception and more false alarms. tau_j up -> less perception and higher precision. This gives Diagnostic Sciences a control surface rather than a fixed diagnostic flag.

12. Intended Scope Quality Score

For each diagnostic g_j, define Y_j(x)=1 if the intended failure mode of g_j is present. Diagnostic quality can then be decomposed into measurable terms.

$$Y\_j(x)=1 \text{ iff the intended failure mode of } g\_j \text{ is present}$$

Metric	Formula	Question answered
Scope sensitivity	Sens(g_j)=P[g_j(e)>=tau_j | Y_j=1]	Does it detect what it is supposed to detect?
Scope specificity	Spec(g_j)=P[g_j(e)<tau_j | Y_j=0]	Does it stay quiet when it should?
Severity calibration	Cal(g_j)=1-E[|g_j(e)-q_j(x)|]	Does signal magnitude match true severity?
Marginal perception gain	MPG(g_j)=I(Y_j;z_j | z_{-j})	Does it add new information?
Intended decision influence	IDI(g_j)=P[Pi^{+g_j} != Pi^{-g_j} | Y_j=1]	Does it affect MAVS when it should?
Unintended decision influence	UDI(g_j)=P[Pi^{+g_j} != Pi^{-g_j} | Y_j=0]	Does it affect MAVS when it should not?

The overall diagnostic quality score can be written as:

$$Q(g\_j) = \text{alpha\_1 Sens} + \text{alpha\_2 Spec} + \text{alpha\_3 Cal} + \text{alpha\_4 MPG} + \text{alpha\_5 IDI - alpha\_6 UDI - alpha\_7 Redundancy - alpha\_8 Instability}$$

Interpretation of the Quality Score
A high-quality diagnostic detects its intended failure mode, does not fire randomly outside scope, has severity proportional to real severity, adds new perception not already captured, changes MAVS decisions when it should, avoids damage when it should not act, is not redundant, and remains stable under irrelevant perturbations.

13. Diagnostic Set Theory

A full diagnostic system G is not just a bag of individually good diagnostics. A set of individually strong diagnostics can still be bad if the functions are redundant, contradictory, collectively too severe, or unstable when composed. Therefore Diagnostic Sciences evaluates diagnostic sets as systems.

$$Q_x(G) = J_x(M_G) - \lambda_1 C(G) - \lambda_2 R(G) - \lambda_3 N(G)$$

Term	Meaning
$J_x(M_G)$	Objective performance of MAVS using diagnostic set G.
$C(G)$	Complexity cost.
$R(G)$	Redundancy among diagnostics.
$N(G)$	Noise or instability introduced by G.

Redundancy can be estimated using mutual information or correlation:

$$R(G) = \sum_{i \neq j} I(z_i; z_j)$$

$$R(G) = \sum_{i \neq j} \mathrm{Corr}(z_i, z_j)$$

A good diagnostic set has high coverage, low redundancy, high objective gain, stable severity, and appropriate breadth. The formal set-selection problem is:

$$G^* = \arg\max_{G \subseteq D} Q_x(G)$$

Diagnostic Set Principle
The best G is not necessarily the largest G. The best G is the set that maximizes objective-aligned perception while controlling redundancy, instability, and governance damage.

14. Objective-Relative MAVS Optimization

A MAVS configuration can be represented as:

$$\eta = (G, A, W, \Theta, \tau_{hard}, \alpha, \lambda, \delta, \theta_0, \tau_G)$$

Parameter	Role
G	Chosen diagnostic set.
A	Severity aggregation rule.
W	Rebalancing rule.
Θ	Threshold policy.
τ_{hard}	Hard veto boundary.
α	Diagnostic weights.
λ	Severity strictness.
δ	Mitigation strength.
θ_0	Base threshold.
τ_G	Diagnostic firing thresholds.

The objective-relative optimal configuration is:

$$\eta_x^* = \arg\max_{\eta \in H} J_x(M_\eta)$$

Objective	Possible form	Expected configuration tendency
Accuracy	J_{acc} = Accuracy(Pi)	Lower severity sensitivity, fewer hard vetoes, more mitigation, more specificity.

Robustness	J_rob = E_{c~C}[Perf(M_eta,c)]	Stronger diagnostics, corruption-aware weights, higher lambda, more cautious threshold.
Safety	J_safety = 1 - UnsafeAcceptanceRate	High diagnostic breadth, strong hard veto, high lambda, low tolerance for correlated failure.
Balanced	beta_1 Accuracy - beta_2 UnsafeAcceptance - beta_3 FalseRejection + beta_4 Stability - beta_5 Complexity	Pareto-tuned governance under explicit tradeoffs.

No Universal Diagnostic Configuration
Usually G*_safety != G*_accuracy and G*_robustness != G*_recall. Diagnostic Sciences predicts this because each objective induces a different ordering over MAVS configurations.

15. Diagnostic Sciences Search Pipeline

The practical Diagnostic Sciences pipeline converts the theory into a research workflow. It identifies which diagnostics should exist, how they should be scoped, how they should be measured, and how they should be composed into MAVS configurations.

1. Define the objective x, such as robustness under high corruption or minimizing unsafe acceptance under correlated failure.
2. Define the failure-mode universe F = {F_1, F_2, ..., F_q}.
3. Generate candidate diagnostics D_cand = {g_1, ..., g_M}.
4. Assign each diagnostic an intended scope Omega_j.
5. Measure individual quality Q_x(g_j): sensitivity, specificity, calibration, perception gain, intended influence, unintended influence, redundancy, and instability.
6. Compose diagnostic sets G subset D_cand rather than assuming every candidate should be used.
7. Tune governance parameters A, W, Theta, tau_hard, lambda, delta, theta_0, and alpha.
8. Evaluate whole MAVS under clean, corrupted, adversarial, and correlated-failure regimes.
9. Select eta_x*, the best MAVS pattern for objective x.

Example failure-mode universe:

Failure mode	Description
F_1	Correlated specialist failure.
F_2	Input corruption.
F_3	Adversarial perturbation.
F_4	Distribution shift.
F_5	Specialist disagreement.
F_6	Overconfident consensus.

The output of the pipeline can be a family of parameterized MAVS configurations: MAVS_accuracy, MAVS_robustness, MAVS_safety, and MAVS_balanced. Each configuration is not a different architecture; it is a different optimized setting of the same governance system.

16. Theorem Program

Theorem 1 - Diagnostic Perception Extension

If I(Y_j; z_j | z_{-j}) > 0, then adding g_j strictly expands MAVS diagnostic perception with respect to failure mode Y_j. In effect, G^{+j} can distinguish cases that G^{-j} could not distinguish.

Theorem 2 - Monotone Governance Pressure

If A and Theta are monotone and g_j(phi) increases while other variables are fixed, then acceptance cannot become easier. Formally: z_j' >= z_j -> a' >= a -> theta' >= theta, so R - theta' <= R - theta.

Theorem 3 - Intended Influence Condition

A diagnostic g_j has nonzero architectural influence if P[Pi^{+g_j} != Pi^{-g_j}] > 0. It has intended influence if P[Pi^{+g_j} != Pi^{-g_j} | Y_j=1] > P[Pi^{+g_j} != Pi^{-g_j} | Y_j=0].

Theorem 4 - Redundant Diagnostic Theorem

If I(Y_j; z_j | z_{-j}) = 0 and z_j does not affect W or Theta beyond z_{-j}, then adding g_j does not improve objective performance except by noise or regularization artifact.

Theorem 5 - Safety-Recall Tradeoff

For d_j = 1[g_j(e) >= tau_j], decreasing tau_j weakly increases diagnostic breadth but may decrease specificity. Therefore maximum perception and maximum specificity are generally competing modes.

Theorem 6 - Objective-Relative Optimality

There does not exist a universally optimal diagnostic configuration G* unless all objectives induce the same ordering over MAVS configurations. Usually safety, accuracy, robustness, and recall require different optima.

These are theorem seeds rather than final formal proofs. They define the proof program for Diagnostic Sciences: prove when diagnostics expand perception, prove monotonic influence through governance, characterize redundancy, and formalize why objective-relative configurations differ.

17. Practical Implications for MAVS Research

Diagnostic Sciences turns MAVS-GC from a fixed governance configuration into a searchable family of governance configurations. This directly supports future MAVS research because it gives a method for designing and validating diagnostics rather than merely reporting benchmark outcomes.

The current MAVS-GC evidence path already supports the need for this addition. Chapter 10A prevents overclaiming by showing that MAVS-GC is competitive but not universally dominant under clean predictive correctness. Chapter 10B provides the strongest empirical signal by showing safer failure behavior under corruption and specialist failure. Chapter 10C adds the stability story by showing stronger stability preservation under corruption than under clean conditions.

Program	Interpretation for Diagnostic Sciences
Foundation Arc	Supplies the formal governance architecture that Diagnostic Sciences extends.
Chapter 9	Shows that diagnostics, severity, mitigation, hard veto, and traces can behave according to intended semantics.
Chapter 10A	Constrains claims: governance does not automatically maximize clean accuracy.

Chapter 10B	Motivates objective-relative diagnostics: safer failure under corruption is a diagnostic-governance outcome.
Chapter 10C	Motivates stability metrics as part of diagnostic quality and set composition.

The main experimental implication is that future MAVS validation should not only compare MAVS-GC against baselines. It should also run diagnostic ablations: remove one g_j, add one g_j, vary tau_j, measure IDI and UDI, compute perception gain, and evaluate objective-specific eta_x* under clean, corrupted, adversarial, and correlated-failure regimes.

The strongest immediate experimental design is therefore a diagnostic ablation matrix: rows are candidate diagnostics, columns are objectives and failure modes, and cells report sensitivity, specificity, calibration, MPG, IDI, UDI, redundancy, instability, unsafe acceptance, false rejection, accuracy, and stability. This would transform MAVS-GC validation from system-level benchmarking into diagnosis-level science.

18. Coverage and Traceability Audit

This section verifies that the Diagnostic Sciences draft has not been omitted. The main body reorganizes and explains the material, while Appendix A preserves the entire original diagnostic draft verbatim.

Original draft section	Where it appears in this document
Opening / right next layer	Executive Summary and Core Thesis.
MAVS existing equations	Section 2, MAVS-GC Substrate.
1. Diagnostics are MAVS sensors	Section 4.
2. Correlation failure case	Section 5.
3. Diagnostic Sciences base definition	Section 6.
4. Intended scope	Section 7.
5. Diagnostic quality as intended influence	Section 8.
6. Perception extension	Section 9.
7. Diagnostic influence pathway	Section 10.
8. Breadth and specificity	Section 11.
9. Intended scope quality score	Section 12.
10. Diagnostic sets G	Section 13.
11. MAVS optimization for objective x	Section 14.
12. Search pipeline	Section 15.
13. Theorem seeds	Section 16.
14. Clean definition	Executive Summary and Section 18.
15. Exact formulation sharpened	Core Thesis, Central Doctrine, and Section 18.

The MAVS-GC overview PDF is also integrated and preserved. Its system tuple, pipeline, definitions, properties, evidence path, Chapter 10B results, 10A/10C interpretation, limitations, repository links, and source artifacts are included in Section 2, Section 17, and Appendix B.

Final Clean Definition
Diagnostic Sciences is the formal study of diagnostic functions as perception-extending sensors inside governance architectures. In MAVS-GC, a diagnostic function g_j maps the shared representation or trace evidence into a nonnegative signal z_j, expanding the system ability to detect failure-relevant structure. A diagnostic is high-quality when it adds nonredundant perception within an intended scope and produces intended influence on governance decisions under a chosen objective, while minimizing unintended severity inflation, false rejection, instability, and redundancy.

Sharpened Formulation

g_j is high-quality iff it detects intended hidden structure and beneficially changes MAVS behavior for objective x.
G*_x = argmax_{G,A,W,Theta} J_x(M_{G,A,W,Theta}).
The best diagnostics are not chosen by vibes; they are selected by how much they improve the governed system for a defined purpose.

Appendix A. Original Diagnostic Sciences Draft - Preserved

The following appendix preserves the supplied Diagnostic Sciences draft in full so that no source content is lost during restructuring.

```
Yes. This is the right next layer.

What you are describing is **Diagnostic Sciences as the formal science of perception-extension inside
MAVS**.

MAVS already has:

[
G={g_1,\dots,g_k}
]

[
g_j(\phi)\rightarrow z_j
]

[
z=(z_1,\dots,z_k)
]

[
a=A(z)
]

[
\theta=\Theta(a,m)
]

[
\Pi=\mathbf{1}[a<\tau_{\text{hard}}]\cdot\mathbf{1}[R\ge\theta]
]

The MAVS documents already define this: diagnostics produce the nonnegative flag vector (z), severity
is aggregated by (A), and the threshold policy makes higher severity stricter while mitigation can
only boundedly relax it.

Now Diagnostic Sciences asks:

**What makes (G) good? What makes a single (g_i) good? How do we design, compare, compose, and
optimize diagnostic functions so MAVS becomes better at some target (x)?**

That is the exact formulation.

---

# 1. First principle: diagnostics are MAVS sensors

In MAVS:

[
F={f_i}
]

is the specialist/predictive layer.
```

$$
G={g_j}
$$

is the diagnostic/perceptual layer.

So:

$$
F \Rightarrow \text{answer evidence}
$$

$$
G \Rightarrow \text{condition evidence}
$$

The diagnostic layer does not merely predict the answer. It detects the **state of the input, model agreement, corruption, uncertainty, contradiction, instability, adversarialness, anomaly, or domain-risk**.

So the correct statement is:

$$
G \text{ expands MAVS' diagnostic perception over } \phi.
$$

Each diagnostic function is a sensor:

$$
g_j:\Phi(\mathcal{X})\rightarrow \mathbb{R}_{\ge 0}
$$

It transforms hidden or latent failure evidence into an observable governance signal:

$$
\phi \mapsto z_j
$$

The full diagnostic system is a sensor array:

$$
G(\phi)=(g_1(\phi),...,g_k(\phi))=z
$$

Then (A) compresses this into severity:

$$
a=A(z)
$$

Then (Θ) converts severity into governance pressure:

$$
\theta=\theta_0+\lambda a-\delta m
$$

So yes, formally:

$$
G\uparrow \Rightarrow z\text{-space expands} \Rightarrow \text{MAVS perception expands}
$$

but only if the added diagnostics are high-quality.

Bad sensors expand dimensionality, not perception.

Good sensors expand perception.

That distinction is the birth of Diagnostic Sciences.

---

2. Correlation failure case

You said:

> In correlation failure, MAVS shut everything down: 100% false rejection rate, 0 false positive acceptance rate.

That is actually a very important diagnostic phenomenon.

Suppose all specialists share the same flawed understanding:

$$
r_1,r_2,\dots,r_N
$$

are correlated in the wrong direction.

A normal ensemble may treat agreement as confidence:

$$
\text{agreement} \Rightarrow \text{accept}
$$

But MAVS can detect that the agreement itself is suspicious if diagnostics fire:

$$
z_{\text{corr}}=g_{\text{corr}}(\phi,r,F)
$$

Then:

$$
z_{\text{corr}}\uparrow \Rightarrow a\uparrow \Rightarrow \theta\uparrow
$$

possibly:

$$
a\ge\tau_{\text{hard}}
$$

then:

$$
\Pi=0
$$

So MAVS chooses shutdown.

That produces:

$$
\text{False Positive Acceptance}=0
$$

but:

$$
\text{False Rejection}=100\%
$$

This is not automatically “good” or “bad.” It depends on the objective.

For safety-critical governance:

$$
0\% \text{ unsafe acceptance}
$$

may be ideal.

For high-recall systems:

$$
100\% \text{ false rejection}
$$

may be unacceptable.

Therefore Diagnostic Sciences must evaluate diagnostics relative to an objective:

$$
x\in{\text{accuracy},\text{robustness},\text{safety},\text{recall},\text{precision},\text{stability},
\text{calibration},...}
$$

This is the central doctrine:

$$
\text{Diagnostic quality is objective-relative.}
$$

A diagnostic can be excellent for safety and terrible for recall.

---

3. Diagnostic Sciences: base definition

Let the diagnostic science layer be:

$$

```
\mathfrak{D}=(\mathcal{D},\mathcal{S},\mathcal{Q},\mathcal{O},\mathcal{B})
]

where:

[
\mathcal{D}
]

is the space of possible diagnostic functions,

[
\mathcal{S}
]

is the scope theory,

[
\mathcal{Q}
]

is the quality theory,

[
\mathcal{O}
]

is the optimization theory,

[
\mathcal{B}
]

is the breadth/specificity control theory.

A diagnostic function is:

[
g:\Phi(\mathcal{X})\rightarrow\mathbb{R}_{\ge 0}
]

or more generally:

[
g:\mathcal{E}\rightarrow\mathbb{R}_{\ge 0}
]

where (\mathcal{E}) is the evidence space available to the diagnostic.

For basic MAVS:

[
\mathcal{E}=\phi
]

For richer MAVS:

[
```

\mathcal{E}=(\phi,r,s,w,z_{\neg j},\text{metadata},\text{trace history})
]

So a diagnostic can inspect:

[
\phi
]

the shared representation,

[
r
]

specialist support,

[
s
]

raw specialist scores,

[
w
]

weights,

[
z_{\neg j}
]

other diagnostics,

[
T
]

trace/history.

Then:

[
g_j(e)=z_j
]

where:

[
z_j=0
]

means no detected issue,

[
z_j>0
]

means some detected diagnostic signal,

$$
z_j\gg 0
$$

means strong detected signal.

---

4. Intended scope of a diagnostic function

This is the important part.

A diagnostic function should not be judged in the abstract. It must have an **intended scope**.

Define the intended scope of g_j as:

$$
\Omega_j=(\mathcal{F}_j,\mathcal{C}_j,\mathcal{R}_j,\mathcal{I}_j)
$$

where:

$$
\mathcal{F}_j
$$

is the failure-mode family it is meant to detect,

$$
\mathcal{C}_j
$$

is the context/domain where it applies,

$$
\mathcal{R}_j
$$

is the response behavior it is supposed to trigger,

$$
\mathcal{I}_j
$$

is the intended influence on MAVS.

Example:

$$
g_{\text{corr}}
$$

Correlation failure diagnostic.

Its intended scope:

$$
\mathcal{F}_{\text{corr}}={\text{specialists agree due to shared flawed representation}}
$$

$$
\mathcal{C}_{\text{corr}}={\text{high agreement but low independent evidence/diversity}}
$$

$$
\mathcal{R}_{\text{corr}}={\text{raise severity sharply; possibly hard veto}}
$$

$$
\mathcal{I}_{\text{corr}}={\text{reduce unsafe acceptance under correlated failure}}
$$

Then quality is measured against that scope.

A diagnostic is bad if it fires outside its intended scope too often.

A diagnostic is weak if it fails inside its intended scope.

A diagnostic is dangerous if its influence does not match its evidence quality.

---

5. Diagnostic quality as intended influence

Your definition is strong:

> A diagnostic is high-quality according to how much intended influence it has on the overall MAVS architecture, meaning how much it extends MAVS' perception to detect more information/signals.

We can formalize that.

Let:

$$
\mathcal{M}
$$

be the MAVS system without diagnostic (g_j).

Let:

$$
\mathcal{M}^{+g_j}
$$

be MAVS with (g_j) added.

The diagnostic quality of (g_j) for objective (x) is:

$$
Q_x(g_j)=J_x(\mathcal{M}^{+g_j})-J_x(\mathcal{M})
$$

where (J_x) is the system objective.

For robustness:

$$
Q_{\text{rob}}(g_j)=J_{\text{rob}}(\mathcal{M}^{+g_j})-J_{\text{rob}}(\mathcal{M})
$$

For safety:

$$
Q_{\text{safety}}(g_j)=J_{\text{safety}}(\mathcal{M}^{+g_j})-J_{\text{safety}}(\mathcal{M})
$$

For accuracy:

$$
Q_{\text{acc}}(g_j)=J_{\text{acc}}(\mathcal{M}^{+g_j})-J_{\text{acc}}(\mathcal{M})
$$

This is **marginal diagnostic value**.

But we also need intended influence:

$$
I(g_j;\mathcal{M})=\mathbb{E}\left[|\Pi_{\mathcal{M}^{+g_j}}(x)-\Pi_{\mathcal{M}}(x)|\right]
$$

This measures how often adding (g_j) changes the final decision.

But changing decisions is not always good.

So quality is not just influence.

Quality is:

$$
\text{quality}=\text{intended influence}-\text{unintended damage}
$$

Formally:

$$
Q_x(g_j)=\Delta J_x(g_j)-\gamma U(g_j)
$$

where:

$$
\Delta J_x(g_j)=J_x(\mathcal{M}^{+g_j})-J_x(\mathcal{M})
$$

and:

$$
U(g_j)
$$

is unintended influence.

Examples of unintended influence:

$$
\text{false rejection inflation}
$$

$$
\text{accuracy collapse}
$$

$$
\text{noise sensitivity}
$$

$$
\text{redundancy}
$$

$$
\text{unnecessary severity inflation}
$$

$$
\text{instability under small perturbations}
$$

So the refined definition is:

$$
\boxed{
Q_x(g_j)=\text{objective gain caused by }g_j\text{ within its intended scope}
-\text{damage caused outside its intended scope}
}
$$

That is the core of Diagnostic Sciences.

---

6. Perception extension

Now define perception extension.

Let MAVS without g_j observe diagnostic state:

$$
z_{-j}
$$

With g_j:

$$
z=(z_{-j},z_j)
$$

The diagnostic perception extension of (g_j) is the amount of relevant signal it adds beyond existing diagnostics:

[
PE(g_j)=I(Y_{\Omega_j};z_j\mid z_{-j})
]

where:

[
Y_{\Omega_j}
]

is the latent truth of whether the intended failure mode is present.

This is conditional information gain.

Meaning:

[
PE(g_j)
]

is high if (g_j) tells MAVS something useful that other diagnostics did not already tell it.

If:

[
I(Y_{\Omega_j};z_j\mid z_{-j})\approx 0
]

then (g_j) is redundant.

If:

[
I(Y_{\Omega_j};z_j\mid z_{-j})>0
]

then (g_j) extends perception.

This is your “sensor” idea mathematically.

A sensor is high-quality if it detects relevant latent structure not already captured by the system.

---

7. Diagnostic influence pathway

A diagnostic influences MAVS through this chain:

[
g_j(\phi)=z_j
]

[
z_j \rightarrow A(z)=a
]

$$
a\rightarrow \Theta(a,m)=\theta
$$

$$
\theta\rightarrow \Pi
$$

Also sometimes:

$$
z_j\rightarrow W(i,\phi,z)=w_i
$$

$$
w_i\rightarrow R
$$

So there are two influence channels:

Severity channel

$$
z_j\rightarrow a\rightarrow \theta\rightarrow \Pi
$$

Rebalancing channel

$$
z_j\rightarrow w_i\rightarrow R\rightarrow \Pi
$$

Therefore diagnostic influence should be decomposed:

$$
I_{\text{total}}(g_j)=I_{\theta}(g_j)+I_R(g_j)
$$

where:

$$
I_{\theta}(g_j)=\mathbb{E}[|\theta^{+g_j}-\theta^{-g_j}|]
$$

and:

$$
I_R(g_j)=\mathbb{E}[|R^{+g_j}-R^{-g_j}|]
$$

Then final decision influence:

$$
I_{\Pi}(g_j)=\mathbb{P}[\Pi^{+g_j}(x)\ne \Pi^{-g_j}(x)]
$$

A diagnostic function is architecturally meaningful if:

$$
I_{\Pi}(g_j)>0
$$

or at least:

$$
I_{\theta}(g_j)>0
$$

or:

$$
I_R(g_j)>0
$$

But again: influence must be useful.

So:

$$
\text{High-quality diagnostic}=\text{perception extension}+\text{correct intended influence}
$$

---

8. Breadth and specificity

You need a way to increase/decrease breadth.

Define diagnostic breadth as the measure of cases where the diagnostic fires:

$$
B(g_j;\tau)=\mathbb{P}[g_j(e)\ge \tau]
$$

High breadth:

$$
B(g_j;\tau)\text{ large}
$$

The diagnostic fires across many cases.

Low breadth:

$$
B(g_j;\tau)\text{ small}
$$

The diagnostic fires only in narrow cases.

Specificity should measure how concentrated the diagnostic is on its intended failure mode:

$$
S(g_j)=\mathbb{P}[Y_{\Omega_j}=1\mid g_j(e)\ge\tau]
$$

Sensitivity/recall:

$$
R_c(g_j)=\mathbb{P}[g_j(e)\ge\tau\mid Y_{\Omega_j}=1]
$$

So:

$$
\text{breadth}=\mathbb{P}[\text{fires}]
$$

$$
\text{specificity}=\mathbb{P}[\text{true intended failure}\mid \text{fires}]
$$

$$
\text{sensitivity}=\mathbb{P}[\text{fires}\mid \text{true intended failure}]
$$

Now you can tune a diagnostic.

Let diagnostic have a parameter (ρ):

$$
g_j^{(\rho)}
$$

where (ρ) controls breadth.

For example:

$$
z_j=g_j^{(\rho)}(e)
$$

A low threshold gives broad perception:

$$
\tau_j \downarrow \Rightarrow B(g_j;\tau_j)\uparrow
$$

A high threshold gives specificity:

$$
\tau_j \uparrow \Rightarrow S(g_j;\tau_j)\uparrow
$$

So the formal tradeoff is:

$$
\tau_j\downarrow \Rightarrow \text{more perception, more false alarms}
$$

$$
\tau_j\uparrow \Rightarrow \text{less perception, higher precision}
$$

This gives three diagnostic modes:

Maximum perception mode

$$
\max B(g_j)
$$

Goal: detect as much as possible.

Use when missing danger is worse than false rejection.

Specificity mode

$$
\max S(g_j)
$$

Goal: only fire when the failure mode is very likely.

Use when false rejection is costly.

Balanced mode

$$
\max F_{\beta}(g_j)
$$

$$
F_{\beta}=(1+\beta^2)\frac{\text{specificity}\cdot \text{sensitivity}}{\beta^2\text{specificity}+\text{sensitivity}}
$$

Use when both false alarms and misses matter.

---

9. Intended scope quality score

For each diagnostic (g_j), define:

$$
Y_j(x)=1
$$

if the intended failure mode of (g_j) is present.

Then define:

Scope sensitivity

$$
\text{Sens}(g_j)=\mathbb{P}[g_j(e)\ge\tau_j\mid Y_j=1]
$$

Does it detect what it is supposed to detect?

Scope specificity

$$
\text{Spec}(g_j)=\mathbb{P}[g_j(e)<\tau_j\mid Y_j=0]
$$

Does it stay quiet when it should?

Severity calibration

If $q_j(x)$ is true failure severity:

$$
\text{Cal}(g_j)=1-\mathbb{E}[|g_j(e)-q_j(x)|]
$$

Does diagnostic magnitude match true severity?

Marginal perception gain

$$
\text{MPG}(g_j)=I(Y_j;z_j\mid z_{-j})
$$

Does it add new information?

Intended decision influence

$$
\text{IDI}(g_j)=\mathbb{P}[\Pi^{+g_j}\ne\Pi^{-g_j}\mid Y_j=1]
$$

Does it actually affect MAVS when it should?

Unintended decision influence

$$
\text{UDI}(g_j)=\mathbb{P}[\Pi^{+g_j}\ne\Pi^{-g_j}\mid Y_j=0]
$$

Does it wrongly affect MAVS outside scope?

Now the quality score:

$$
\boxed{
Q(g_j)=
\alpha_1\text{Sens}
+\alpha_2\text{Spec}
+\alpha_3\text{Cal}
+\alpha_4\text{MPG}
+\alpha_5\text{IDI}
-\alpha_6\text{UDI}
-\alpha_7\text{Redundancy}
-\alpha_8\text{Instability}
}
$$

This is extremely powerful because it does not define diagnostic quality vaguely.

It says a high-quality diagnostic:

1. Detects its intended failure mode.
2. Does not fire randomly outside scope.
3. Has severity proportional to real severity.
4. Adds new perception not already captured.
5. Changes MAVS decisions when it should.
6. Does not damage decisions when it should not.
7. Is not redundant.
8. Is stable under irrelevant perturbations.

That is a real Diagnostic Sciences quality formalism.

---

10. Diagnostic sets (G), not just individual (g_j)

A full (G) is not just many good diagnostics.

A set of individually good diagnostics can still be bad if they are redundant, contradictory, or collectively too severe.

So define quality of the full diagnostic system:

[
Q_x(G)=J_x(\mathcal{M}_G)-\lambda_1 C(G)-\lambda_2 R(G)-\lambda_3 N(G)
]

where:

[
J_x(\mathcal{M}_G)
]

is objective performance,

[
C(G)
]

is complexity cost,

[
R(G)
]

is redundancy,

[
N(G)
]

is noise/instability.

Redundancy can be:

[
R(G)=\sum_{i\ne j} I(z_i;z_j)
]

or more precisely:

[
R(G)=\sum_{i\ne j} \text{Corr}(z_i,z_j)
]

A good diagnostic set has:

[
\text{high coverage}
]

[
\text{low redundancy}
]

[
\text{high objective gain}
]

[
\text{stable severity}
]

[
\text{appropriate breadth}
]

So:

[
G^*=\arg\max_{G\subseteq\mathcal{D}}Q_x(G)
]

This is the formal way to find the best diagnostics for a target.

---

11. Optimization of MAVS for objective (x)

Now define a MAVS configuration:

[
\eta=(G,A,W,\Theta,\tau_{\text{hard}},\alpha,\lambda,\delta,\theta_0,\tau_G)
]

where:

[
G
]

is the chosen diagnostic set,

[
A
]

is severity aggregation,

[
W
]

is rebalancing,

[
\Theta
]

is threshold policy,

[
\tau_{\text{hard}}
]

is hard veto boundary,

[
\alpha
]

are diagnostic weights,

[
\lambda
]

severity strictness,

[
\delta
]

mitigation strength,

[
\theta_0
]

base threshold,

[
\tau_G
]

diagnostic firing thresholds.

Then:

[
\mathcal{M}_\eta

]

is MAVS under that configuration.

For objective (x):

[
\eta_x^*=\arg\max_{\eta\in\mathcal{H}}J_x(\mathcal{M}_{\eta})
]

where (\mathcal{H}) is the parameter/search space.

Examples:

Accuracy optimization

[
J_{\text{acc}}=\text{Accuracy}(\Pi)
]

[
\eta_{\text{acc}}^*=\arg\max_{\eta}\text{Accuracy}(\mathcal{M}_{\eta})
]

Likely result: lower severity sensitivity, fewer hard vetoes, more mitigation, more specificity.

Robustness optimization

[
J_{\text{rob}}=\mathbb{E}*{c\sim\mathcal{C}}\left[\text{Perf}(\mathcal{M}*{\eta},c)\right]
]

Likely result: stronger diagnostics, corruption-aware weights, higher (\lambda), more cautious threshold.

Safety optimization

[
J_{\text{safety}}=-\text{UnsafeAcceptanceRate}
]

or:

[
J_{\text{safety}}=1-\text{UAR}
]

Likely result: high diagnostic breadth, strong hard veto, high (\lambda), low tolerance for correlated failure.

Balanced objective

[
J_{\text{balanced}}
==========

\beta_1\text{Accuracy}
-\beta_2\text{UnsafeAcceptance}

```
-\beta_3\text{FalseRejection}
+\beta_4\text{Stability}
-\beta_5\text{Complexity}
]

This is probably the most useful real-world form.

---

# 12. Diagnostic Sciences search pipeline

The practical pipeline becomes:

## Phase 1 — Define objective

Choose:

[
x
]

Example:

[
x=\text{robustness under high corruption}
]

or:

[
x=\text{minimize unsafe acceptance under correlated failure}
]

## Phase 2 — Define failure-mode universe

[
\mathcal{F}={F_1,F_2,\dots,F_q}
]

Examples:

[
F_1=\text{correlated specialist failure}
]

[
F_2=\text{input corruption}
]

[
F_3=\text{adversarial perturbation}
]

[
F_4=\text{distribution shift}
]

[
```

$$
F_5=\text{specialist disagreement}
$$

$$
F_6=\text{overconfident consensus}
$$

Phase 3 — Generate candidate diagnostics

$$
\mathcal{D}_{cand}={g_1,\dots,g_M}
$$

Each diagnostic gets intended scope:

$$
\Omega_j
$$

Phase 4 — Measure individual quality

For each:

$$
Q_x(g_j)
$$

Measure sensitivity, specificity, calibration, perception gain, intended influence, unintended influence, redundancy, instability.

Phase 5 — Compose diagnostic sets

Search over:

$$
G\subseteq\mathcal{D}_{cand}
$$

Not all diagnostics should be used.

Some are redundant.

Some are too broad.

Some are useful only when paired with others.

Phase 6 — Tune governance parameters

For each (G), tune:

$$
A,W,\Theta,\tau_{\text{hard}},\lambda,\delta,\theta_0,\alpha
$$

Phase 7 — Evaluate whole MAVS

Compute:

$$
J_x(\mathcal{M}_{\eta})
$$

under clean, corrupted, adversarial, and correlated failure regimes.

Phase 8 — Select pattern

Return:

$$
\eta_x^*
$$

the best MAVS pattern for objective (x).

This gives you:

$$
\text{MAVS}_{\text{accuracy}}
$$

$$
\text{MAVS}_{\text{robustness}}
$$

$$
\text{MAVS}_{\text{safety}}
$$

$$
\text{MAVS}_{\text{balanced}}
$$

Each is a different parameterized configuration.

---

13. The key formal theorem seeds

These are the theorem directions.

Theorem 1 — Diagnostic perception extension

If:

$$
I(Y_j;z_j\mid z_{-j})>0
$$

then adding (g_j) strictly expands MAVS diagnostic perception with respect to failure mode (Y_j).

Meaning:

$$
G^{+j}
$$

can distinguish cases that:

$$
G^{-j}
$$

could not distinguish.

Theorem 2 — Monotone governance pressure

If (A) and (\Theta) are monotone and (g_j(\phi)) increases while other variables are fixed, then acceptance cannot become easier.

$$
z_j'\ge z_j \Rightarrow a'\ge a \Rightarrow \theta'\ge \theta
$$

Therefore:

$$
R-\theta'
\le
R-\theta
$$

This is the existing MAVS monotone safety property extended to individual diagnostics.

Theorem 3 — Intended influence condition

A diagnostic (g_j) has nonzero architectural influence if:

$$
\mathbb{P}[\Pi^{+g_j}\ne\Pi^{-g_j}]>0
$$

It has intended influence if:

$$
\mathbb{P}[\Pi^{+g_j}\ne\Pi^{-g_j}\mid Y_j=1]

>

\mathbb{P}[\Pi^{+g_j}\ne\Pi^{-g_j}\mid Y_j=0]
$$

Meaning it changes decisions more often inside its intended scope than outside it.

Theorem 4 — Redundant diagnostic theorem

If:

$$

$$
I(Y_j;z_j\mid z_{-j})=0
$$

and z_j does not affect W or Θ beyond z_{-j}, then adding g_j does not improve objective performance except by noise or regularization artifact.

In plain language:

A diagnostic that tells MAVS nothing new is not scientifically valuable.

Theorem 5 — Safety-recall tradeoff

For thresholded diagnostic:

$$
d_j=\mathbf{1}[g_j(e)\ge\tau_j]
$$

decreasing τ_j weakly increases diagnostic breadth:

$$
\tau_j\downarrow \Rightarrow B(g_j)\uparrow
$$

but may decrease specificity:

$$
\tau_j\downarrow \Rightarrow S(g_j)\downarrow
$$

Therefore maximum perception and maximum specificity are generally competing modes.

Theorem 6 — Objective-relative optimality

There does not exist a universally optimal diagnostic configuration G^* unless all objectives induce the same ordering over MAVS configurations.

Usually:

$$
G^*_{\text{safety}}\ne G^*_{\text{accuracy}}
$$

$$
G^*_{\text{robustness}}\ne G^*_{\text{recall}}
$$

This explains why correlated failure shutdown can be excellent for safety but bad for recall.

---

14. The clean “Diagnostic Sciences” definition

Here is the polished definition:

Diagnostic Sciences is the formal study of diagnostic functions as perception-extending sensors inside governance architectures. In MAVS-GC, a diagnostic function (g_j) maps the shared representation or trace evidence into a nonnegative signal (z_j), expanding the system's ability to detect failure-relevant structure. A diagnostic is high-quality when it adds nonredundant perception within an intended scope and produces intended influence on governance decisions under a chosen objective, while minimizing unintended severity inflation, false rejection, instability, and redundancy.

That is the concept.

---

15. Your exact formulation, sharpened

You can state the core doctrine like this:

$$
\boxed{
\text{Diagnostic quality is not signal magnitude. It is objective-aligned perception extension.}
}
$$

Or:

$$
\boxed{
g_j \text{ is high-quality iff it detects intended hidden structure and beneficially changes MAVS behavior for objective } x.
}
$$

Or:

$$
\boxed{
G^**x=\arg\max*{G,A,W,\Theta}J_x(\mathcal{M}_{G,A,W,\Theta})
}
$$

Meaning:

The best diagnostics are not chosen by vibes.

They are selected by how much they improve the whole governed system for a defined purpose.

This is exactly the layer MAVS needs before deeper theorems.

Appendix B. MAVS-GC Overview Source Text - Preserved

The following appendix preserves the extracted text from the supplied MAVS-GC governance overview PDF.

```
MAVS-GC: Governance-First AI for Failure-Mode Control
A concise research overview for external review and potential support
Prepared by Saif Malik - MAVS Research Program

Mathematical Formulation (MAVS-GC)
System tuple:
M = (X, Phi, F, G, A, W, P, Theta, Pi)

Core pipeline:
x -> Phi(x)=phi -> {f_i(phi)} -> r_i -> z -> a -> w_i,m -> theta -> R -> Pi

Definitions:
s_i = f_i(phi) in [0,1]
r_i = 2s_i - 1 in [-1,1]
z = (g_1(phi),...,g_k(phi)) in R_{>=0}^k
a = A(z), where A is monotone
w_i = W(i, phi, z)
m = max_l p_l(phi) in [0,1]
theta = Theta(a,m) = theta_0 + lambda a - delta m
R(x) = sum_i w_i r_i
Pi(R,theta,a) = 1[a < tau_hard] * 1[R >= theta]
Equivalently: Pi = 0 if a >= tau_hard; otherwise Pi = 1[R >= theta]

Auditable trace:
(r, w, z, a, m, theta, tau_hard, R, Pi)

Key properties:
All-speak evaluation: every specialist evaluates every input.
Monotone safety: higher diagnostic severity cannot make acceptance easier.
Bounded mitigation: mitigation cannot override hard veto because hard veto is inside Pi.
Governance separability: A or Theta can change acceptance without retraining specialists.

Executive Summary

MAVS, the Multi-Adaptive Vetting System, is a governance-first AI architecture. Its core claim is not
that a single model becomes
universally more accurate, but that explicit governance over specialist outputs can improve how
systems behave when evidence
becomes uncertain, contradictory, corrupted, or unstable.
MAVS-GC research overview - concise external review version

 The MAVS-GC formulation separates specialist prediction from output governance. All specialists
evaluate every input; diagnostics
produce red flags; severity is aggregated; contextual weights and bounded mitigation influence a
governed threshold; and the final
decision is made through an auditable consensus trace. This distinguishes MAVS from static ensembles
and routing-based
Mixture-of-Experts systems.
The current research program has completed a Foundation Arc, a synthetic validation chapter, and
three real-benchmark programs:
Chapter 10A for clean predictive correctness, Chapter 10B for robustness under corruption, and
Chapter 10C for reproducibility and
stability. The strongest empirical result is Chapter 10B: under specialist-failure and
high-corruption regimes, Pure MAVS-GC
substantially reduced unsafe acceptance while maintaining high accuracy.
```

Chapter

Question

1-8

Can MAVS be specified as a formal
governance architecture?

9

Do MAVS-GC mechanisms behave according
to intended formal semantics?

10A
10B
10C

Result

Does governance improve clean predictive
correctness?
Does governance fail more safely under
adverse conditions?
Does governance preserve reproducibility and
stability?

Foundation Arc completed: mission, thesis,
identity, formal objects, metrics, theorem
roadmap, cost model.
Synthetic validation completed; governance
altered decisions, preserved hard-veto
behavior, and produced complete traces.
Negative-to-mixed: competitive, but not
universally superior.
Supported and verified; strongest evidence
layer, especially under specialist failure.
Clean effects limited; corruption-condition
stability evidence supported.

1. What MAVS-GC Is

A MAVS system is represented as M = (X, Phi, F, G, A, W, P, Theta, Pi). X is the input space, Phi is a shared feature map, F is a set
of always-on specialists, G is a diagnostic system, A aggregates diagnostic flags into severity, W rebalances specialist influence, P
provides bounded mitigating evidence, Theta maps severity and mitigation into an acceptance threshold, and Pi produces the final
decision.
The key architectural distinction is regulated consensus. Specialists provide calibrated scores s_i in [0,1], converted into supports r_i
= 2s_i - 1. The system computes a governed consensus R = sum_i w_i r_i and accepts only when R meets the governed threshold
theta. The trace exposes r, w, z, a, m, theta, R, and the final decision.
This means MAVS-GC is better interpreted as a governance framework than as a fixed detector set. The current benchmarks

evaluate one concrete configuration, not the entire space of possible diagnostics, severity operators, mitigation strategies,
rebalancers, and threshold policies.

2. Evidence Path: From Formalism to Benchmarks

The Foundation Arc establishes the research identity: intelligence generation and intelligence governance are distinct concerns.
MAVS elevates governance into a first-class computational object rather than treating final acceptance as a static threshold after
model scoring.
Chapter 9 then isolates governance behavior using synthetic specialists instead of trained models. This design removed
model-training artifacts and tested whether diagnostics, severity, mitigation, hard veto, and trace generation behave as intended. In
the false-positive trap, mean aggregation accepted 100% of unsafe cases, static weighted aggregation accepted 85%, while
governance mechanisms materially reduced unsafe acceptance. Hard-veto compliance reached 100%, with no observed violations.
Chapter 10 moved to real benchmark datasets: Breast Cancer Wisconsin, Adult Income, Credit Card Fraud, and Bank Marketing.
The comparison systems were Single Model, Mean Ensemble, Static Weighted Ensemble, Veto MAVS, and Pure MAVS-GC. This
produced three subprograms: 10A for clean accuracy, 10B for robustness under corruption, and 10C for reproducibility and stability.

Program
10A - Accuracy

10B - Robustness
10C - Reproducibility

Main interpretation

MAVS-GC is competitive but does not establish universal
clean-condition predictive superiority. It changes the error profile: often
increasing precision or reducing false positives while reducing
recall/F1.
MAVS-GC demonstrates the clearest empirical advantage: it frequently
fails more safely under corruption, especially under specialist failure.
Clean-condition reproducibility improvements are limited, but stability
preservation becomes stronger as corruption increases.

MAVS-GC research overview - concise external review version

3. Chapter 10B: Highlighted Robustness Results

Chapter 10B is the main empirical signal to highlight. The hypothesis was that MAVS-GC fails more safely under adverse conditions.
The verified result supports this hypothesis across multiple datasets, corruption families, benchmark splits, governance traces,
robustness curves, and evaluation metrics. The strongest pattern appeared under specialist-failure corruption.

Figure 1. Under specialist failure, Pure MAVS-GC maintained 89.95% accuracy and 1.35% unsafe acceptance. Ensemble baselines
clustered near 74% accuracy and roughly 27% unsafe acceptance, while the single-model baseline had 59.46% accuracy and
45.42% unsafe acceptance.

Figure 2. Under high corruption (corruption >= 0.6), Pure MAVS-GC maintained 85.30% accuracy with 0.45% unsafe acceptance.
Mean/Veto/Static baselines were near 43.24% accuracy with 67.61% unsafe acceptance; the single model dropped to 23.22%
accuracy with 91.78% unsafe acceptance.

Regime	Pure MAVS-GC
Specialist failure	89.95% accuracy; 1.35% unsafe acceptance
High corruption >= 0.6	85.30% accuracy; 0.45% unsafe acceptance
Baselines	Mean/Veto ~74.31% accuracy; ~27.29% unsafe acceptance; Single Model 59.46% / 45.42% Mean/Veto/Static 43.24% accuracy; 67.61% unsafe acceptance; Single Model 23.22% / 91.78%
Relative signal	Unsafe acceptance ~20.2x lower than Mean/Veto and ~33.6x lower than Single Model. Unsafe acceptance ~149x lower than ensemble-like baselines and ~202x lower than Single Model.

The central scientific interpretation is not that MAVS-GC is universally more accurate. It is that governance changes failure behavior.
Under stress, MAVS-GC rejects more cautiously, suppresses unsafe acceptance, and preserves decision quality more effectively
than the evaluated aggregation baselines. The Chapter 10B report therefore identifies MAVS-GC as a failure-management,
robustness, and safety-oriented governance architecture rather than a pure accuracy-maximization architecture.

4. What 10A and 10C Add to the Interpretation

Chapter 10A is important because it prevents overclaiming. Under clean benchmark conditions, MAVS-GC did not demonstrate
universal predictive-correctness superiority. It improved accuracy over Veto MAVS in 2 of 8 comparisons, over Static Weighted

Ensemble in 0 of 8 comparisons, and produced positive metric deltas in 79 of 288 benchmark comparisons. This positions
MAVS-GC as competitive under normal conditions, not dominant.
Chapter 10C adds the stability story. Clean-condition reproducibility gains were limited, but corruption-condition stability effects were
stronger. Mean corrupted prediction stability was 0.971615 for Pure MAVS-GC versus 0.952713 for the comparison mean; decision
stability was 0.975770 versus 0.958762; consensus stability was 0.979332 versus 0.963946; and trace stability was 0.967976 versus
0.959693 for Veto MAVS. The chapter concludes that MAVS-GC preserves behavioral consistency under stress rather than
universally reducing variance.

5. Current Limitations and Support Needed

The current evaluation is rigorous but bounded. It uses four tabular datasets, a fixed suite of corruption families, controlled split and
audit structure, reproducibility manifests, and verified artifact trails. It does not establish production-scale behavior, LLM-agent
behavior, universal robustness superiority, or cross-domain generalization beyond the tested benchmark suite.
The most valuable next step is external-scale validation: larger datasets, additional modalities, LLM/agent specialist settings,
adversarial expansions, ablation matrices, and independent replication. Support from a research lab could help test whether the
observed failure-management and stability-preservation effects survive under larger-scale and more realistic AI systems.
Specific support requested: research feedback, compute credits, review of experimental design, guidance on scalable evaluation
settings, and potential collaboration on applying governance-first evaluation to LLM agents or safety-critical multi-model systems.

One-Sentence Summary

MAVS-GC is a governance-first architecture whose current evidence suggests competitive clean-condition performance, strong
failure-mode control under corruption, and stronger stability preservation under adverse conditions, with Chapter 10B providing the
clearest verified empirical signal.

Repository Links

Chapter 9 Synthetic Benchmark Program:
https://github.com/MAVS-RESEARCH/MAVS-Chapter-9
Chapter 10A Accuracy Benchmark Program:
https://github.com/MAVS-RESEARCH/MAVS-Chapter-10A
Chapter 10B Robustness Benchmark Program:
https://github.com/MAVS-RESEARCH/MAVS-Ch10B
Chapter 10C Reproducibility and Stability Program:
https://github.com/MAVS-RESEARCH/MAVS-Ch10C

Source Artifacts Used

-
-

MAVS Foundation Arc, Chapters 1-8.
MAVS formal definition and MAVS-GC calculus.
MAVS-GC research overview - concise external review version

-

-
-
-

Chapter 9 Synthetic Benchmark Program and completion report.
Chapter 10A Accuracy Benchmark completion report.
Chapter 10B Robustness completion report, robustness highlights, and corrected final numbers report.
Chapter 10C Reproducibility and Stability completion report.

MAVS-GC research overview - concise external review version