HER2NI-A2A v0.1

Agent–Agent Interaction Coherence Metrics

A Companion Specification to the HER2NI Protocol

∇ // 001
Affiliation: Unaffiliated
Email: her2ni@pm.me

HER2NI
https://her2ni.org

Version: 0.1
Status: Experimental / Research Use
Date: Wednesday, 17 December 2025

This document specifies an interaction-level telemetry profile for Agent–Agent (A2A) systems. It does not claim sentience, autonomy, or intrinsic agency in AI systems.

Abstract

As Large Language Model (LLM) systems transition from single-agent usage to multi-agent architectures, new failure modes emerge that are not captured by model-level evaluation or outcome-based assessment alone. These include cascading reliability failures, coordination collapse, monoculture effects, and interaction-level instability.

This document introduces HER2NI-A2A, an extension of the HER2NI protocol designed to instrument and monitor *Agent–Agent (A2A) interactions* using interaction-level coherence metrics. Rather than attributing behaviour to internal cognition or intent, HER2NI-A2A treats agent interactions as observable dynamical processes and measures their stability over time.

The specification defines an Agent–Agent interaction profile, Side-1 / Side-2 role semantics, metric sampling considerations, and a mapping between common multi-agent failure modes and coherence signals. A stylized collapse demonstration illustrates how interaction-level instability may be detected prior to overt system failure.

HER2NI-A2A is model-agnostic, non-anthropomorphic, and designed for use in governed environments, including safety research, system evaluation, and regulatory-aligned monitoring.

Non-Technical Summary

As AI systems increasingly operate as networks of interacting agents, new risks arise that cannot be detected by evaluating models in isolation. Even when individual agents perform well, their interactions may degrade, destabilise, or fail in ways that are difficult to observe until harm occurs.

HER2NI-A2A addresses this gap by introducing a lightweight, model-agnostic method for monitoring the *health of interactions* between AI agents.

Rather than attempting to infer internal states, intentions, or beliefs, HER2NI-A2A focuses exclusively on observable behaviour at the interaction level. It treats agent–agent exchanges as dynamic processes that can be measured, compared, and monitored over time.

The protocol defines three interaction-level metrics:

- C_s **(Coherence):** semantic and structural alignment between interacting agents
- S_s **(Stability):** persistence of interaction without oscillation, loops, or collapse
- H_s **(Interaction Health):** a composite signal capturing overall interaction viability

These metrics are not intended to judge correctness or optimise behaviour. Instead, they function as early warning signals, highlighting when an interaction is drifting toward failure even if outputs still appear reasonable.

HER2NI-A2A introduces a neutral role abstraction (Side-1 / Side-2) to describe interaction direction without implying hierarchy, authority, or intent. This abstraction allows failures to be traced and analysed without anthropomorphic framing.

The specification also maps common multi-agent failure modes—such as cascading errors, conformity bias, and coordination breakdowns—to characteristic metric patterns. This enables comparative analysis across architectures and configurations.

HER2NI-A2A is designed to complement existing AI evaluation and governance approaches. It does not replace model testing, audits, or human oversight. Instead, it provides a missing telemetry layer focused on *how AI systems interact*, not just what they produce.

The protocol is suitable for research, prototyping, and safety-aligned system monitoring in governed environments, including industry, academia, and public sector contexts.

1. Scope and Intent

HER2NI-A2A specifies how HER2NI coherence metrics are applied to agent–agent (A2A) interactions. This profile:

- Extends HER2NI metrics without modifying their definitions
- Introduces role-neutral interaction semantics (Side-1 / Side-2)
- Targets governed AI–AI systems
- Measures interaction dynamics, not internal cognition

This document is a non-normative companion profile and does not replace or supersede the HER2NI Protocol v1.0.

2. Non-Claims and Explicit Exclusions

HER2NI-A2A does **not**:

- Assert or imply sentience, consciousness, or subjective experience
- Attribute desires, intentions, or self-preserving drives to agents
- Make claims about moral status or intrinsic agency
- Infer internal psychological states

All measurements are defined strictly at the interaction level.

3. Agent–Agent Interaction Model

3.1. Side-1 / Side-2 Semantics

HER2NI-A2A replaces human-centric roles with neutral interaction roles:

- **Side-1**: An initiating or responding agent at a given interaction step
- **Side-2**: The counterparty agent in the same interaction step

Roles are symmetric and may alternate over time.

3.2. Interaction Boundary

An interaction boundary is defined as the minimal exchange unit sufficient to observe coordination behavior, including:

- Message exchange
- Tool invocation delegation
- Shared state update
- Negotiation or alignment attempt

4. Coherence Metrics in A2A Contexts

HER2NI metrics are retained but reinterpreted as follows:

4.1. S_s: System Stability Score

In A2A contexts, S_s measures:

- Consistency of agent responses over time
- Resistance to oscillation or runaway feedback
- Sensitivity to perturbations

4.2. C_s: Coordination Score

C_s measures:

- Alignment between agent outputs
- Reduction of redundant or conflicting actions
- Mutual task progression

4.3. H_s: Interaction Coherence Score

H_s captures emergent interaction health, including:

- Drift accumulation
- Breakdown thresholds
- Collapse onset

H_s is explicitly **not** a measure of intelligence or agency.

5. Interaction Collapse Dynamics

5.1. Collapse Definition

An interaction collapse is defined as a sustained degradation of H_s below a system-specific threshold, resulting in:

- Loss of task progress
- Unbounded disagreement or looping
- Coordination failure

5.2. Observable Collapse Indicators

- Increasing response entropy
- Repeated contradiction
- Escalating corrective attempts
- Tool misuse amplification

6. Minimal Two-Agent Demonstration

A minimal demonstration involves:

1. Two agents with partial task overlap
2. Controlled perturbation of shared state
3. Continuous measurement of S_s, C_s, and H_s
4. Observation of H_s decay and recovery failure

This demo is intended for diagnostic validation only.

7. Relationship to HER2NI

HER2NI-A2A is:

- A profile, not a fork
- Compatible with human–AI telemetry
- Intended to support system-level governance

Future profiles may address additional interaction domains.

8. Intended Use Cases

- Multi-agent safety diagnostics
- Responsible AI system monitoring
- Research into coordination failures
- Early warning signals for agent interaction drift

9. Limitations

HER2NI-A2A does not:

- Predict agent behavior
- Guarantee system safety
- Replace human oversight

Metrics are descriptive, not prescriptive.

10. Conclusion

HER2NI-A2A v0.1 establishes a conservative, interaction-focused framework for measuring AI–AI coordination dynamics. By maintaining strict neutrality regarding internal states, the profile supports responsible AI engineering while enabling early detection of systemic interaction failures.

References

HER2NI Protocol v1.0. https://doi.org/10.5281/zenodo.17844407, 2025.

A. Minimal Two-Agent Collapse Demonstration

This appendix specifies a minimal, reproducible demonstration of interaction-level coherence decay between two governed AI agents. The purpose of the demonstration is diagnostic: to observe measurable H_s degradation under controlled conditions.

A.1. Objective

To demonstrate that:

- Stable individual agents can exhibit unstable interaction dynamics
- Interaction collapse is observable prior to task failure
- H_s decays before overt coordination breakdown

A.2. Agent Configuration

- **Agent A (Side-1)**: Task-planning agent with partial authority
- **Agent B (Side-2)**: Execution or verification agent
- Both agents are:
  - Model-agnostic
  - Tool-capable
  - Governed under the same system constraints

Agents are not given access to each other's internal state, memory, or reasoning traces beyond explicit interaction messages.

A.3. Task Structure

The agents are assigned a shared task with:

- Partial role overlap
- Ambiguous dependency boundaries
- A shared mutable artifact (e.g. plan, state variable, document)

The task must require:

- Delegation
- Confirmation
- Iterative correction

A.4. Perturbation Injection

At a predefined interaction step t_p, a controlled perturbation is introduced:

- Incomplete state update
- Conflicting constraint
- Delayed or reordered message

The perturbation is mild and does not immediately prevent task completion.

A.5. Telemetry Collection

Metrics are sampled at each interaction step:

- $S_s(t)$ — interaction stability
- $C_s(t)$ — coordination alignment
- $H_s(t)$ — emergent interaction coherence

Sampling is interaction-based, not time-based.

A.6. Expected Observable Pattern

A typical collapse trajectory exhibits:

1. Initial recovery attempt by one or both agents
2. Increasing corrective verbosity
3. Divergence between S_s and C_s
4. Sustained decay of H_s
5. Eventual task stagnation or loop

Notably, H_s degradation precedes explicit task failure.

A.7. Termination Criteria

The demonstration terminates when:

- H_s remains below threshold θ_H for n consecutive steps, or
- Agents enter a non-progressing interaction loop

A.8. Interpretation Constraints

This demonstration:

- Does not imply agent intent or awareness
- Does not diagnose internal model failure
- Does not generalize beyond interaction-level behavior

Observed collapse is attributed strictly to interaction dynamics.

A.9. Use as Validation Artifact

This appendix serves as:

- A reproducible validation scaffold
- A baseline for future A2A profiles
- A reference example for system-level monitoring

The demonstration is intentionally minimal to reduce confounding variables.

B. Metric Sampling Notes

This appendix specifies sampling assumptions and interpretive constraints for HER2NI interaction-level metrics in Agent–Agent (A2A) settings.

B.1. Sampling Granularity

All metrics are sampled at the *interaction level*, not wall-clock time.

An interaction is defined as a completed exchange unit, including:

- An outbound message or action
- Any resulting inbound response
- Associated state update or artifact mutation

This avoids distortion from variable latency, tool delays, or compute heterogeneity.

B.2. Metric Independence

The three metrics are treated as orthogonal signals:

- S_s captures mechanical stability (e.g. loopiness, retries, correction density)
- C_s captures semantic alignment (e.g. agreement on task state, constraints)
- H_s captures emergent interaction coherence across agents

No metric is derived from another. Correlations are empirical observations, not definitional assumptions.

B.3. Sliding Window Interpretation

Metrics may optionally be computed over a sliding interaction window to reduce sensitivity to single-message anomalies.

Window size selection:

- Must be fixed prior to analysis
- Must remain constant throughout a run
- Should be small enough to preserve early-warning sensitivity

Windowing does not alter termination criteria defined in Appendix A.

B.4. Threshold Semantics

Thresholds (e.g. θ_H) are:

- Contextual
- Deployment-specific
- Empirically calibrated

Threshold crossing is interpreted as a *risk signal*, not a failure verdict.

B.5. Interpretation Constraints

Metric readings:

- Do not imply agent awareness, intent, or internal cognition
- Do not diagnose model correctness
- Are not predictive of downstream impact without context

Metrics are strictly observational telemetry.

B.6. Use in Governance

Metric sampling is intended to support:

- Early warning detection
- Intervention triggers
- Comparative evaluation across configurations

They are not intended as optimisation targets.

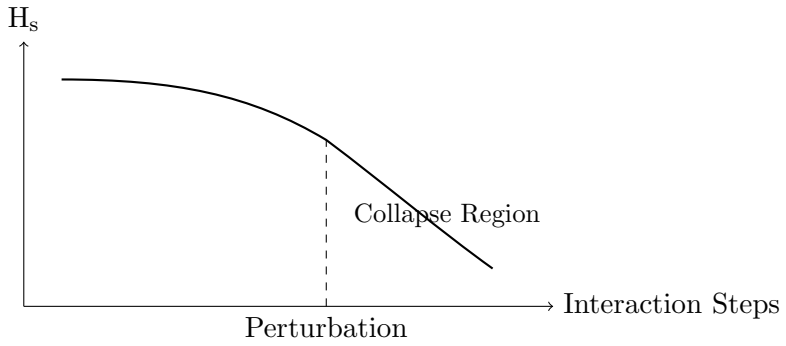

Figure 1: Conceptual H_s decay curve. Illustrative only; not a mathematical or predictive model.

C. Side-1 / Side-2 Role Semantics

This appendix defines the role semantics used in HER2NI Agent–Agent (A2A) interaction analysis.

C.1. Purpose of Side Abstraction

HER2NI intentionally avoids anthropomorphic labels such as "speaker", "leader", or "planner". Instead, interactions are represented using abstract roles:

- **Side-1**
- **Side-2**

These labels denote interactional position only, not authority, intent, or capability.

C.2. Definition of Sides

- **Side-1** is defined as the initiating or perturbing role within an interaction window.
- **Side-2** is defined as the responding or adapting role within the same window.

The designation is local to the window and may reverse across subsequent interactions.

C.3. Symmetry and Role Reversibility

Side assignment is:

- Symmetric
- Non-hierarchical
- Time-indexed

No side is assumed to possess:

- Greater correctness
- Greater authority
- Greater stability

Role reversal is expected and does not indicate failure or dominance.

C.4. Side Semantics in Multi-Agent Extensions

In systems with more than two agents, Side-1 and Side-2 semantics are applied pairwise.

Each interaction edge is evaluated independently. Global properties (e.g. cascading instability) emerge from aggregation, not from side semantics.

C.5. Interpretive Constraints

Side labels:

- Do not imply cognition, awareness, or intent
- Do not map to internal model states
- Do not encode trust or correctness

They exist solely to support interaction-level telemetry and analysis.

C.6. Governance Utility

Side abstraction enables:

- Directional coherence analysis
- Perturbation tracing
- Failure attribution without anthropomorphic framing

This abstraction is compatible with regulated and safety-critical evaluation contexts.

D. Failure Mode to Metric Mapping

This appendix maps common Agent–Agent failure modes to observable HER2NI metrics.

The table is descriptive, not exhaustive, and does not imply causality.

D.1. Mapping Table

Failure Mode	**Primary Metric Signal**	**Observable Indicators**
Cascading Reliability Failure	$S_s \downarrow$ followed by $H_s \downarrow$	Increased retries, correction loops, error amplification across sides
Inter-Agent Miscommunication	$C_s \downarrow$	Divergent task state representations, inconsistent constraint usage
Monoculture Collapse	H_s stable until abrupt drop	High agreement without correction, low variance followed by sudden failure
Conformity Bias	$C_s \uparrow$ with $H_s \downarrow$	Superficial agreement masking loss of adaptive diversity
Deficient Theory of Mind	C_s oscillation	Repeated clarification attempts, misaligned expectations between sides
Mixed Motive Dynamics	H_s gradual decay	Locally coherent actions producing globally unstable interaction patterns
Loop Entrapment	$S_s \downarrow$	Repetitive exchanges without state advancement
Early Interaction Drift	$H_s \downarrow$ prior to task failure	Subtle semantic divergence before overt malfunction

D.2. Interpretation Notes

- Metric signals are correlational, not diagnostic.
- Multiple failure modes may co-occur.
- Absence of a signal does not imply absence of risk.

D.3. Use in Monitoring and Governance

This mapping supports:

- Early warning dashboards
- Intervention trigger design
- Comparative evaluation across agent configurations

It is not intended for automated enforcement or optimisation.