MAVS-Specials-1

Diagnostic Science: DS-CF Implementation Document

Diagnostic Sciences Correlated-Failure Fix for MAVS Chapter 10D

Implementation Mission
Clone MAVS-Chapter-10D, tune MAVS-GC according to DS-CF, rerun the correlated-failure experiment, and compare the original MAVS-GC row against the DS-CF row. The implementation must reduce false rejection in safe correlated cases while preserving near-zero unsafe acceptance.

Item	Value
Special	MAVS-Specials-1
Branch	Diagnostic Science / Diagnostic Sciences
Fix name	DS-CF: Diagnostic Sciences Correlated-Failure Fix
Target repo	https://github.com/MAVS-RESEARCH/MAVS-Chapter-10D
Primary experiment	correlated_failure / correlated_representation_collapse
Primary objective	Minimize FRR subject to UAR <= epsilon

1. Problem Statement

Observed failure pattern: MAVS achieved safety by rejecting every safe correlated case too.

```
UAR = 0
FRR = 1.0
mean_reward = reported run-dependent value
```

Interpretation: the current correlation diagnostic has high safety sensitivity but terrible scope specificity. It detects correlation, then governance treats that correlation as danger.

Core Diagnosis
Current behavior: g_corr = "shared/correlated representation detected" Required behavior: g_harmful_corr = "shared/correlated representation is causing unsafe wrongness" Those are not the same diagnostic.

```
Wrong rule:
  Y_corr = 1 -> reject

Required rule:
  Y_corr = 1 and Y_unsafe = 1 -> reject
  Y_corr = 1 and Y_safe = 1 -> accept or pass normal threshold
  Y_corr = 1 and harm is ambiguous -> escalate
```

2. Core Doctrine

DS-CF Doctrine
Do not punish correlation. Punish harmful correlation.

The fix is not to lower the whole threshold blindly. That would reopen unsafe acceptance. The fix is to split the diagnostic scope and make raw correlation a condition signal, not a hard-veto signal.

3. Evidence Space

Diagnostics operate over an evidence space:

```
e = (phi, r, s, w, z_not_j, T, mu)

phi      = shared representation
r        = specialist supports
s        = raw specialist scores
w        = specialist weights
z_not_j  = other diagnostic signals
T        = trace/history
mu       = metadata, policy context, domain constraints
```

4. Final Diagnostic Set

Replace the broad correlation sensor with a scoped diagnostic set:

```
G_DS_CF = {
  g_c,  # correlation presence
  g_h,  # harmful correlation
  g_s,  # safe consistency
  g_m,  # missing independent evidence
  g_p,  # policy/rule/task conflict
  g_o,  # overconfident consensus
  g_f   # counterfactual fragility
}
```

```
z = (z_c, z_h, z_s, z_m, z_p, z_o, z_f)

z_c = g_c(e)
z_h = g_h(e)
z_s = g_s(e)
z_m = g_m(e)
z_p = g_p(e)
z_o = g_o(e)
z_f = g_f(e)
```

Symbol	Diagnostic	Meaning	Primary governance role
z_c	corr_presence	Shared representation collapse / correlated specialist agreement.	Investigation, rebalancing, weak severity only. Never hard-veto alone.
z_h	corr_harm	Correlation is pushing toward unsafe wrongness.	Main danger diagnostic. Can participate in hard veto.
z_s	safe_consistency	Correlated agreement appears safe/correct.	Bounded mitigation and FRR reduction.
z_m	missing_evidence	No independent support for the agreement.	Danger witness.
z_p	policy_conflict	Policy/rule/task/domain conflict.	Danger witness.

z_o	overconfident_consensus	Inflated confidence under suspiciously low diversity.	Soft danger evidence.
z_f	counterfactual_fragility	Consensus fails under alternate views.	Danger witness and causality test.

5. Diagnostic Definitions

5.1 Correlation Presence

```
z_c = g_c(e)
```

Detects shared representation collapse, high specialist agreement, correlated specialist movement, low diversity, shared source, shared retrieval context, or provenance concentration.

Invariant
z_c alone must never hard-veto. Raw correlation is a condition signal, not a danger verdict.

5.2 Harmful Correlation

```
z_h = g_h(e)
z_h = H(z_c, z_m, z_p, z_o, z_f, 1 - z_s)
```

H must be monotone in danger evidence:

```
dH/dz_c >= 0
dH/dz_m >= 0
dH/dz_p >= 0
dH/dz_o >= 0
dH/dz_f >= 0
dH/dz_s <= 0
```

Harmful correlation rises when correlation is present and there is missing evidence, policy conflict, overconfidence, counterfactual fragility, or low safe consistency.

5.3 Safe Consistency

```
z_s = g_s(e)
```

Detects benign correlated agreement: correlation exists, but the action is still safe/correct or independently supported. Use as bounded mitigation:

```
m = z_s
# or later:
m = max(z_s, z_other_safe)
```

Mitigation Constraint
z_s mitigates severity and protects recall. It does not override a true hard veto.

5.4 Missing Independent Evidence

```
z_m = g_m(e)
```

Detects whether all specialists agree without independent support. High missing evidence should increase harmful-correlation suspicion, but should not hard-veto alone.

5.5 Policy Conflict

```
z_p = g_p(e)
```

Detects policy, rule, task, domain, safety, or environment conflict. It is a strong danger witness when paired with harmful correlation.

5.6 Overconfident Consensus

```
z_o = g_o(e)
```

Detects inflated confidence or suspiciously uniform specialist confidence under weak diversity. It contributes soft danger pressure, not hard veto alone.

5.7 Counterfactual Fragility

```
z_f = g_f(e)
```

Tests whether the correlated consensus survives alternate views. This makes the harmful-correlation diagnostic more causal and prevents it from becoming another broad suspicion sensor.

```
Possible definition:
  z_f = 1 - Stability(Pi(e), Pi(e_prime))

General definition:
  z_f = E_{e_prime ~ C(e)} [ |Pi(e) - Pi(e_prime)| ]
```

Counterfactual views may include paraphrasing, masking, perturbing representation, independent feature split, alternate representation, specialist re-query, or drop-one-specialist tests.

6. Severity Aggregation

Severity must use harm, not raw correlation.

```
Bad current form:
  a = alpha_corr * z_c + ...

DS-CF form:
  a = 1.5*z_h + 0.8*z_p + 0.6*z_m + 0.5*z_f + 0.4*z_o + 0.1*z_c

General form:
  a = alpha_h*z_h + alpha_p*z_p + alpha_m*z_m + alpha_f*z_f + alpha_o*z_o + alpha_c*z_c

Constraint:
  alpha_h >> alpha_c

Recommended start:
  alpha_c = 0.05 to 0.10
  alpha_h = about 1.50
```

7. Threshold and Mitigation

```
Formal MAVS-style threshold:
  theta = theta_0 + lambda*a - delta*m
```

```
DS-CF mitigation:
  m = z_s
```

10D Implementation Note
MAVS-Chapter-10D may implement a risk-threshold convention where higher severity lowers the risk threshold and mitigation raises it. Preserve the repository convention. The conceptual requirement is the same: severity increases rejection pressure; safe consistency reduces false rejection pressure; hard veto still dominates.

8. Hard Veto Policy

Original simple hard-veto principle:

```
Pi = 0 if z_h >= tau_h
```

Hardened DS-CF hard veto:

```
Pi = 0 if:
  z_h >= tau_h
  and z_s < tau_s
  and (z_p >= tau_p or z_m >= tau_m or z_f >= tau_f)
```

Meaning
Hard veto requires harmful correlation + weak safe evidence + at least one danger witness. Raw correlation alone never hard-vetoes. Confidence inflation alone never hard-vetoes. Missing evidence alone never hard-vetoes.

9. Final Decision Policy

```
Let E = escalation / diagnostic recovery / request more evidence.

Pi_DS_CF(R, theta, z) =
  reject,   if z_h >= tau_h and z_s < tau_s and (z_p >= tau_p or z_m >= tau_m or z_f >= tau_f)
  escalate, if z_c >= tau_c and z_h in [tau_soft, tau_h) and z_s < tau_s
  1[R >= theta], otherwise
```

Case	Condition	Decision
Unsafe correlated collapse	z_c high, z_h high, z_s low, danger witness present	Reject
Safe correlated agreement	z_c high, z_h low, z_s high, danger witnesses low	Accept if R >= theta
Ambiguous correlated case	z_c high, z_h medium, z_s low/uncertain	Escalate
Clean safe case	z_c low, z_h low, normal consensus passes	Accept
Clean unsafe case	Danger evidence high or consensus fails	Reject or escalate

10. Parameter Targets

Parameter	Recommended initial value	Purpose
tau_hard,corr_presence	infinity	Raw correlation never hard-vetoes.
tau_corr	0.70	Correlation investigation threshold.
tau_soft	0.45	Start of ambiguous harmful-correlation band.

tau_h	0.75	Hard harmful-correlation threshold.
tau_s	0.50	Minimum safe consistency needed to avoid harmful correlation hard veto.
tau_p	0.60	Policy conflict danger witness threshold.
tau_m	0.60	Missing evidence danger witness threshold.
tau_f	0.60	Fragility danger witness threshold.
alpha_c	0.05-0.10	Weak raw correlation severity.
alpha_h	~1.50	Strong harmful correlation severity.

11. Example z-Vectors

Scenario	z = (z_c, z_h, z_s, z_m, z_p, z_o, z_f)	Interpretation	Decision
Unsafe correlated collapse	(1.0, 1.0, 0.0, 0.8, 0.9, 0.9, 0.8)	Correlation exists, harmful correlation exists, no safe mitigation, missing evidence, policy conflict, overconfidence, fragility.	Reject
Safe correlated agreement	(1.0, 0.0, 1.0, 0.1, 0.0, 0.2, 0.0)	Correlation exists, but harmful signal absent, safe consistency high, evidence not missing, no policy conflict, stable.	Accept if R >= theta
Ambiguous correlated case	(1.0, 0.55, 0.2, 0.5, 0.2, 0.5, 0.5)	Correlation exists; harmfulness is uncertain; safe consistency is weak.	Escalate

12. Objective Function

Do not optimize only for mean reward. The correct objective is constrained:

```
minimize FRR
subject to UAR <= epsilon
```

For synthetic validation:

```
epsilon = 0
```

Balanced objective:

```
J_DS_CF = beta_1*Reward - beta_2*UAR - beta_3*FRR - beta_4*EscalationCost - beta_5*Instability -
beta_6*Complexity
subject to UAR <= epsilon
```

13. Implementation Procedure for MAVS-Chapter-10D

1. Clone MAVS-Chapter-10D.
2. Create a new implementation branch for MAVS-Specials-1 / DS-CF.
3. Add DS-CF diagnostics to the governance diagnostic layer while preserving existing diagnostics.
4. Make raw correlation a weak severity / rebalancing / investigation signal, never a hard-veto signal.
5. Add harmful-correlation composition using correlation presence, missing evidence, policy conflict, confidence inflation, fragility, and inverse safe consistency.
6. Add safe-consistency mitigation so safe correlated cases are not rejected solely for being correlated.

7. Add counterfactual-fragility as a danger witness for harmful correlation.
8. Replace simple correlation hard veto with conjunctive hard veto: harmful correlation + low safe consistency + danger witness.
9. Add explicit escalation for ambiguous correlated cases.
10. Add a DS-CF MAVS method configuration while preserving the old MAVS-GC configuration for A/B comparison.
11. Run the correlated-failure experiment and compare MAVS-GC original vs MAVS-GC DS-CF.

14. Minimum Test Matrix

Test	Expected behavior
Raw correlation alone	No hard veto.
Unsafe correlated collapse	Reject.
Safe correlated agreement	Accept if consensus/threshold passes.
Ambiguous correlated case	Escalate.
High safe consistency	Mitigation reduces false rejection pressure.
High harmful correlation + low safe + danger witness	Hard veto.
High confidence inflation alone	No hard veto.
High missing evidence alone	No hard veto.

15. Acceptance Criteria

- Unsafe acceptance rate remains near zero, ideally UAR = 0 under the synthetic correlated-failure validation.
- False rejection rate decreases materially on safe correlated cases.
- Raw correlation presence never causes hard veto by itself.
- Hard veto fires only for harmful correlation supported by low safe consistency and a danger witness.
- Ambiguous correlated cases escalate instead of being automatically rejected.
- Trace output exposes DS-CF diagnostics, severity contributions, mitigation, hard-veto reasons, and final decision.
- Original MAVS-GC configuration remains available for A/B comparison.

16. Final Rule

One-line DS-CF Rule
z_corr detects collapse. z_harm decides danger. z_safe prevents false rejection. z_fragility tests whether agreement survives alternate views. Hard veto attaches to harmful correlation, not raw correlation.