Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Evaluating Neural Network Repairs in ACAS Xu: Static Fidelity and Off-Target Encounter Effects

Vimuth Fonseka

vimuth_fonseka@hotmail.com

Independent researcher

Ronnie Marlon Abeykoon

Independent researcher

Research Article

Keywords: neural network repair, ACAS Xu, aircraft collision avoidance, static fidelity, paired encounter evaluation, off-target effects

Posted Date: November 26th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11374296/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License

Additional Declarations: No competing interests reported.

Abstract

This study evaluates whether static fidelity measures favour the same neural-network repair as aircraft encounter outcomes. Twelve released repairs of a local output constraint (property 2) from four Airborne Collision Avoidance System (ACAS) Xu networks are tested in a planar research prototype. Four initial populations contain 278,528 draws, with baseline-defined inclusion and explicit numerically unresolved outcomes. For network N3,1, matched controller inference gives deep_opt_early_exit (deep) greater original-advisory agreement than early_exit (early): 99.3669% versus 99.3097%. In an 8,192-case fourteen-family mixture, early recovers 54 failures and introduces 30; deep recovers 14 and introduces 95. Replaying every case for the full 150-second horizon preserves all classifications and the 105-failure contrast, ruling out early stopping as its explanation in this cohort. The contrast falls to one after omitting two driving families. A frozen 28,781-case extension instead gives deep fourteen fewer known failures; central 95% resampling sensitivity intervals span zero. On its 1,693 unperturbed centres, the repairs tie with immediate response, whereas early has 21 and seven fewer failures under one- and four-second command delays. These are sensitivity settings, not calibrated aircraft responses. All extension centres have intruder/ownship speed ratio above two, and repaired-property-domain exposure remains zero. The released inputs, outcomes and programs support reproducible comparison of static fidelity with off-target encounter effects. The findings establish why repair assessments must specify encounter coverage, stopping rules and response assumptions; they provide neither a general repair ranking nor an operational flight-safety improvement.

1. Introduction

Repairing a neural component in an aircraft controller creates an engineering question: does a candidate preferred by a static measure also improve the controller's encounter outcomes? A network property constrains outputs on a specified input region. Agreement fidelity measures preservation of the original network's decisions on a test grid. Encounter evaluation follows the aircraft states resulting from a sequence of advisories and checks a separation criterion. These measurements concern different objects, and their relationship must be tested for a declared controller and encounter population. We assess unchanged released repair checkpoints using the version-pinned Specrepair evaluation and early Airborne Collision Avoidance System (ACAS) Xu controller sources [1, 2].

The original controller is a reference for fidelity, but it need not be successful on the encounter criterion. An advisory change can recover a failing encounter or introduce a failure in a previously passing encounter. Consequently, preserving more reference decisions and achieving fewer encounter failures are different objectives. Their relationship depends on where disagreements occur, the resulting dynamics, the encounter population and the measure used. A complete assessment should therefore retain recoveries and regressions, including cases that challenge a preferred configuration.

Aircraft-controller research already includes encounter-based evaluation. Julian et al. compare neural approximation with encounter performance [3], and Manzanas Lopez et al. examine switching-controller ACAS properties [4]. Bak and Tran distinguish network properties from closed-loop safety in an idealised ACAS Xu model [5]. Pang et al.'s MDPFuzzer author version evaluates ACAS repairs through renewed fuzzing and normal-case tests [6]. CoMPAQt studies aircraft trajectories after quantization and classification-preserving transformations [7]. These studies establish integrated evaluation; the present work assesses a specified panel of released property-repair artefacts.

Boetius et al. supply the checkpoints and static evaluation used here [8]. Their configurations instantiate the SpecRepair counterexample-guided framework [9], with different counterexample searchers. We evaluate the supplied artefacts without retraining them or introducing a repair algorithm. This separation makes the empirical question reproducible: what happens to a fixed candidate comparison after the artefacts are substituted into the same advisory-dependent controller?

Preservation and repair selection are established concerns. Ishimoto et al. predict repairs and breaks and perform ACAS-specific method selection [10]. BIRDNN's published study uses behaviour imitation with explicit repair/preservation objectives, repair-success and drawdown measurements, and layer and neuron-budget comparisons [11]. NCCDA and SPGR address preservation-aware repair and distribution or source/path choices [12, 13]. Counterfactual causality-based repair and Shapley-guided search provide further localisation and repair precedents [14, 15]. Our study does not claim priority for these mechanisms.

System-level repair also precedes this experiment. ContrRep repairs controller weights against temporal-logic system correctness [16], and ISAR addresses preservation during control-policy repair [17]. Stefani et al. generate aviation testing scenarios from an Operational Design Domain (ODD) model and evaluate ACAS-related configurations in FlightGear [18]. That framework advances scenario generation; this paper examines static measures and paired encounter outcomes of a released checkpoint panel.

We study three configurations for each of four ACAS Xu networks, replacing one component at a time in a public controller prototype. Static agreement is reconstructed on the source grid; encounter inclusion is defined by the untouched controller's exposure, so repaired paths cannot select their own evaluation population. The initial uniformly sampled source-law, unsafe-anchor and two challenge populations remain separate. A detailed N3,1 comparison selected after the initial outcomes is followed by a frozen broader-catalogue test of the same comparison.

The contribution is a reproducible evaluation of a fixed repair-checkpoint panel that connects static agreement, paired encounter recoveries and regressions, and sensitivity to encounter construction and response assumptions. Aligning network inference preserves the N3 agreement preference, while complete full-horizon replay rules out the original stopping convention as the explanation for the contrary encounter preference. The frozen extension and a subsequent, prospectively specified command-delay check then delimit that preference: it does not extend consistently across the tested conditions. The released inputs, outcomes and programs allow these comparisons to be repeated without retraining or substituting checkpoints. Sections 2 and 3 describe the methods and findings; Section 4 discusses their engineering interpretation and physical limits.

1.1 ACAS Xu prototype and encounter geometry

ACAS Xu is the unmanned-aircraft variant of the Airborne Collision Avoidance System X family. In the early horizontal research prototype considered here, neural networks approximate a lookup-table advisory policy [3]. Five continuous inputs describe the horizontal encounter: range ρ (ft), intruder bearing θ relative to ownship heading (rad), intruder heading ψ relative to ownship heading (rad), ownship speed v (ft/s), and intruder speed v_i (ft/s). Ownship is the aircraft receiving an advisory; the intruder is the other aircraft.

Each network produces five advisory scores. The prototype selects the minimum score, with first-index tie-breaking: clear of conflict (COC), weak left, weak right, strong left or strong right. Weak and strong commands impose ideal turn rates of ± 1.5 and ± 3 degrees per second, respectively; COC gives straight flight in this model. These commands update the planar constant-speed motion for one second before the next decision [2].

The previous advisory and τ, the time to loss of vertical separation in the network-selection model, choose a network rather than entering its five-dimensional continuous input vector. Five previous-advisory modes and nine τ slices yield 45 networks. The pinned selector uses τ values 0, 1, 5, 10, 20, 50, 60, 80 and 100 s [2]. Indices are one-based in the label N*i*,*j*: *i* identifies the previous advisory and *j* its τ slice. Thus N3,1 follows weak right at τ = 0. This co-altitude experiment retains τ = 0 throughout, exercises the five first-slice networks, and substitutes repairs only in N2,1–N5,1.

Figure 1 Encounter geometry and the evaluated ACAS Xu prototype. (a) Horizontal encounter geometry; (b) advisory-dependent prototype feedback. The five continuous inputs are range ρ, relative bearing θ, relative heading ψ, and speeds v/v_i. Previous advisory and τ select one of 45 networks; the minimum of five scores selects the next advisory. The co-altitude study uses the five τ=0 networks and ideal one-second motion updates. Geometry is schematic, not a recorded trajectory. The schematic was drawn with Python ReportLab from the pinned controller definitions [2, 3], using code drafted; no primary research image was generated or altered

2. Methods

2.1 Controller, checkpoint panel and study chronology

The panel contains optimal, early_exit and deep_opt_early_exit for N2,1, N3,1, N4,1 and N5,1. These are related configurations of four networks. The label optimal names a source search configuration; it does not assert global optimality. Deep identifies a configuration combining SpecAttack falsification with early-exit verification, rather than a deeper repaired layer [8]. The study keeps every checkpoint unchanged.

Specrepair is pinned to commit **261eacc661ba8128a3e2fc576c42bfd90a1ea33f** [1]; the aircraft simulator to **502eddbd7269f5f84c7d6d3387135980bf9997ad** [2]. Each candidate replaces one of the installed 45 networks, retaining the other 44. The co-altitude prototype exercises five τ = 0 advisory modes. The initial advisory is clear of conflict, ownship heading is π/2, and the intruder flies straight. Planar constant-speed dynamics use one-second commands: straight flight or turns of ± 1.5 or ± 3 degrees per second. No aircraft performance envelope or deployed implementation is evaluated.

The repaired property φ2 originates in the ACAS verification benchmark [19]; its exact pinned implementation controls this experiment. Within its bounded support, its domain includes radius $\rho \geq 55{,}947.691$ ft, ownship speed ≥ 1,145 ft/s and intruder speed ≤ 60 ft/s. It requires clear of conflict not to have the maximum output score, whereas advisory selection uses the minimum score. Domain membership and property satisfaction therefore do not require selection of clear of conflict. Source-reported repair status is held constant; no property proof is rerun.

A conversion control in an initial attempt found an advisory disagreement between a converted Open Neural Network Exchange (ONNX) repair and its released checkpoint. That attempt was stopped. An amendment fixed before the revised repaired outcomes switched all twelve repairs to direct one-dimensional float32 CPU PyTorch calls with normalize=False. The simulator's normalisation and first-index argmin are retained; unpatched components use the original runtime. Sealed original baselines are reused and checked, and stopped-attempt repaired outputs are excluded. The evaluated system is consequently a mixed-runtime prototype.

The panel and highest published agreement choices were fixed before the main repaired outcomes. Detailed N3 analysis, static reconstruction, family sensitivity, duration checks and inference alignment followed the main findings and are post hoc. The subsequent extension fixes the primary N3 comparison after those known results and before its new outputs. It is a prospective extension of a known question, without a public

preregistration or a blind new hypothesis. Supplementary Material (Online Resource 1), Section S1, records the estimands and version pins; Section S7 records the numerical controls and execution chronology, and Section S9 identifies the controlling programs.

The programs were executed through the Codex agent on a local computer; their outputs were checked using the controls and immutable records described in Sections 2.3–2.6 and Supplementary Material, Sections S7, S9 and S10. Python ReportLab renders Fig. 1 from pinned controller definitions and Figs. 2 and 3 directly from the recorded values in Tables 3–5. The accompanying figure reproduction program and its exact input tables specify this deterministic drawing procedure. AI-generated interpretations are distinguished from executed calculations; the human authors remain responsible for scientific verification and approval.

2.2 Separate encounter populations and baseline-defined inclusion

The initial design contains 278,528 draws, excluding calibration and feasibility cases. Four populations use independent recorded draw streams and are analysed separately (Table 1). There is no pooled safety rate. The anchor population selects uniformly among fourteen known unsafe README seeds, then applies support-conditioned perturbations. These are held-out perturbations of known families, not unseen encounter families. Rejected jitter preserves the selected family and is redrawn inside the fixed support.

Table 1
Initial encounter populations; their denominators remain separate.

Population	Draws	Frozen construction
Source law	262,144	Uniform radius 60,760–63,160 ft; intruder bearing and heading over a full circle; ownship speed 100–1,200 ft/s; intruder speed 0–1,200 ft/s
Anchor mixture	8,192	Uniform choice among fourteen known unsafe seeds; intruder x/y ± 200 ft, heading ± 0.005 rad and speeds ± 1%, within source support
Target-speed challenge	4,096	Source radius and angles; ownship speed 1,145–1,200 ft/s and intruder speed 0–60 ft/s
Uncontrolled close-approach challenge	4,096 accepted	Source proposals with straight-line closest approach ≤ 1,000 ft at a time in (0,150] s; proposal cap 2,097,152

For network n, E_n is the set of encounters invoking it at least once in the untouched baseline trajectory. Its three candidates share E_n even if repaired paths subsequently invoke different components. A nonexposed candidate retains the original outcome only after original-replacement and nonexposure controls pass. Such cases remain in whole-population tables. Target exposure T_n additionally requires an invocation inside φ2's domain; a zero eligible denominator makes the target-effect estimate undefined.

The source-law regression estimand is PASS→FAIL among baseline-PASS cases in E_n. The anchor recovery estimand is FAIL→PASS among baseline-FAIL cases in E_n. Whole-cohort failure changes use a common population denominator, retaining all numerical states. Conditional recovery and regression percentages are not subtracted because their denominators differ. Draw keys, case identifiers and input fingerprints allow exact reconstruction of each inclusion set.

2.3 Encounter classification and uncertainty

The initial simulation retains the source's 150-second maximum and early termination when sampled separation is increasing and exceeds 500 ft. Failure means a minimum separation below 500 ft within the retained held-command segments. This is the benchmark criterion, not a regulatory separation rule or operational accident probability. Straight-segment minima are analytic; turning-segment minima use root refinement. Each reported minimum is checked independently through the source matrix dynamics.

A separate threshold procedure uses the relative-distance Lipschitz constant $L = v_o + v_i$. For interval [*a*,*b*], its lower bound is $max(0, \frac{d(a) + d(b) - L(b-a)}{2})$, less a 10^{-7}-ft floating-point guard. A source-checked point below 499.99 ft is a failure witness. A pass requires lower bounds above 500.01 ft on all held-command intervals after subdivision. Depth 24 or 100,000 midpoint evaluations ends an unresolved check as UNKNOWN. The ± 0.01-ft band is numerical caution; the procedure is not an outward-rounded formal certificate.

All nine PASS/FAIL/UNKNOWN transitions are retained. Conservative failure-difference envelopes allow unresolved labels either binary assignment; these are numerical bounds rather than confidence intervals. Known-failure recovery excludes baseline UNKNOWN only from that conditional stratum. Keeping its full-population record prevents boundary cases from being silently assigned a favourable outcome.

Source-law regression bounds use one-sided exact Clopper–Pearson intervals with α = 0.05/12. The fixed research screen requires at least 1,094 eligible cases and an upper bound ≤ 0.5%; it is a precision screen, not an acceptable aviation failure rate. Anchor recovery uses two-sided exact intervals at confidence 1 − 0.05/12 and a separate eligible-count screen of 400. Each family has its own nominal simultaneous target under the intended independent Monte Carlo model; there is no joint operational confidence claim across the two families.

The original anchor draws independently choose among fixed centres and draw conditional jitter. Those intervals concern further eligible draws from that artificial mixture, not uncertainty over centre selection, operational prevalence or the post hoc choice of N3. Pointwise margin bootstraps

use 2,000 paired resamples, retaining shared case IDs. The separate post hoc family-composition sensitivity analysis resamples fourteen family totals with replacement 100,000 times, using PCG64 seed 20261003. Reported resampling endpoints are the 2.5th and 97.5th percentiles, forming central 95% sensitivity intervals. None of these calculations samples independent repair methods.

2.4 Static measures and matched N3 inference

The highest published agreement choices are deep for N2,1 and N3,1, and optimal for N4,1 and N5,1. Following the main outcomes, the complete 30^5=24,300,000-point source grid is reconstructed for each network. The original-property filter is preserved, and equal-point agreement is the exact agreeing count divided by the retained count. Every candidate within a network sees identical retained inputs. This removes batch-size weighting while retaining the source's input normalisation and output-denormalisation path. Published mean absolute output error (MAE), with lower values preferred, remains a secondary descriptive comparator; it is not remeasured as controller-path MAE.

The later N3 diagnostic reads 41,516 stored trace records for 10,379 baseline-exposed cases, yielding 827,477 invocations on 235,564 distinct exact input tuples. It compares source/controller normalisation, batch/scalar calls, normalised/denormalised outputs and authoritative original ONNX decisions. A separate 4,160-input grid control sample checks implementation paths. This recorded trajectory union is outcome-conditioned; it is not an independent validation population.

After a predeclared timing screen, every retained N3 grid point is evaluated through the actual encounter decision paths: single-input original ONNX and scalar repaired PyTorch inference with the simulator's normalisation, normalised outputs and first-index argmin. The raw grid and retained-input filter match the earlier reconstruction. The retained denominator is 24,299,779. Exact minimum ties are recorded, allowing conservative finite-grid order bounds without claiming arithmetic invariance on other backends or hardware.

2.5 Broader catalogue and robustness controls

The pinned public seed list contains 1,707 source-selected unsafe centres. The extension uses all 1,693 outside the original fourteen; those fourteen form a separate replication stratum. Each centre contributes its unchanged input and sixteen prekeyed, support-conditioned perturbations of the same widths used above. PCG64 pseudorandom-generator keys [20261004,104,seed,jitter_index] determine 28,781 extension and 238 replication cases before outcomes. New centres are disjoint from the old fourteen, but remain selected using unsafe-controller evidence.

Every new baseline and repair runs the full 150 seconds with early termination disabled. All baselines and exposure sets are sealed before repaired paths. Inference, classification and baseline inclusion remain fixed. The primary estimand is the equal-centre mean deep-minus-early known-failure difference across seventeen cases per centre. Replication is separate; the original source-law cohort is not an extension regression denominator.

All twelve candidates, centre-only/jitter-only subsets, centre counts and pretreatment geometry are retained. Geometry includes intruder/ownship speed-ratio bins, straight-motion closest approach clipped to [0,150] s, and ownship ≤ 200/intruder ≥ 600 ft/s. The two largest absolute centre contributions are jointly omitted using ascending-seed tie-breaking. A thirty-exposed-centre rule screens coverage rather than statistical significance.

Two thousand paired jitter resamples hold each centre case fixed; ten thousand centre resamples assess hypothetical composition sensitivity. Keys are [20261004,1041] and [20261004,1042]. For both analyses, the 2.5th and 97.5th percentiles define central 95% sensitivity intervals. These target different questions from numerical UNKNOWN and do not estimate fleet risk. The previous 8,192 draws and the new fixed-centre cohorts are never pooled.

Original stopping controls suppress early termination for the first 128 baseline-PASS/exposed source and first 128 baseline-FAIL/exposed anchor cases per network. A separate post hoc duration check runs all 661 baseline-PASS/N3-exposed anchor cases and 56 previously observed recovery cases through 300 seconds. The latter group is outcome-selected. These controls test specific stopping or duration concerns, rather than every case or infinite-horizon safety.

Every patched invocation is compared with a separately loaded identical checkpoint using the same backend. This checks wiring and inference convention, not independent arithmetic. Interrupted broader-study execution preserves its completed records and completes only missing records under a dated storage-monitor amendment. Population, horizon, inference, thresholds and estimands are unchanged. Per-record checks cover the full result; continuation diagnostic counters cover only the completed remainder. Supplementary Material, Section S7, retains the full execution history.

2.6 Complete stopping control and command-delay sensitivity

After the broader study, two focused controls are specified before their new simulations. First, all 8,192 original anchor inputs are regenerated and each is run with the baseline and N3 early/deep controllers for 150 seconds without early stopping. This includes cases that encounter N3 only after their former stopping time. The original outcomes are reconstructed from each full trajectory's first source-stopping prefix. Direct source replays on sixteen fixed cases and the original UNKNOWN case check prefix equivalence for all three controllers; the complete reconstructed baseline and paired totals are then compared with the recorded study. The comparison concerns whole-cohort counts; the original conditional exposure denominators remain attached to their original protocol.

Second, every unchanged extension centre (1,693) and every replication centre (fourteen) is evaluated separately at command delays of zero, one and four seconds, without jitter, always for 150 seconds. At integer second t, the current state and previous issued advisory select the network and new advisory. The aircraft applies the advisory issued d seconds earlier, retaining clear-of-conflict for the first d seconds. Issued and applied advisories are logged separately; delayed motion does not overwrite the controller's previous-issued-advisory memory. All other source dynamics, speeds, thresholds and inference conventions remain fixed. Each repair is paired with the baseline at the same delay on the complete fixed centre set.

The one- and four-second settings have precedent in aircraft detect-and-avoid evaluation [22], but the present command-transport delay is a sensitivity model, not a calibrated pilot, autopilot or airframe response. No delay compensation, roll model or operational encounter distribution is added. Zero-delay paths are checked against the existing runner for every centre and controller. Applied-command trajectories receive the same independent continuous-threshold check, and every issued repaired-network advisory is checked with a separately loaded checkpoint. All conditions and numerical UNKNOWN outcomes are retained. Supplementary Material, Section S10, supplies the complete results and implementation checks; Online Resource 2 supplies the new case records and programs.

3. Results

3.1 Original outcomes, coverage and source-law regression

The untouched controller's classifications differ substantially among the four initial populations (Table 2). N3 exposure includes 5,504 source cases and 2,637 anchors; the latter contain 661 PASS, 1,975 FAIL and one UNKNOWN. No target-domain invocation occurs in any baseline or candidate, including the target-speed challenge. These findings therefore concern off-target consequences of φ2 repair artefacts. The target-effect denominator is zero, rather than a measured zero benefit.

Table 2
Untouched-controller outcomes. The extension and replication are separate later cohorts.

Population	Cases	PASS	FAIL	UNKNOWN
Source law	262,144	262,141	3	0
Anchor mixture	8,192	2,230	5,961	1
Target-speed challenge	4,096	4,096	0	0
Uncontrolled close-approach challenge	4,096	4,092	4	0
Later extension	28,781	10,585	18,195	1
Separate old-centre replication	238	57	181	0

Every candidate has zero source-law PASS→FAIL regressions in its baseline-PASS/exposed stratum, with no repaired UNKNOWN there. The eligible counts per variant are 4,220 for N2,1, 5,502 for N3,1, 3,533 for N4,1 and 1,289 for N5,1. Their family-adjusted one-sided upper bounds are respectively 0.129789%, 0.099562%, 0.155007% and 0.424283%, meeting the fixed research screen. This provides no zero-risk conclusion. In the separate uncontrolled close-approach challenge, N2 early introduces one failure. All target-speed classifications remain passing despite their zero target exposure.

3.2 Complete panel and the original N3 comparison

Table 3 retains all twelve candidates and the contrary extension alongside their source-grid measures. Static agreement is equal-point reconstruction; MAE is the published secondary comparator. Recovery/regression pairs concern original FAIL→PASS/PASS→FAIL on the common cohort, while known-failure change includes unresolved transitions where present. These counts do not establish a statistically complete ranking of independent repair methods.

Table 3
All twelve released checkpoints. Rec/reg means recoveries/new failures; ΔF is the known-failure change against the original controller on the whole cohort. Static values are rounded for presentation; exact grid counts, stored MAE values and all nine-cell outcomes are in the supplementary tables and reproduction inventory.

Network	Repair	Agreement (%)	MAE	Anchor rec/reg	Anchor ΔF	Extension rec/reg	Extension ΔF
N2,1	Optimal	99.5734	0.098801	137/130	−7	531/352	−180
N2,1	Early	99.3634	0.151783	160/137	−23	812/443	−369
N2,1	Deep	99.5960	0.108055	154/136	−18	586/361	−225
N3,1	Optimal	98.9968	0.216561	39/107	+ 68	212/420	+ 208
N3,1	Early	99.3097	0.139732	54/30	−24	395/307	−88
N3,1	Deep	99.3669	0.171256	14/95	+ 81	391/289	−102
N4,1	Optimal	99.8319	0.103449	12/2	−10	89/43	−46
N4,1	Early	99.8105	0.071018	8/2	−6	100/34	−66
N4,1	Deep	99.8067	0.098856	7/2	−5	99/31	−68
N5,1	Optimal	99.6545	0.124276	1/0	−1	44/4	−40
N5,1	Early	99.5343	0.125779	1/0	−1	73/23	−50
N5,1	Deep	99.6278	0.089642	1/0	−1	52/12	−40

Note. Optimal, Early and Deep denote **optimal**, **early_exit** and **deep_opt_early_exit**, respectively. MAE is mean absolute output error. Both ΔF columns use their own whole-cohort denominator; negative values mean fewer known failures. For N2,1 Optimal in the extension, 352 new failures minus 531 recoveries gives − 179; one additional FAIL→UNKNOWN transition reduces the known-FAIL count by one, giving the reported − 180. Complete nine-cell transitions are in Supplementary Material, Section S6.

The main N3 source-path reconstruction gives deep 24,145,944 agreements and early 24,132,034 out of 24,299,779 retained inputs, a 13,910-point advantage. In the original anchors, early instead recovers 54/1,975 baseline-failing exposed cases and introduces 30 failures among 661 baseline-passing exposed cases. Deep recovers 14 and introduces 95. Recovery-family exact intervals are [1.796885%,3.957880%] for early and [0.285679%,1.438131%] for deep. These address the fixed-mixture recovery estimand, not a universal repair preference. Supplementary Material, Section S2, gives all static measures and anchor totals; Section S3 gives complete transitions for all four initial populations, and Section S4 gives the N3 paired outcomes and fourteen family contributions.

Direct early/deep pairing gives 106 early-PASS/deep-FAIL cases and one contrary case, with 2,148 passing under both, 5,936 failing under both and one shared UNKNOWN. Deep has 105 more known failures, with the deliberately conservative numerical envelope [104,106]. The paired encounter assessment therefore differs from agreement fidelity in this fixed cohort (Fig. 2). Published N3 MAE instead favours early and agrees with the original encounter ordering; this is not a failure of static metrics as a class.

Other networks do not repeat a universal N3 pattern. Original N4 optimal has the most recoveries with the same two regressions as its alternatives. All original N5 recovery/regression counts tie. For N2, published MAE favours optimal while the known-failure change favours early. These descriptive contrasts explain why the detailed N3 analysis is labelled post hoc and why the complete panel remains visible.

3.3 Inference alignment, margins and original-family sensitivity

The completed controller-path grid check preserves N3's agreement order (Table 4). Deep has 13,907 more agreements than early, or 0.057230973 percentage points. Conservative assignment of original/repair exact minimum ties gives the positive count envelope [13,903,13,911]. The historical source-path and matched controller-path values are distinct measurements; neither is silently substituted for the other.

Table 4
Complete retained N3 grid through the encounter controller's actual scalar inference paths.

N3 repair	Grid agreements	Retained points	Agreement (%)	Original ties	Repair ties
Optimal	24,055,999	24,299,779	98.996781000%	1	2
Early	24,132,035	24,299,779	99.309689195%	1	1
Deep	24,145,942	24,299,779	99.366920168%	1	1

Note. Configuration labels follow Table 3. Ties are exact minimum-score ties under the recorded decision paths.

All 827,477 stored N3 advisory occurrences are reproduced. Four nonzero batch/scalar path comparisons concern two distinct stored inputs, including exact or near ties; all complete score/input records are retained. The 4,160 static control inputs have matching diagnostic advisories. Full enumeration, rather than that sample alone, supplies the matched agreement order. Independent aggregation reproduces the diagnostic summaries and retained-grid records.

Zero source-law regressions coexist with substantial decision and margin changes. Among the 5,502 baseline-PASS/N3-exposed source cases, at least one advisory changes in 59.51% for optimal, 46.91% for early and 44.46% for deep. Mean repaired-minus-original minimum-separation changes are + 432.27, + 175.14 and + 125.21 ft; their most negative changes are − 8,105.82, − 7,925.84 and − 9,784.80 ft. Positive averages coexist with negative tails, while the classified outcomes remain above the benchmark boundary. Complete margin descriptors, exact published MAE values and stopping controls are in Supplementary Material, Section S5.

Early retains fewer original-anchor failures than deep after each single-family omission, but two families dominate the contrast. All 30 early and all 95 deep new failures belong to seed 1084018, and 46 of early's 54 recoveries belong to 1379646. Omitting both leaves eight early and seven deep recoveries with no regressions: the 105-failure difference shrinks to one. Removing recovery family 1379646 alone changes early's known-failure difference against baseline from − 24 to + 22. A single-omission result therefore does not establish robustness to all compositions.

The post hoc central 95% family-composition sensitivity interval [− 1,262] also spans zero. Its mass at or below zero is 0.04949, which is a resampling sensitivity descriptor rather than a calibrated p-value. The original fixed-mixture binomial intervals and this hypothetical composition analysis answer different questions. Their separation prevents strong Monte Carlo precision conditional on fourteen centres from being read as knowledge about unseen families.

All twenty-four original stopping-control rows retain classifications and minima within the 0.01-ft tolerance. The 717-case selected duration study produces 2,868 original/N3 trajectories, retaining the observed regression and recovery counts at 150 and 300 seconds. Its latest minimum is at 102.161224 s. This addresses the selected stopping/duration concern, without creating a full-anchor 300-second ranking or an infinite-horizon conclusion.

3.4 The broader catalogue gives a small opposite difference

All 1,693 extension centres reproduce as original FAIL. Their 27,088 jitter cases create the later cohort totals in Table 2. N3 exposure covers 18,168 cases at 1,491 centres, with 6,953 PASS and 11,215 FAIL. The single baseline UNKNOWN is N3-nonexposed and retains its original record. Each candidate still has zero target-domain exposure.

Early N3 recovers 395 failures and introduces 307; deep recovers 391 and introduces 289. Their whole-cohort known-failure counts are 18,107 and 18,093. The fourteen-failure difference favouring deep is − 0.048643 percentage points, with numerical envelope [− 15,−13]. This reverses the original direction. Deep has eighteen fewer regressions but four fewer recoveries, so the lower total does not mean that it dominates both objectives. Table 5 compares the three cohorts, and Fig. 3 displays their distinct failure differences and numerical envelopes.

Table 5
Primary N3 comparison across the separate cohorts. The last column applies the respective two-driver omission rule.

Cohort	Cases	Centres	Early rec/reg	Deep rec/reg	Deep−early known FAIL	Numerical envelope	After two-driver omission
Original anchor mixture	8,192	14	54/30	14/95	+ 105	[104,106]	+ 1
Later extension	28,781	1,693	395/307	391/289	−14	[− 15,−13]	−12
Separate replication	238	14	0/1	0/2	+ 1	[1, 1]	0

Centre-only extension recoveries tie at fifty each. The entire − 14 difference comes from jitter cases, which have 345/307 early and 341/289 deep recovery/regression pairs. Across centres, 160 favour early in total, 161 favour deep and 1,372 tie. Of 339 centres with some classification disagreement, eighteen have cancelling totals. Thus centre count and classification-disagreement count describe different aspects of the evidence.

The two largest absolute extension contributions are − 8 and + 6. Omitting both leaves twelve fewer deep failures across 28,747 cases, envelope [− 13,−11]. Persistence of this small opposite sign does not validate the original early preference. The separate 238-case replication has no N3 recovery, one early regression and two deep regressions; its one-case difference disappears under the prescribed omission. It is a new small centre/jitter sample, not the original 8,192 draws.

The full-panel results in Table 3 also retain the extension's additional UNKNOWN for N2 optimal: its known-failure change of − 180 includes a FAIL→UNKNOWN transition, in addition to 531 recoveries and 352 regressions. This distinction is why all nine-cell tables accompany summary counts. Secondary candidates are not promoted to replace the frozen N3 comparison after viewing the results.

3.5 Geometry, sampling sensitivity and implementation checks

Every extension centre has intruder/ownship speed ratio greater than two. Of these, 1,636 satisfy the slow-ownship/fast-intruder descriptor; all nonzero N3 centre-total differences lie there. The two lower speed-ratio bins are empty (Table 6). More public seeds therefore expand the catalogue without establishing broad physical coverage.

Table 6
Centre-defined pretreatment geometry. The speed, closest-approach and slow/fast descriptors overlap and are not added as independent datasets.

Centre stratum	Centres	Cases	Deep−early FAIL	Numerical envelope
Intruder/ownship speed ratio < 0.5	0	0	0	[0,0]
Intruder/ownship speed ratio 0.5–2	0	0	0	[0,0]
Intruder/ownship speed ratio > 2	1,693	28,781	−14	[− 15,−13]
Straight-motion closest-approach distance < 500 ft	185	3,145	+ 8	[8, 8]
Straight-motion closest-approach distance ≥ 500 ft	1,508	25,636	−22	[− 23,−21]
Slow-ownship/fast-intruder condition: no	57	969	0	[0,0]
Slow-ownship/fast-intruder condition: yes	1,636	27,812	−14	[− 15,−13]

Note. The slow/fast condition is ownship speed ≤ 200 ft/s and intruder speed ≥ 600 ft/s. Straight-motion closest approach is evaluated over 0–150 s before controller intervention.

Straight-motion closest approach below 500 ft favours early by eight cases, whereas the ≥ 500-ft stratum favours deep by twenty-two. These are pretreatment descriptors, not controlled-path minima or a complete causal explanation. The central 95% known-count paired jitter resampling endpoints are [− 0.201522,+0.111271] percentage points, and the hypothetical centre-composition endpoints [− 0.298808,+0.211945]. Both span zero, separately from the strictly negative finite-catalogue numerical envelope.

The initial revised run retains 278,528 baselines, executes 111,771 exposed repair replays and 10,752 controls, and preserves the failed conversion attempt. Source-matrix and threshold checks cover 401,054 records, including three known-source controls, with no classification disagreement. The broader run executes 217,635 exposed repair replays and 392 controls. All 5,265,103 saved patched invocations agree with the separately loaded same-backend reference. Independent checks reproduce 377,247 case/system rows, ninety-six panel rows and 20,484 centre/candidate records. These counts measure implementation checks, not independent physical validation.

The resumed broader run retains every completed group and finishes the missing records under the unchanged protocol. No partial or outcome-selected replacement cohort enters the result. Supplementary Material, Section S7, records the interruption, continuation checks and execution details.

3.6 Stopping-rule control and response sensitivity

The complete stopping control executes 24,576 full trajectories and reproduces the historical outcomes. All 8,192 cases continue beyond their former stopping time. Baseline N3 exposure rises from 2,637 to 3,355 cases, but every baseline and repaired classification and continuous minimum remains unchanged. Early retains 54 recoveries/30 regressions and deep 14/95: the + 105 deep-minus-early contrast, [104,106] numerical envelope, two-family-omission result of + 1 and central 95% family-composition interval [− 1,262] all persist. Early stopping therefore does not explain the original N3 contrast or its difference from the full-horizon extension within these tested inputs.

The delay control executes 15,363 full trajectories across the 1,707 fixed centres and three controllers. On the extension centres, immediate-response recoveries tie at fifty each. At one second, early recovers 36 failures and introduces eighteen, whereas deep recovers twenty-two and introduces twenty-five; early has 21 fewer known failures. At four seconds the respective recovery/regression pairs are 10/6 and 4/7, giving seven fewer early failures (Table 7). No delayed-centre result is numerically unresolved. The small replication set remains separate.

Table 7
Command-delay sensitivity on unchanged centres, with a common 150-second horizon.

Centre set (cases)	Delay (s)	Baseline FAIL	Early rec/reg	Deep rec/reg	Deep−early FAIL
Extension (1,693)	0	1693	50/0	50/0	0
Extension (1,693)	1	1319	36/18	22/25	+ 21
Extension (1,693)	4	644	10/6	4/7	+ 7
Replication (14)	0	14	0/0	0/0	0
Replication (14)	1	12	1/0	0/0	+ 1
Replication (14)	4	7	0/0	0/0	0

Note. Each row compares repairs with the baseline at that row's delay. Positive differences mean fewer early failures. UNKNOWN is zero throughout; no jitter cases enter this table. The three delay conditions reuse the same centres and are not independent samples.

The baseline itself changes from 1,693 extension failures at zero delay to 1,319 and 644 at one and four seconds. These centres were selected as unsafe under immediate response, so this decrease cannot establish that adding delay improves operational safety. The controlled finding is that an unchanged static fidelity comparison corresponds to different paired encounter outcomes when the response assumption changes. The full-horizon and delay controls have zero target-domain exposure.

4. Discussion

4.1 What the empirical comparison establishes

The initial N3 result demonstrates a particular disagreement between original-advisory agreement and encounter outcomes. A highest-agreement screen chooses deep, while the fixed original anchors favour early on both recoveries and regressions. Matching controller inference preserves the static order, so a loss of that N3 grid order under the recorded scalar convention does not explain away the comparison. The screening rule is an illustration for these artefacts, not a claim that aircraft developers rely on fidelity alone.

The broader catalogue changes the inference a reader should draw. Deep has slightly fewer known failures in its centre-plus-jitter mixture, while centre-only recoveries tie, centre signs nearly balance and sampling-sensitivity intervals span zero. The command-delay control further shows that the centre-only tie becomes an early preference at the two tested nonzero delays, although the static scores are unchanged. The evidence supports reporting the cohort, response assumption and objective for each candidate assessment. It does not support recommending early or deep generally.

The reference network is fallible under the benchmark criterion. Preserving a decision on a failing encounter and altering it on a passing encounter can have different consequences, even if both changes affect the same fidelity total. Static agreement and separation outcomes therefore remain useful but distinct measurements. Neither the small panel nor its observed ordering validates a fidelity threshold or a monotonic safety relationship. Published N3 MAE's agreement with the original encounter ordering further prevents a general claim that static metrics fail.

4.2 Population, protocol and controller interaction

The original family concentration is material. Its large 105-failure contrast reduces to one when the two main drivers are removed, and a family-composition resampling interval spans zero. The extension expands the number of centres while remaining concentrated in high speed-ratio, mostly slow-ownship/fast-intruder geometry. A strong difference in a fixed artificial mixture is thus compatible with weak knowledge about other encounter compositions. Numerical precision does not resolve this coverage problem.

The old and new cohorts differ in construction and centre membership. The complete stopping control now places the original N3 comparison under the extension's 150-second convention and preserves every outcome and minimum. Thus the stopping-rule difference does not account for the measured sign change in these cohorts. This does not identify one geometric feature as its cause, nor does it establish robustness to other response models. The separate delay experiment changes only command transport while retaining the centre inputs, and demonstrates response dependence under that declared model.

Stored representatives show how changed commands propagate through the deterministic controller. In one recorded original anchor, a deep advisory change increases minimum separation from 472.231537 to 501.763188 ft; in another, it decreases separation from 890.195872 to 471.682705 ft. Both replace an initial strong-right command with weak right after an identical prefix. These contrary consequences connect advisory strength with evolving geometry, without identifying switching as an isolated causal mediator. The source labels early and deep do not denote a general repair-layer safety effect.

Baseline-defined inclusion is important to the interpretation. Defining exposure after repair could let a candidate's altered path change its comparison population. Here the original path fixes inclusion and the nine-cell transitions retain boundary uncertainty. Even so, conditional recovery and regression remain different estimands. Total failure counts express the declared cohort composition; they are not automatically a utility function appropriate for operations with unknown exposure weights.

4.3 Relationship to prior work and the increment's scope

Approximation/encounter comparisons, closed-loop ACAS testing, controller repair and preservation are already established [3, 4, 5, 6, 7, 8, 9, 10, 16, 17]. Boetius et al. provide the released repair configurations and original static objectives [8], while SpecRepair supplies their underlying procedure [9]. This study's increment is an empirical assessment of that unchanged artefact panel under specified integrated-controller and inference conventions. The complete broader result is part of that increment, including its failure to sustain the original directional conclusion. A closed-loop ACAS Xu benchmark also studies original-controller behaviour with Dubins and F-16 models [20]. These precedents establish integrated evaluation without supplying this paired repair-checkpoint comparison.

BIRDNN's published article makes repair/preservation trade-offs explicit and reports pointwise property improvement and drawdown under retraining and layer-wise or cross-layer fine-tuning [11]. Its repair experiments do not report aircraft encounter trajectories. Counterfactual causality-based repair (CCBR) uses causal attribution and neuron localisation, balancing property violations against accuracy loss penalised by semantic distance; its ACAS results are pointwise violation and positive-property-accuracy rates [14]. NCCDA, SPGR and Shapley-guided NSRepair likewise preclude treating localisation, path guidance or preservation as a new mechanism here [12, 13, 15]. Those pointwise metrics are not presumed equivalent to our source-grid original-advisory agreement or trajectory separation classification. Supplementary Material, Section S8, records the complete published BIRDNN and CCBR versions used.

Stefani et al.'s ODD-based FlightGear study demonstrates an alternative route to aviation scenario construction [18]. Its model-based framework and aircraft simulator are different tested objects from these released checkpoint comparisons. Likewise, an automotive temporal-logic controller-repair experiment [16] is a relevant precedent without being the same benchmark. The present benchmark complements these approaches by making the outcomes and inference controls for a fixed repair panel available for reuse.

The benchmark adds a controlled distinction between two possible explanations of repair preference. The complete replay removes a stopping-rule explanation for the original contrast; the delay check demonstrates dependence on the timing of applied commands even when static fidelity is unchanged. These controls strengthen the empirical connection between network measures and a stated aircraft-controller model. Researchers can reuse the released inputs and outcomes to test candidate-selection measures without treating the original controller as an infallible reference. The contribution remains this documented comparison, rather than a new repair algorithm or general repair-selection rule.

4.4 Physical validity and assurance limits

The prototype is planar and constant speed, assumes a straight intruder and ideal held turn rates, and exercises five $\tau = 0$ modes. The added one- and four-second command delays probe one response assumption; they do not validate sensing uncertainty, pilot behaviour, roll dynamics, vertical manoeuvres or aircraft performance feasibility. The demanding source speeds and fixed turn rates further limit the physical interpretation. A longer replay or a fixed command queue cannot establish an aircraft response envelope or deployed-system performance.

Zero $\varphi 2$-domain exposure is also a substantive limit. The experiment cannot measure the repaired property's encounter benefit, because the required invocations never occur. Off-target changes can still be observed and compared, but do not validate or refute the source's formal-property result. Zero exposure in this catalogue is not a universal reachability proof: relative speed magnitude alone does not impose a lower bound on radial closing rate, and controller switching depends on trajectory geometry.

Numerical controls strengthen implementation confidence within their scope. Same-backend comparisons check checkpoint wiring; a separate geometric procedure checks threshold decisions; complete N3 enumeration retains the agreement order despite documented rare path discrepancies. None supplies independent floating-point verification or guarantees on another platform. The original conversion failure remains relevant and is preserved, rather than hidden by a successful later configuration.

The 500-ft criterion is a research benchmark classification. It is not a regulatory finding, certification result or accident probability. Source-law bounds, finite UNKNOWN envelopes and resampling sensitivity also concern different defined objects. Their combination cannot be described as an operational confidence level. The study therefore supports comparison of evidence about this controller and these encounters, with a clear separation between local-property status, measured separation outcomes and authority to deploy an aircraft system.

4.5 Implications for evaluating repaired aircraft controllers

For aircraft-controller evaluation, the practical outcome is a sequence of checks on the evidence used to compare repairs. First align static inference with the controller's actual decision path. Then report recoveries and regressions against a common baseline and retain unresolved cases. Finally hold inputs fixed while checking the stopping horizon and response convention, before extending the encounter construction. In this panel, inference alignment and full-horizon replay preserve the original contrast, while encounter construction and command delay change the

comparison. A favourable net count is therefore interpretable only with its component transitions and the assumptions under which it was measured.

The broader test also shows why encounter coverage must accompany numerical precision. The fourteen-family difference depends strongly on two drivers, and the later catalogue gives a small opposite direction even after the stopping convention is equalised. Its centre-only response checks remain conditioned on unsafe, high speed-ratio inputs. The released case records and programs make these specific challenges available for reuse; evaluation for a particular aircraft would still require a justified operating envelope and a validated response model.

5. Conclusion

This version-pinned assessment compares twelve released ACAS Xu repair checkpoints through static measures and paired encounter outcomes. Matched N3 inference retains greater agreement for deep, while the original fourteen-family mixture favours early; published MAE also favours early. Replaying all original cases for 150 seconds preserves every classification and the 105-failure contrast, which falls to one after removing two drivers. The later extension gives fourteen fewer deep failures in its centre-plus-jitter mixture. On its unchanged centres, the zero-delay tie becomes 21 and seven fewer early failures with one- and four-second command delays. These findings separate a ruled-out stopping explanation from observed encounter- and response-dependent preferences. The released benchmark supports reproducible off-target repair assessment, with explicit numerical uncertainty and coverage. It supplies neither a general repair ranking nor an operational flight-safety improvement.

Declarations

Funding

This research received no specific grant from funding agencies in the public, commercial or not-for-profit sectors.

Competing interests

The authors declare no competing interests.

Data and code availability

The original software and released checkpoints are publicly identified in References 1, 2 and 8. The original and broader-study research data and standalone Python programs are publicly available in Zenodo, version 1.0.0 [21], at https://doi.org/10.5281/zenodo.23202447. That package includes all 29,019 broader-study input vectors, 377,247 case/system outcome rows, 20,484 centre/candidate records and resampling draws. Supplementary Material, Sections S1–S10, is supplied as Online Resource 1. Online Resource 2 contains the additional static, inference, duration and family checks, and the new stopping-rule and delay outcomes, summaries and executable programs. The new controls are supplied with this article and are not part of the earlier Zenodo version. File captions and instructions appear in Section S9 and the supporting README. Full trajectory databases can be regenerated with the documented programs.

AI assistance

During the preparation of this work, OpenAI Codex (recorded model identifier gpt-6.1-sol, OpenAI) assisted with spelling, grammar correction, language editing of the manuscript, document preparation and setting up the coding environment. The human authors retain full responsibility for verification, scientific interpretation, review and editing of the content, and approval of the version submitted.

Ethics approval and consent

Not applicable. This computational study involves no human participants, personal participant data or animals; consent to participate and consent for publication are not applicable.

Author Contribution

V.F. contributed to methodology and validation, focusing on the aeronautical interpretation, encounter assumptions and aircraft-controller limitations.R.M.A. contributed to conceptualisation, methodology, software, formal analysis, data curation, validation and visualisation, focusing on the AI and neural-network evaluation. R.M.A. was responsible for preparing the programs, analyses and original manuscript draft, with AI assistance as disclosed in the manuscript.Both authors reviewed and edited the manuscript.

References

1. Boetius, D., Leue, S., Sutter, T.: Specrepair software and released repair checkpoints [software] Version pinned to commit 261eacc661ba8128a3e2fc576c42bfd90a1ea33f. (2024). https://github.com/sen-uni-kn/specrepair/tree/261eacc661ba8128a3e2fc576c42bfd90a1ea33f. Accessed 4 October 2026; experiment identity is the recorded immutable commit
2. Bak, S., Early ACAS Xu controller prototype [software]: quantized_nn_backreach, simulation/acasxu_dubins.py. Version pinned to commit 502eddbd7269f5f84c7d6d3387135980bf9997ad. (2022). https://github.com/stanleybak/quantized_nn_backreach/blob/502eddbd7269f5f84c7d6d3387135980bf9997ad/simulation/acasxu_dubins.py. Accessed 4 October 2026; experiment identity is the recorded immutable commit
3. Julian, K.D., Kochenderfer, M.J., Owen, M.P.: Deep Neural Network Compression for Aircraft Collision Avoidance Systems, Journal of Guidance, Control, and Dynamics 42(3) 598–608. (2019). https://doi.org/10.2514/1.G003724. Version-specific companion: *Deep Neural Network Compression for Aircraft Collision Avoidance*, arXiv 1810.04240v1 (2018), https://doi.org/10.48550/arXiv.1810.04240
4. Manzanas Lopez, D., Johnson, T.T., Bak, S., Tran, H.-D., Hobbs, K.L.: Evaluation of Neural Network Verification Methods for Air-to-Air Collision Avoidance, Journal of Air Transportation 31(1) 1–17. https://doi.org/10.2514/1.D0255. Version-specific companion: (2023). https://www.verivital.com/research/lopez2022jat.pdf. Accessed 5 October 2026
5. Bak, S., Tran, H.-D., Neural Network Compression of ACAS Xu Early Prototype Is Unsafe: Closed-Loop Verification Through Quantized State Backreachability, NASA, Formal Methods: 280–298. (2022). https://doi.org/10.1007/978-3-031-06773-0_15. Version-specific companion: *Closed-Loop ACAS Xu NNCS is Unsafe: Quantized State Backreachability for Verification*, arXiv 2201.06626v1 (2022), https://doi.org/10.48550/arXiv.2201.06626
6. Pang, Q., Yuan, Y., Wang, S., 31st ACM SIGSOFT International Symposium on Software Testing and Analysis: MDPFuzz: Testing Models Solving Markov Decision Processes, Proceedings of the 378–390. (2022). https://doi.org/10.1145/3533767.3534388. Version-specific companion: *MDPFuzzer: Finding Crash-Triggering State Sequences in Models Solving the Markov Decision Process*, arXiv 2112.02807v1 (2021), https://doi.org/10.48550/arXiv.2112.02807
7. Kabaha, A., Drachsler, D., Cohen: Quantization with Guaranteed Floating-Point Neural Network Classifications, Proceedings of the ACM on Programming Languages 9(OOPSLA2) Article 340, 1893–1920 (28 pages). (2025). https://doi.org/10.1145/3763118
8. Boetius, D., Leue, S., Sutter, T., A Robust Optimisation Perspective on Counterexample-Guided Repair of Neural Networks:, Proceedings of Machine Learning Research 202 2712–2737. https://proceedings.mlr.press/v202/boetius23a.html. Associated experimental artefact, version 1.0.0: (2023). https://doi.org/10.5281/zenodo.7938547. Accessed 5 October 2026
9. Bauer-Marquart, F., Boetius, D., Leue, S., Schilling, C.: SpecRepair: Counter-Example Guided Safety Repair of Deep Neural Networks, Model Checking Software, SPIN 2022 79–96. https://doi.org/10.1007/978-3-031-15077-7_5. (2022). Extended author version: arXiv 2106.01917v5, https://doi.org/10.48550/arXiv.2106.01917
10. Ishimoto, Y., Kondo, M., Ma, L., Ubayashi, N., Kamei, Y.: Repairs and Breaks Prediction for Deep Neural Networks. ACM Trans. Softw. Eng. Methodol. **34**(4), 1–42 (2025). https://doi.org/10.1145/3702983
11. Liang, Z., Wu, T., Zhao, C., Liu, W., Xue, B., Yang, W., Wang, J., Huang, W.: BIRDNN: Behavior-Imitation Based Repair for Deep Neural Networks. Neural Netw. **183** (2025). Article 106949 https://doi.org/10.1016/j.neunet.2024.106949
12. Bao, L., Wang, Y., Sun, T.: NCCDA: Neuron-wise Class-Conditional Distribution Alignment for Deep Neural Network Repair, Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining, volume 2 115–126. (2026). https://doi.org/10.1145/3770855.3818021
13. Bao, L., Wang, Y., Sun, T.: SPGR: Source-Path Guided Repair for Deep Neural Networks, ICASSP 2026 — IEEE International Conference on Acoustics, Speech and Signal Processing 2316–2320. (2026). https://doi.org/10.1109/ICASSP55912.2026.11460506
14. Liu, F., Du, X., Ding, H., Qian, J.: Towards robust neural networks: Exploring counterfactual causality-based repair. Expert Syst. Appl. **257** (2024). Article 125082 https://doi.org/10.1016/j.eswa.2024.125082
15. Du, X., Zhang, Z., Wang, X., Liu, L., Qian, J.: Neural Network Repair With Shapley-Guided Search. CAAI Trans. Intell. Technol. **11**(3), 816–834 (2026). https://doi.org/10.1049/cit2.70147
16. Lyu, D.: Quality Assurance of AI-Enabled Cyber-Physical Systems, doctoral dissertation, Kyushu University (December 2024), Chap. 7, printed pages 130–151. https://hdl.handle.net/2324/7363810. Accessed 4 October 2026
17. Lu, P., Cleaveland, M., Sokolsky, O., Lee, I., Ruchkin, I.: Repairing neural network-based control policies with safety preservation, Research Directions. Cyber-Physical Syst. **3**, e5 (2025). https://doi.org/10.1017/cbp.2025.10003
18. Stefani, T., Christensen, J.M., Girija, A.A., Gupta, S., Durak, U., Köster, F., Krüger, T., Hallerbach, S.: Automated scenario generation from Operational Design Domain model for testing AI-based systems in aviation. CEAS Aeronaut. J. **16**(1), 197–212 (2025). first published online 16 November 2024 https://doi.org/10.1007/s13272-024-00772-4

19. Katz, G., Barrett, C., Dill, D.L., Julian, K., Kochenderfer, M.J., Reluplex: An Efficient SMT Solver for Verifying Deep Neural Networks, Verification, C.A.: CAV (2017) 97–117. (2017). https://doi.org/10.1007/978-3-319-63387-9_5. Extended author version: arXiv 1702.01135v1, https://doi.org/10.48550/arXiv.1702.01135
20. Sheikhi, S., Bak, S., Closed-Loop, A.C.A.S., Xu Neural Network, Verification: EPiC Series in Computing 96 1–8. https://doi.org/10.29007/vf8z. Publisher full text: (2023). https://easychair.org/publications/open/mmSLW. Accessed 5 October 2026
21. [dataset], V., Fonseka, R.M., Abeykoon: Evaluating Neural Network Repairs in ACAS Xu: Static Fidelity and Off-Target Encounter Effects - Data and Reproduction Materials, Zenodo, version 1.0.0 (2026). https://doi.org/10.5281/zenodo.23202447
22. Bahari Kordabad, A., Ghosh, A., Soudjani, S., Stroeve, S., Intent-Aware: MPC for Aircraft Detect-and-Avoid with Response Delay: A Comparative Study with ACAS Xu, SESAR Innovation Days paper 49. (2025). https://hdl.handle.net/10921/1817. Version used to specify the sensitivity settings: arXiv:2503.23518v1 (30 March 2025). https://arxiv.org/abs/2503.23518v1. The proceedings version retains the one- and four-second settings in Section II-B (pp. 2–3). Accessed 9 October 2026

Figures

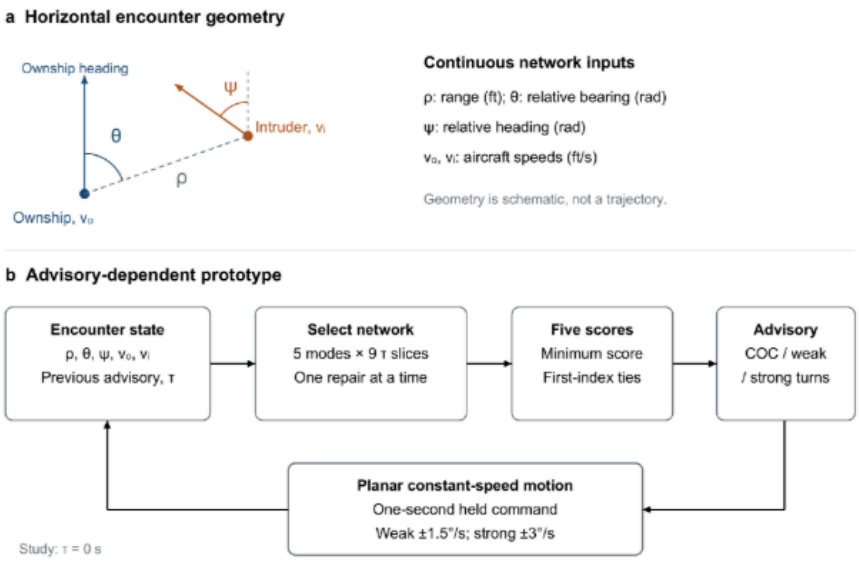

Figure 1

Encounter geometry and the evaluated ACAS Xu prototype. (a) Horizontal encounter geometry; (b) advisory-dependent prototype feedback. The five continuous inputs are range ρ, relative bearing θ, relative heading ψ, and speeds v_{own}/v_{int}. Previous advisory and τ select one of 45 networks; the minimum of five scores selects the next advisory. The co-altitude study uses the five τ=0 networks and ideal one-second motion updates. Geometry is schematic, not a recorded trajectory. The schematic was drawn with Python ReportLab from the pinned controller definitions [2,3], using code drafted; no primary research image was generated or altered

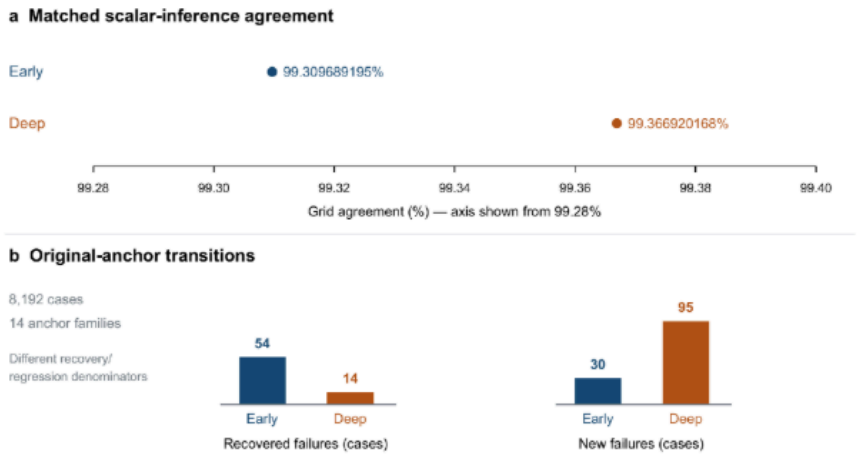

Figure 2

N3 static agreement and original-anchor encounter transitions. (a) Matched scalar agreement: early 99.309689195%, deep 99.366920168%, on the same 24,299,779 retained grid points (Table 4). (b) The 8,192-case fourteen-family mixture gives early 54 recoveries/30 new failures and deep 14/95 (Table 3). Axes distinguish percentages and counts; recovery and regression denominators differ. Drawn from these unchanged values with Python ReportLab using code drafted.

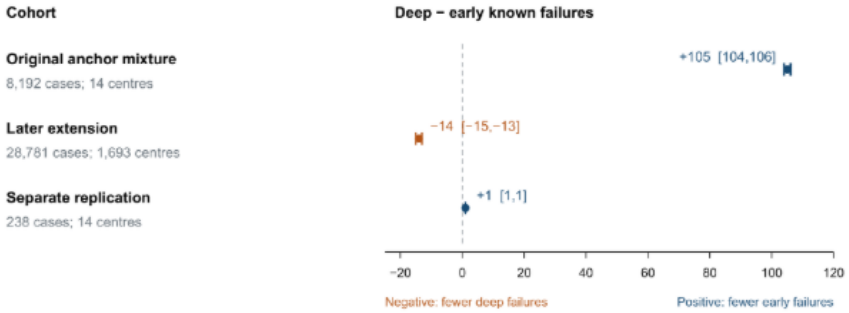

Figure 3

Deep-minus-early known-failure differences in three separate cohorts. Positive values favour early; negative values favour deep. Original anchors give +105 with numerical envelope [104,106], the later extension −14 with [−15,−13], and the separate replication +1 with [1,1] (Table 5). Envelopes allow binary assignment of numerical UNKNOWN and are not confidence intervals. Case and centre counts identify the different cohort constructions, which are not pooled. Plots were drawn with Python ReportLab from the recorded table values, using code drafted, with no new experiment or generated observations

Supplementary Files

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

- ESM1.pdf
- ESM2.zip