[CAPABILITY PROFILE · AUTHREX-03]

ADARA · ADVERSARIAL DECEPTION-AWARE REASONING

Estimate the probability that the system's inputs are being manipulated and modulate authority downward before a deceptive input becomes an authorized action.

MODELED **TESTED**

Name variants: Deposit title: Adversarial Deception-Aware Risk Architecture (Zenodo record 19043924). Same framework.

PRIMARY FUNCTION	Deception prior P(adversarial) computed from input anomalies, temporal correlation, cross-sensor consistency, and Bayesian mission history; authority adjustment A_adj = A_hmaa x (1 - lambda x P_deception); phantom-fleet module against fabricated hostile scenarios.
MISSION AREAS	Counter-UAS engagement under spoofing; maritime and air domain awareness against decoys; command and control against injected false tracks; electronic-warfare resilience of perception.
EVIDENCE ANCHORS	Zenodo DOI 10.5281/zenodo.19043924; seeded browser simulation; formal-coverage record: ADARA_XSI_EXACT.tla FAITHFUL; ADARA.tla and ADARA_IMPL_V2.tla surrogate countermodels; ADARA_IMPL.tla withdrawn.
CURRENT STANDING	Demonstrated in simulation; one faithful formal model of the exact cross-sensor inconsistency function; a strong-form counterexample is published and drives a stated requirement; not independently reproduced.
SELF-ASSESSED TRL	TRL 3. No formal TRL determination has been performed.
INTELLECTUAL PROPERTY	U.S. provisional application 64/110,218, filed July 13, 2026 (receipt 78658472), reported as filed; unexamined; confers no enforceable rights.

[MISSION PROBLEM]

Adversaries manipulate what an autonomous system perceives: spoofed positions, decoy tracks, blinded cameras, fabricated hostile scenarios. A governance layer that trusts the perception layer inherits its errors, and a trust score that looks only for disagreement can be defeated by an attack that makes every channel agree. Honest confusion (a degraded sensor) and hostile manipulation call for different responses, and an architecture that cannot tell them apart either over-restricts on every fault or under-restricts under attack. The consequence in counter-UAS is an engagement recommendation built on a manufactured track.

[OPERATIONAL RELEVANCE]

- Counter-UAS engagement decisions where the track feeding the recommendation may be spoofed or decoyed.
- Maritime and air domain awareness against decoys and AIS or ADS-B manipulation.
- Command and control against injected false tracks in a common operating picture.
- Contested electromagnetic environments in which perception attacks are expected rather than exceptional.

[CAPABILITY DESCRIPTION]

ADARA runs in parallel with HMAA on the SATA output. It computes a proactive deception prior from four evidence streams and uses it to modulate the authority tier downward as the probability of manipulation rises. The Phantom Fleet detection module identifies AI-hallucinated hostile forces from correlated anomalies across sensor modalities so that a fabricated fleet does not drive an engagement. Deception is a first-class input to authority, not a detection alarm handled elsewhere. The formal record adds a requirement learned from a counterexample: deception detection must weight agreement, not only disagreement, because identical lies across all channels read as perfect consistency to a spread-based score.

[ADARA · ARCHITECTURE, ASSURANCE, EVIDENCE]

[TECHNICAL ARCHITECTURE]

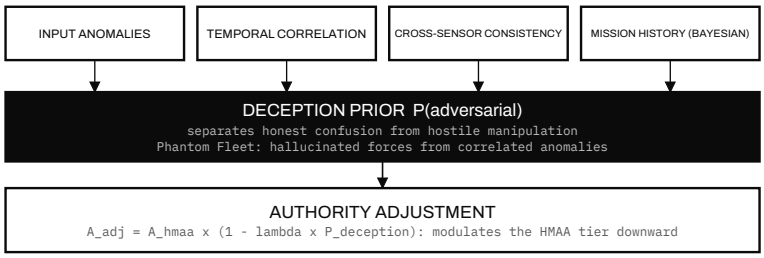

Figure 1. ADARA technical architecture. Encoding: solid = implemented; dashed = specified; gray = external; filled = enforcement point.

[HOW IT WORKS]

1. Receive the SATA per-source trust vector and the raw inconsistency features.
2. Compute input-anomaly, temporal-correlation, and cross-sensor-consistency scores; update the Bayesian mission-history prior.
3. Combine into P(adversarial), separating the honest-confusion hypothesis from the hostile-manipulation hypothesis.
4. Run the Phantom Fleet check (correlated anomalies across modalities) on any newly asserted hostile scenario.
5. Apply the authority adjustment: A_adj = A_hmaa x (1 - lambda x P_deception), which can only reduce the HMAA tier.
6. Emit the adjusted authority and the deception prior to the enforcement point and to FLAME (longer deliberation under suspected deception).
7. Record the prior, its components, and the adjustment to the ledger.

[TECHNICAL DIFFERENTIATION]

Anomaly detection and runtime monitoring flag deviations and leave the response to someone else; ADARA converts the flag into an authority change with a recorded outcome. Redundancy and voting cannot see an attack that makes every channel agree; ADARA's stated requirement is that agreement itself be weighed as evidence. The comparison is structural and is supported by the published counterexample; no claim is made about the performance of any named product.

[ASSURANCE MECHANISMS: DESIGN PROPERTY VERSUS ESTABLISHED PROPERTY]

Mechanism	Status
Adjustment can only reduce authority (monotone downward)	Design property; exercised in simulation
Exact cross-sensor inconsistency function	FAITHFUL formal model (ADARA_XSI_EXACT.tla imports and checks the implemented function)
Detection of identical-lie attacks	NOT established: published strong-form counterexample (all channels lie identically, conflict reads zero); requirement recorded
Deception prior recorded with components	Implemented in the seeded console
Near-minimum observation variant (ADARA_XSI_NEAR)	NOT FAITHFUL; result not relied upon

[TECHNICAL EVIDENCE]

Evidence	Result	Source / artifact
Seeded browser simulation	Deterministic replay from seed; synthetic deception scenarios	burakoktenli.com/adara-simulation
Faithful formal model of the inconsistency function	ADARA_XSI_EXACT.tla verdict FAITHFUL	authrex.systems/formal-coverage
Strong-form counterexample	Naive form held, strong form failed: identical readings read as agreement regardless of baselines	formal-coverage record
Withdrawn surrogate	ADARA_IMPL.tla withdrawn: discarded per-sensor baselines and the standard-deviation divisor	formal-coverage record
Zenodo deposit	Specification, source, simulation data	DOI 10.5281/zenodo.19043924

[EVIDENCE SNAPSHOT]

EVIDENCE SNAPSHOT	
Implemented	Yes (simulation engine)
Executable artifact	Yes
Formal model	Partial (one faithful module)
Simulation evidence	Yes
Hardware validation	No
Field validation	No
Independent validation	No
Government evaluation	No
Operational deployment	No

[LIMITATIONS]

- The adversary is scripted (spoofing, jamming, drift injection); adaptive adversaries that respond to the system are not modeled.
- Identical-lie attacks across all channels are a published open failure mode; the requirement is stated but the repair is not implemented in the shipped engine.
- The lambda weighting is a design parameter tuned on synthetic data; no calibration against recorded deception episodes exists.
- Simulation and formal-model evidence only; all data synthetic; no hardware-in-the-loop test, no field validation, no independent replication, no deployment.
- Not independently reproduced or technically verified by another party; second-environment reruns are confirmation by the same producer.

[TECHNOLOGY READINESS (SELF-ASSESSED; NO FORMAL TRA)]

CURRENT ASSESSED TRL	TRL 3 (self-assessed against the DoD nine-level scale; no formal technology readiness assessment has been performed).
BASIS	Critical function (deception prior modulating authority) is implemented in a seeded simulation and one component function has a faithful formal model; no laboratory validation against recorded or live deception has been performed.
EVIDENCE SUPPORTING	concept defined; partial formal model; simulation implementation
EVIDENCE REQUIRED FOR NEXT TRL	(1) Government-defined deception scenarios with ground truth (Phase 2) to measure false-restriction and miss rates. (2) Implementation and faithful-model check of the agreement-weighting repair. (3) Red-team evaluation against an adaptive adversary in a software-in-the-loop environment.

[INTEGRATION PROFILE]

INPUTS	SATA trust vector and inconsistency features; asserted hostile scenarios; mission history.
OUTPUTS	P(adversarial); adjusted authority; deception components; ledger entry.
INTERFACES	Parallel tap on the SATA output; modulation input to the HMAA tier and FLAME window.
DEPLOYMENT	Software module; browser console for evaluation.
SECURITY BOUNDARY	Inside: prior computation, phantom-fleet check. Outside: perception stack, track sources.
DATA REQUIREMENTS	Synthetic today; recorded spoofing or decoy episodes with ground truth would enable calibration.

[TRANSITION ROADMAP]

Phase	Activity for this capability	Who can execute it	Exit evidence
1 Technical review	Review of the artifacts listed in the evidence table	Program alone, with a government reviewer	Reviewer findings; correction log entries
2 Government-defined simulation	Sponsor scenarios on the seeded console	Program with sponsor scenarios	Deterministic, replayable results; hash-chained records
3 Degradation and failure injection	Evidence lapse, spoofing, delayed human response	Program with sponsor parameters	False-restriction, miss, and contraction-latency figures
4 Government testbed / SIL	Integration with a representative controller or interface	Requires sponsor interface and testbed	SIL integration report
5 Hardware-in-the-loop	Representative interference; timing on real hardware	Requires sponsor laboratory or rig	Measured latencies; TRL 4 to 5 evidence
6 Relevant-environment demonstration	Independent evaluation and red team	Requires independent evaluator	Independent report
7 Program integration	Only after earlier gates succeed	Requires transition partner	Not claimed to be available

[GOVERNMENT ENGAGEMENT REQUEST (SPECIFIC)]

- Government-defined deception scenarios with ground truth for Phase 2 simulation evaluation.
- Red-team assessment by an electronic-warfare or counter-UAS test organization.
- Guidance on acceptable false-restriction rates for the target mission, which determines lambda.

[REVIEWER QUESTIONS THIS PROFILE ANSWERS]

What is it: see capability description. Why does DoD care: see operational relevance. Where would it fit: see integration profile. Does it exist: see evidence snapshot. What evidence supports it: see technical evidence. Has anyone independent tested it: no. Has it been tested in the real world: no. What does it rely on: see limitations and security boundary. What can cause it to fail and what does it do then: see assurance mechanisms and the published counterexamples. How mature: see the TRL block. What would the government need to provide: see the request above. Next measurable milestone: first item under evidence required for next TRL.