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.

Starlight: Predicting Design Ranking Reversals in an Autonomous Cars from Measured Steering Gains

Aarav Jalan

aaravjalan.chess@gmail.com

Adani International School

Darsh Zaveri

Apple Global School

Research Article

Keywords: Autonomous robotics Robot, design selection, Simulation uncertainty, Matched Monte Carlo experiments, Simultaneous confidence intervals, Robust decision-making

Posted Date: November 25th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11830339/v1

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

Additional Declarations: The authors declare no competing interests.

Starlight: Predicting Design Ranking Reversals in an Autonomous Car from Measured Steering Gain

Aarav Jalan[1,*] **Darsh Zaveri**[2]

[1]Adani International School, Ahmedabad, Gujarat, India
[2]Apple Global School, Ahmedabad, Gujarat, India

[*]**Corresponding author:** Aarav Jalan,
aarav.adani2353@adaniinternational.edu.in

Abstract

Engineers often choose designs through trial and error, but small calibration changes can reverse their performance rankings. Limited tests may also make uncertain choices appear decisive. Starlight investigates how recognizing these reversals can lead to more reliable decisions.

We compared two bundles, A and D, each containing seven robot-design features, across 752,000 simulated missions. Our decision procedure used paired failure-rate differences and simultaneous uncertainty bounds, balancing the costs of further testing, incorrect selections and withholding judgment. It could select a winner, accept either design, reject both as inadequate or request more evidence. We compared it with seven baselines and three ablations. Paired physical docking trials tested Clock, which used preset turns, and Compass, which corrected heading through feedback, under controlled steering mismatch. Linkage-change trials and blind ratings assessed hardware effects, while colour sorting tested generalizability.

A 5% steering-gain increase reversed simulated rankings: A led D by 1.050 percentage points at gain 1.200, while D led A by 2.925 at 1.260. The method withheld selection in 29 of 66 conditions and reduced overall decision cost by up to 33% against seven baselines. Clock's mean heading shifted 6.42° under mismatch, compared with 0.48° for Compass. The ten-unit audit matched eight held-out references but made one unsupported winner call. In three-colour sorting, the framework favoured calibrated detection (S2) over fixed-threshold detection (S1), while accepting detection with rechecking (S3) as an alternative.

To sum it up, Starlight shows how accounting for uncertainty improves calibration-sensitive robot-design decisions.

Keywords: Autonomous robotics; Robot design selection; Simulation uncertainty; Matched Monte Carlo experiments; Simultaneous confidence intervals; Robust decision-making.

Journal categories: (2) Systems Modelling/Simulation/Control/Computer-Aided Design; (3) Robotic Modelling/Robot Control; (9) Decision Support Systems.

Acknowledgements. The authors thank Sunil Solanki (Robofun Lab India) for mentorship throughout the project and assistance with physical robot construction and integration, and Deyaan Agrawal (Ahmedabad University) for coding guidance and implementation support.

1 Introduction

When two robotic designs are evaluated under the same nominal calibration, the one with fewer failures is easily called "better." Yet this decision confounds two different questions: *which design wins at this setting?* and *does the ordering remain defensible as a relevant assumption changes?* In a compact autonomous car, sensing, commanded steering, realized wheel angle, numerical

integration, and parking success form a chain in which modest variation at one interface can affect the final mission outcome.

Starlight arose from a student-built Ackermann-steered autonomous vehicle and its engineering test bench. The broader work included perception, steering, movement, docking, and safety investigations. The scientific focus of the present paper is the *identification and statistical certification of sensitivity in a paired design ranking*. The physical robot motivates perturbations and reveals engineering constraints; it does not calibrate the complete simulator to a real-world probability distribution.

The contributions are threefold. First, we provide an explicit finite-family comparison of two seven-feature source-informed design bundles under coupled changes to steering gain, disturbance intensity, numerical timestep, and selected model forms. Second, we apply a simple simultaneous exact-binomial construction to paired mission failures, and use an independent confirmation bank selected by a frozen search rule to check opposing sign witnesses. Third, we connect the statistical result to a decision policy that does not interpret a non-significant difference as equivalence or a supported winner as automatically adequate.

The central finding is not that one design is universally superior. Rather, the experiments demonstrate that within the evaluated simulator family the preferred design can reverse, even when independent finite-sample intervals support each direction at two specified settings.

2 Related Work

2.1 Simulation fidelity is distinct from Monte Carlo precision

Sargent [1] distinguishes simulation verification from validation, while Jakobi et al. [2] describe the longstanding reality gap in robot simulation. Domain and dynamics randomization have been used to improve transfer of learned robotic policies [3, 4], but such transfer is not demonstrated in Starlight. Kadian et al. [5] study whether success in simulation predicts physical performance; Li et al. [6] demonstrate ways to improve sim-to-real policy evaluation in manipulation. Badithela et al. [7] develop an approach that combines simulations with limited real rollouts to correct performance inference using imperfect simulators. These contributions underscore an important boundary for the present work: many simulated trajectories may reduce *sampling error within a model* without reducing *model error against a physical system*. Our intervals therefore refer only to the declared simulator specifications. We do not estimate a simulation-to-hardware correction or claim physical risk coverage.

2.2 Ranking-and-selection, paired experiments, and uncertainty

Using common random numbers to compare simulated alternatives is well established [8]. Paired-proportion interval methods exploit the structure of jointly observed binary outcomes [9]; our chosen construction instead emphasizes a transparent finite-sample joint-coverage guarantee. Distributionally robust selection asks how a choice behaves across input uncertainty classes [10]. Starlight does not solve a general distributionally robust optimization problem; it asks whether two opposing, simultaneously supported signs can be exhibited in a finite, explicitly enumerated family. Recent robot policy comparison research provides sophisticated, including sequential, statistical methods [11]; our design instead fixes sample sizes and uses direct binomial bounds, keeping the inference auditable. We distinguish confidence intervals for a difference from affirmative equivalence evidence, following the general principle of specifying practical equivalence margins [12]. More powerful step-down multiplicity procedures also exist [13], but this paper retains a predetermined Bonferroni allocation for direct auditability. Unlike anytime-valid confidence sequences [14], our bounds do not license arbitrary sequential stopping.

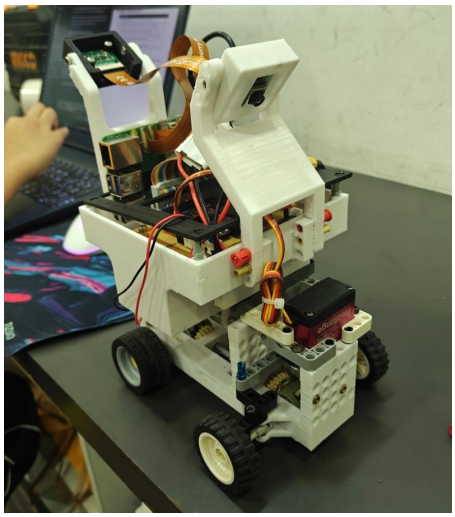


Fig. 1 Physical Starlight test vehicle. The physical measurements and the simulated A/D architecture comparison are distinct evidence streams: physical controller tests do not establish a physical A/D ranking reversal.

3 Robot Platform and Simulator

3.1 Autonomous robot and instrumentation

The physical Starlight platform is an approximately 18.5 cm long, 12.8 cm wide, camera-guided, Ackermann-style autonomous car developed through iterative mechanical and software revisions. Its measurement work includes steering-command versus road-wheel response, end-to-end sensing-to-actuation latency, visual docking pose estimation, paired timed versus feedback-heading control, and a reversible linkage modification. Different hardware versions occur across the development record; a physical result should not be presumed to apply unchanged to every hardware revision.

3.2 Source-informed surrogate and architecture contrast

The simulator is a mechanistic, source-informed *surrogate*, with geometric course projection replacing absent pixel-level image projection. It models motion, obstacle geometry, cue uncertainty, steering command limits, steering response, delayed actuation, event-based lap counting, and parking-phase transitions. Its output is a binary mission outcome with an explicit failure code. The simulator does not execute the deployed Raspberry Pi image-processing program verbatim, and several physical constants and disturbance scales are engineering assumptions, not population-calibrated distributions.

The design matrix has seven binary features in the following order: repeatable-mount attenuation; bounded servo commands; fused perception scoring; apparent-area gating; magenta event cooldown behaviour; a rear-conditioned initial parking trigger; and heading-terminated parking. Bundle A sets all seven flags to zero; bundle D sets all seven to one. Intermediate B/C designs are defined in the code, but the present primary matched comparison uses A and D. Consequently, the A/D contrast concerns *bundled architectures* and does not identify the causal contribution of any one switch.

Table 1: The declared robot mission simulation and its model-input grid.

Dimension	Implementation in the computational experiment
Primary comparison	A and D run on the same generated scenario and disturbance vector; failure is binary.
Steering gain g	11 values from 1.200 to 1.300 in increments of 0.010.
Disturbance scale s	0.015, 0.030, and 0.050; ten named disturbance coordinates.
Integration step Δt	0.010 s and 0.040 s.
Within-cell sampling	4,000 matched pairs for each of 66 grid settings.
Model-form modifications	Nonlinear steering, interpolated fractional delay, triangular disturbances, expanded clearance envelope.
Independent confirmation	At most eight frozen selections; 8,000 new matched pairs for each of eight retained settings.

4 Experimental and Statistical Methods

4.1 Matched binary risk estimand

For model setting $\theta \in \Theta$, let $Y_{A,i}(\theta), Y_{D,i}(\theta) \in \{0, 1\}$ denote mission failure indicators for the two architectures on the same scenario i, with 1 indicating failure. Define

$$X_i(\theta) = Y_{D,i}(\theta) - Y_{A,i}(\theta) \in \{-1, 0, 1\}. \tag{1}$$

The population failure-risk difference and its matched estimate are

$$\begin{aligned} \Delta(\theta) &= \mathbb{P}(Y_D = 1) - \mathbb{P}(Y_A = 1), \\ \widehat{\Delta}(\theta) &= \frac{n_+(\theta) - n_-(\theta)}{N}. \end{aligned} \tag{2}$$

where n_+ counts D-only failures and n_- counts A-only failures. A positive Δ favours A; a negative value favours D. Pairs in which both designs succeed or both fail contribute zero to the difference, although they remain essential to absolute-reliability assessment. The same scenario seed and disturbance vector are supplied to A and D to support a matched comparison. The binomial argument below assumes independent scenario pairs *within each setting*; it does not require independence of comparisons across model settings.

4.2 Simultaneous Clopper–Pearson construction

Let $p_+(\theta) = \mathbb{P}(X = 1)$ and $p_-(\theta) = \mathbb{P}(X = -1)$. Each corresponding marginal count is binomial across independent matched pairs, even though the two discordance events are mutually exclusive within a pair. For a fixed family of K settings with family error budget α, construct two-sided Clopper–Pearson marginal intervals for p_+ and p_-, each with noncoverage allocation $\alpha/(2K)$ [15]. Thus each tail uses $\alpha/(4K)$. Denote the resulting endpoints by (ℓ_+, u_+) and (ℓ_-, u_-); then

$$\mathcal{I}(\theta) = [\, \ell_+(\theta) - u_-(\theta),\ u_+(\theta) - \ell_-(\theta)\,]. \tag{3}$$

By a union bound across $2K$ marginal intervals, $\mathbb{P}\{\Delta(\theta) \in \mathcal{I}(\theta) \text{ for all } \theta \in \Theta\} \geq 1 - \alpha$. No independence between the two discordance counts or between different settings is required for this multiplicity argument. The intervals are conservative because the component intervals are combined by subtraction.

A **strict A certificate** is assigned if the lower endpoint exceeds zero; a **strict D certificate** if the upper endpoint is below zero; otherwise the ranking is **unresolved**. In particular, unresolved *does not mean* the designs are equivalent. For a finite family, one certified A setting and one certified D setting jointly demonstrate *non-uniform ordering* on the joint coverage event.

Proposition 1 (Finite-family sign validity)**.** *Under independent within-setting scenario pairs and correctly computed binary outcomes, the sign classifications based on* (3) *are simultaneously valid for all K declared specifications with probability at least $1 - \alpha$.*

Proof. For each setting, both discordance counts have binomial marginal laws. Exact Clopper–Pearson intervals provide their allocated marginal coverage. The union bound gives simultaneous coverage of all $2K$ probabilities; interval subtraction then contains each $p_+ - p_-$ on that event. A strict sign stated from an interval excluding zero is correct on the same event. □

As a supplementary check, bounded-difference Hoeffding bounds can be formed using

$$\varepsilon_N = \sqrt{\frac{2\log(2K/\alpha)}{N}}, \qquad \left[\widehat{\Delta} - \varepsilon_N, \widehat{\Delta} + \varepsilon_N\right] \cap [-1, 1], \tag{4}$$

for $X_i \in [-1, 1]$ [16]. The reported classifications here use the Clopper–Pearson construction.

4.3 Predeclared grid, model forms, and independent confirmation

The primary experiment crosses 11 steering gains, three disturbance scales, and two timesteps. Its 66 settings each contain 4,000 matched pairs, yielding 528,000 missions. Four alternative model assumptions are then evaluated at three gains each with 4,000 pairs, yielding 96,000 additional missions. Three existing baseline cells are reused for comparison in the 15-row model-form report and are not counted as new missions.

After evaluating the fixed primary grid, the confirmation-selection rule retains a nominal setting, extrema of the observed risk difference, a near-nominal negative point, and gain midpoints favouring adjacent observed sign changes, subject to de-duplication and a maximum of eight settings. The eight selected settings are rerun using a fresh master random seed and 8,000 matched pairs each. In total, the study records 376,000 pairs or 752,000 simulated robot missions. The primary and confirmation seeds are 202610031 and 202610032, respectively; they define reproducible pseudorandom sampling streams but do not validate the physical distributions.

The family error budgets are 0.02 (primary grid), 0.01 (model forms), and 0.01 (independent confirmation). An additional 0.01 is reserved in the protocol for a separate CartPole experiment; those episodes are excluded from the robot mission total. The sum is 0.05 across the four declared families under the stated sampling assumptions.

Proposition 2 (Independent post-selection confirmation)**.** *If the selection rule uses only the exploratory bank and the confirmation bank is independent of it, the simultaneous confidence statement for up to eight selected settings remains valid conditionally on the exploratory observations.*

Proof. After conditioning on the exploratory bank, the selected settings become fixed. The independent confirmation samples retain their specified sampling distributions, so the finite-family confidence argument applies conditionally. Integrating over exploratory selections preserves the advertised coverage. □

4.4 Five-action engineering policy

A ranking interval answers only a relative-performance question. A separate executable decision layer uses a minimum acceptable reliability, a practical equivalence margin, and declared testing and error costs. Its precedence is: **reject both** if both reliability upper bounds fall below the minimum; **accept either** if both lower reliability bounds meet the minimum and the entire difference interval lies within the equivalence margin; **select A** if the difference interval is strictly positive and A passes the reliability threshold; **select D** under the symmetric negative condition; otherwise **request more evidence**. A further test is *permitted* only if its declared cost fits the budget and does not exceed the specified error cost. This is an operational rule, not a claimed optimal value-of-information calculation. The present paper evaluates the interval evidence; it does not claim validated cost savings from the policy.

4.5 Physical experiments and scope

The physical records comprise: 126 steering-calibration readings; 200 response-latency events; 105 camera pose-error frames; paired Clock (preset-duration) and Compass (heading-feedback) controller tests across separately recorded sessions; an A–B–A′ mechanical-linkage series of 12 runs per phase; 24 emergency-stop trips; a ten-unit decision-audit ledger; and a separate colour-sorting experiment. The docking setup uses ArUco-type fiducial pose references [17]. The physical controller pairs and simulated A/D architecture pairs answer different questions and must not be numerically pooled. The sensor and hardware tests are summarized descriptively; released processed tables do not by themselves authenticate instrument binaries, photographs, or all session-level confounders.

4.6 Manuscript preparation transparency

OpenAI ChatGPT assisted with drafting and revising the manuscript text and with LaTeX preparation. The scientific results and their interpretation remain the responsibility of the named authors; the generated content is not a source of experimental evidence. This disclosure concerns manuscript preparation, not a claimed use of generative AI to produce the measured experimental data.

5 Results

5.1 A finite-family ranking is not stable across model settings

Of the 66 primary-grid cells, eight have positive (A-favouring) simultaneous intervals, 29 have negative (D-favouring) intervals, and 29 remain unresolved (Fig. 2). Thus there is no uniformly supported preference for either design across the declared model family. More simulation at one nominal setting would sharpen that setting's estimate, but cannot by itself establish ordering over a family in which opposite signs are supported.

5.2 Independent confirmation of opposing signs

At a common disturbance scale $s = 0.030$ and timestep $\Delta t = 0.040$ s, the confirmed risk difference at $g = 1.200$ is $+1.050$ percentage points ($n_+ = 90$, $n_- = 6$, $N = 8000$), with interval $[+0.518, +1.576]$ pp. At $g = 1.260$, it is -2.925 pp ($n_+ = 59$, $n_- = 293$), with interval $[-3.982, -1.861]$ pp. These settings differ by 0.060 in the specified simulated gain, or 5% relative to 1.200. Both statements use the confirmation-family allocation of $\alpha = 0.01$ across eight comparisons (Fig. 3). The result establishes opposite signs *at these two discrete model settings*, not a continuous tipping point or a measured 5% physical steering change.

The two additional confirmations at $g = 1.200, s = 0.050$ and $g = 1.300, s = 0.015$ yield positive and negative certificates respectively. Recomputed intervals are $[+1.225, +3.019]$ pp and $[-10.122, -7.655]$ pp. Several stored endpoint pairs in the processed confirmation CSV do not reproduce the stated formula at the declared family size; the intervals in this paper are calculated directly from n_+, n_-, N, K, and α. All eight signs remain unchanged after this correction.

5.3 Sensitivity to simulator assumptions

The alternative-model family includes nonlinear steering response, fractional delay interpolation, triangular instead of uniform disturbance sampling, and an expanded collision-clearance envelope. Among its 15 reported comparisons (including three reused baseline cells), five favour A, five favour D, and five are unresolved. This is not a model-form-robust victory for a single architecture. The clearance modification changes the criterion used to judge space around the car; the nonlinear and delay variants modify assumed dynamics, while the triangular variant modifies the disturbance law.

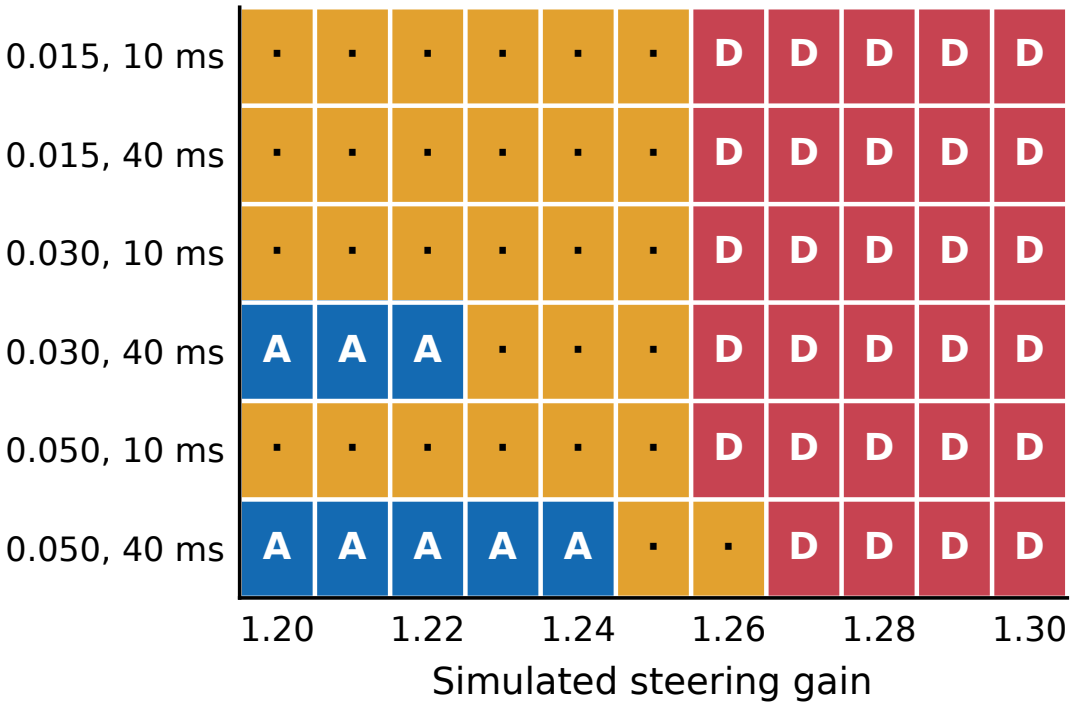


Fig. 2 Primary-grid signs across 66 settings. Cells labelled A denote intervals strictly above zero; D denotes intervals strictly below zero; dots denote unresolved comparisons. Each cell contains 4,000 matched A/D scenarios and belongs to one simultaneous family ($\alpha = 0.02$). The symbols denote model-conditional ordering, not physical superiority.

Table 2: Independent confirmation of ranking witnesses at $s = 0.030$ and $\Delta t = 0.040$ s. Confidence intervals are recalculated, not merely copied from stored rounded endpoints.

Gain	n_+	n_-	$\widehat{\Delta}$ (pp)	Joint interval (pp)
1.200	90	6	$+1.050$	$[+0.518, +1.576]$
1.230	72	28	$+0.550$	$[-0.055, +1.153]$
1.245	87	31	$+0.700$	$[+0.048, +1.350]$
1.255	67	93	-0.325	$[-1.099, +0.450]$
1.260	59	293	-2.925	$[-3.982, -1.861]$
1.280	35	310	-3.438	$[-4.439, -2.427]$

The variants are engineering stress tests, not independently established plausible distributions of physical operating conditions.

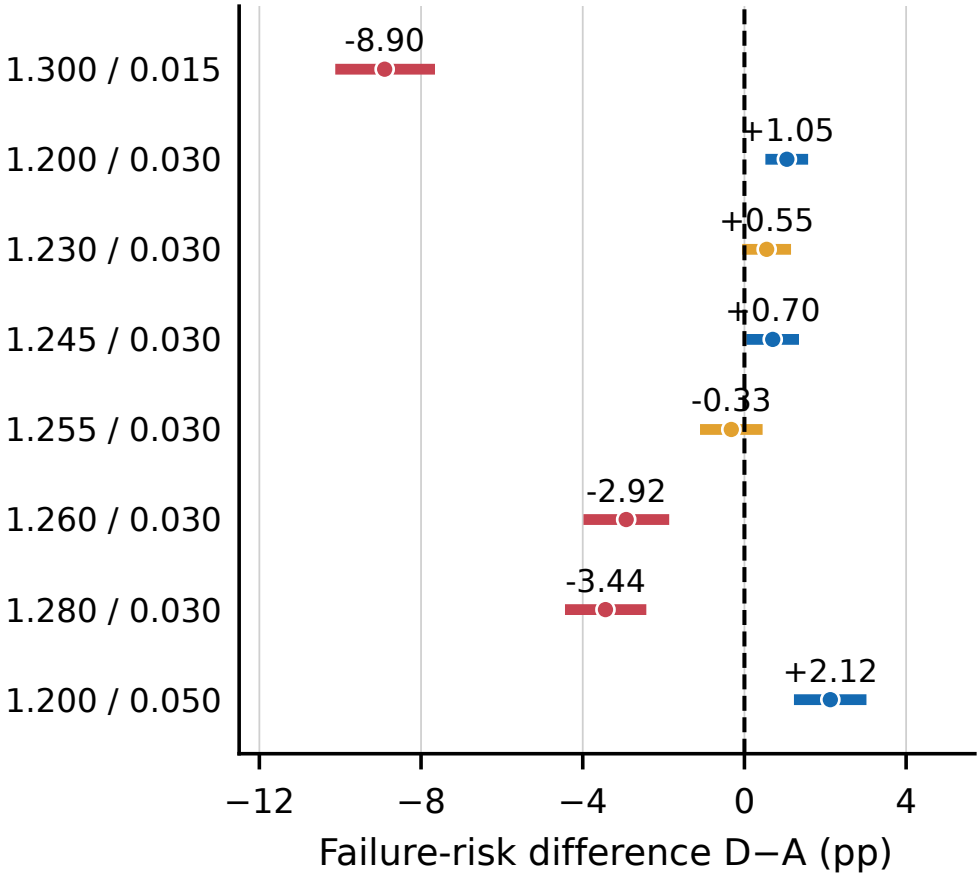


Fig. 3 All eight independently confirmed design comparisons. Points show D-minus-A failure differences; segments are simultaneous Bonferroni–Clopper–Pearson confidence intervals recomputed from the recorded discordant counts (family $\alpha = 0.01$, $K = 8$). Positive intervals favour A, negative intervals favour D, and intervals crossing zero are unresolved. The six settings at $s = 0.030$ share $\Delta t = 0.040$ s; the other two stress settings are labelled explicitly.

5.4 Physical results: useful mechanisms, separate estimands

The physical steering-calibration table contains 126 observations; a fitted command-to-wheel response slope is approximately 1.238, but this measured scale is *not identified with* the simulator's gain g. In the linkage-mismatch cohort, the Clock controller has mean absolute heading shift 6.42° versus 0.48° for Compass (18 matched controller pairs); a separate dial-mismatch pilot has means 6.80° and 1.10° (22 pairs). The linkage-matched cohort has means 6.64° and 0.65° (10 pairs). Because these cohort means do not isolate the effect of a manipulated mismatch, the evidence is descriptive support for differing controller responses, not a clean causal estimate of calibration mismatch or a direct physical A/D reversal.

Additional physical observations place limits on deployment claims (Table 3). The docking-camera positional 95th percentile meets its 2.5 mm criterion narrowly, while heading does not meet its 1.0° criterion. Emergency stopping is confirmed for 23 of 24 logged trips, not all 24. In the 12-run-per-phase reversible linkage experiment, mean response changes from approximately 1.237 at baseline to 1.268 under modification and 1.245 after restoration; mean servo current rises from 0.420 A to 0.538 A in the modified phase. These are measurements of a physical intervention, not proof that the simulated 5% steering parameter change occurred in the hardware.

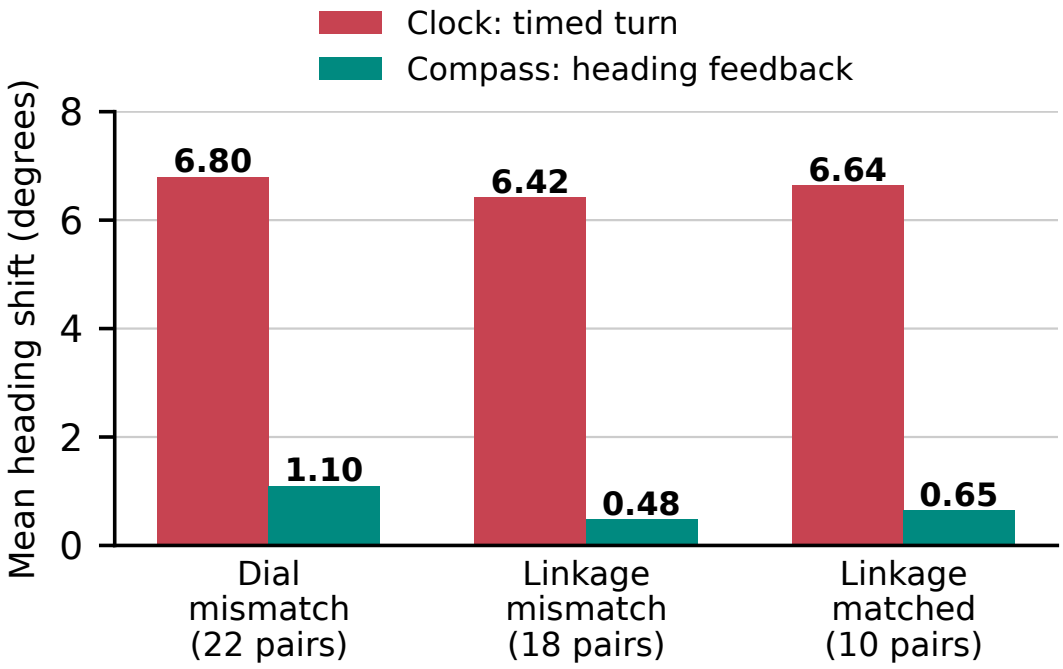


Fig. 4 Separate physical Clock–Compass comparisons. These numbers describe observed absolute heading shifts within each recorded cohort. Matched and mismatched cohorts are not interpreted as a randomized treatment-effect estimate.

Table 3: Arithmetic summaries from the processed physical research tables. These are separate experiments and are not pooled with the simulated A/D missions.

Endpoint	Recorded result	Interpretation
Emergency stop	23/24 confirmed cuts; mean 14.20 ms among 23 successful cuts	One safety failure remains in the logged set.
Actuation latency	3/200 over 33 ms; 0/200 over 50 ms	Deadline counts, not worst-case guarantees.
Docking metrology	Position $P_{95} = 2.480$ mm; heading $P_{95} = 1.286°$	Position passes 2.5 mm target; heading fails 1.0° target.
Reversible linkage	Gain response $1.237 \rightarrow 1.268 \rightarrow 1.245$	Restoration is approximate, not identical.
Decision audit	8/10 agree with stored reference; 1 unsupported winner flag	Small processed audit, not an externally authenticated decision trial.
Colour sorter	S1: 188/238; S2: 202/216; S3: 198/219	Descriptive transfer comparison; non-independent batches.

6 Discussion

6.1 Why two opposite signs carry an engineering message

Suppose a design is declared best using only the nominal simulator. If a small, explicitly chosen model perturbation supports the opposite sign on fresh data, the nominal claim is inadequate as a statement about *all* declared operating assumptions. The conclusion is deliberately asymmetric: two opposing witnesses disprove uniform ordering over a family containing both, whereas proving uniform preference requires evidence for the same strict sign at every family member. This distinction can make unresolved output an intellectually meaningful engineering result rather

than a failure of the procedure.

The result complements, rather than supersedes, earlier work on simulator predictivity, distributionally robust selection, or sequential robotic policy evaluation [5, 10, 11]. The innovation demonstrated here is the traceable connection between a hardware-motivated calibration question, finite-family simultaneous risk comparisons, independent witness confirmation, and an engineering policy that permits nonselection.

6.2 What the physical work does and does not establish

The hardware record demonstrates that steering response, docking metrology, controller behaviour, and safety acceptance limits require explicit measurement. It is consistent with the *motivation* for sensitivity-aware testing. However, the physical Clock–Compass experiment and simulator A/D design comparison are not interchangeable. The physical data do not identify the real-world A/D risk difference, establish the simulator gain distribution, or supply a validated joint sim-to-real calibration. More Monte Carlo missions cannot repair that inferential gap. A claim of predictive deployment would require paired real A/D trials, instrument logs and model calibration against physical outcomes [6, 7].

6.3 Limitations and threats to validity

Model validity. The numerical model is a source-informed surrogate, not an independently validated digital twin. Its geometry, visual evidence proxies, delayed actuation, and probability assumptions can induce ranking changes of their own; this is one reason to distinguish model sensitivity from independent validation [18]. The perturbation set is a finite engineering stress family; no probability law over true physical steering gain is established.

Sampling and multiplicity. The simultaneous intervals use independent matched scenarios within each model cell. Reused random streams across cells are not a multiplicity problem for a union bound, but within-cell dependence or mismatched labels would invalidate the stated binomial calibration. The fixed-size analysis is not an anytime-valid sequential test.

Data provenance. The publicly released physical and simulation CSVs contain processed records or discordant sufficient counts. Source integrity checks and internal arithmetic verification do not authenticate the original physical instrument streams, videos, or generation history of every simulated mission. Several stored confirmation endpoints required recomputation from counts; reporting the corrected intervals improves transparency but does not resolve all provenance questions.

Generalization and causal attribution. A/D compares full design bundles, so feature-level causal attribution is not identified. Clock/Compass and sorting studies concern distinct apparatus and sampling structures; they cannot be pooled into a global performance or industrial efficiency claim. A deployment result cannot be inferred from a prospective industry discussion or a simulator-only cost scenario.

7 Code and Data Availability

The source repository is `https://github.com/Aarav-J-10/Starlight-Ranking-Certificates`. The public structure includes the certificate-analysis package, processed-data reference archive, and simulator source under `scripts/models`. The released tables permit independent checking of reported discordance arithmetic and confidence intervals. The experiment runner stores declared seeds and a frozen protocol; however, full mission-level regeneration from an independently tested, environment-locked public release has not yet been demonstrated. The publicly released data support independent recomputation of the paper's principal risk-difference estimates and statistical intervals. The physical raw sensor streams and a complete independently executed simulator rebuild are not included in the current release.

8 Conclusion

In a source-informed robotic surrogate, a 5% change in declared simulated steering gain from 1.200 to 1.260 changed the direction of the independently supported A/D mission-failure ranking. Across 66 primary model settings, 29 remained unresolved and 37 favoured one of the two designs. This finite-family result shows why a nominal winner can be a fragile engineering conclusion even when its within-setting Monte Carlo estimate is precise. The accompanying physical metrology makes the engineering motivation concrete, while deliberately stopping short of claiming physical A/D reversal or real-world risk coverage. The proposed workflow is therefore best understood as an *evidence discipline*: compare under matched conditions, control uncertainty across declared alternatives, test informative sign witnesses afresh, and separate winner selection from equivalence, adequacy, and the need for more evidence.

References

[1] Sargent, R.G.: Verification and validation of simulation models. *Journal of Simulation* **7**(1), 12–24 (2013). https://doi.org/10.1057/jos.2012.20

[2] Jakobi, N., Husbands, P., Harvey, I.: Noise and the reality gap: the use of simulation in evolutionary robotics. In: *Advances in Artificial Life*, Lecture Notes in Computer Science, vol. 929, pp. 704–720 (1995). https://doi.org/10.1007/3-540-59496-5_337

[3] Tobin, J., Fong, R., Ray, A., Schneider, J., Zaremba, W., Abbeel, P.: Domain randomization for transferring deep neural networks from simulation to the real world. In: *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 23–30 (2017). https://doi.org/10.1109/IROS.2017.8202133

[4] Peng, X.B., Andrychowicz, M., Zaremba, W., Abbeel, P.: Sim-to-real transfer of robotic control with dynamics randomization. In: *Proceedings of the IEEE International Conference on Robotics and Automation*, pp. 3803–3810 (2018). https://doi.org/10.1109/ICRA.2018.8460528

[5] Kadian, A., Truong, J., Gokaslan, A., Clegg, A., Wijmans, E., Lee, S., Savva, M., Chernova, S., Batra, D.: Sim2Real predictivity: does evaluation in simulation predict real-world performance? *IEEE Robotics and Automation Letters* **5**(4), 6670–6677 (2020). https://doi.org/10.1109/LRA.2020.3013848

[6] Li, X., et al.: Evaluating real-world robot manipulation policies in simulation. In: *Proceedings of the Conference on Robot Learning*, PMLR, vol. 270, pp. 3705–3728 (2025). https://proceedings.mlr.press/v270/li25c.html

[7] Badithela, A., Snyder, D., Zha, L., Mikhail, J., O'Kelly, M., Dixit, A., Majumdar, A.: Reliable and scalable robot policy evaluation with imperfect simulators. In: *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, pp. 21728–21735 (2026). https://arxiv.org/abs/2510.04354

[8] Nelson, B.L., Matejcik, F.J.: Using common random numbers for indifference-zone selection and multiple comparisons in simulation. *Management Science* **41**(12), 1935–1945 (1995). https://doi.org/10.1287/mnsc.41.12.1935

[9] Agresti, A., Min, Y.: Simple improved confidence intervals for comparing matched proportions. *Statistics in Medicine* **24**(5), 729–740 (2005). https://doi.org/10.1002/sim.1781

[10] Fan, W., Hong, L.J., Zhang, X.: Distributionally robust selection of the best. *Management Science* **66**(1), 190–208 (2020). https://doi.org/10.1287/mnsc.2018.3213

[11] Snyder, D., et al.: Beyond binary success: sample-efficient and statistically rigorous robot policy comparison. In: *Robotics: Science and Systems XXII* (2026). https://doi.org/10.15607/RSS.2026.XXII.076

[12] Lakens, D.: Equivalence tests: a practical primer for t tests, correlations, and meta-analyses. *Social Psychological and Personality Science* **8**(4), 355–362 (2017). https://doi.org/10.1177/1948550617697177

[13] Holm, S.: A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics* **6**(2), 65–70 (1979).

[14] Howard, S.R., Ramdas, A., McAuliffe, J., Sekhon, J.: Time-uniform, nonparametric, nonasymptotic confidence sequences. *The Annals of Statistics* **49**(2), 1055–1080 (2021). https://doi.org/10.1214/20-AOS1991

[15] Clopper, C.J., Pearson, E.S.: The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika* **26**(4), 404–413 (1934). https://doi.org/10.1093/biomet/26.4.404

[16] Hoeffding, W.: Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association* **58**(301), 13–30 (1963). https://doi.org/10.1080/01621459.1963.10500830

[17] Garrido-Jurado, S., et al.: Automatic generation and detection of highly reliable fiducial markers under occlusion. *Pattern Recognition* **47**(6), 2280–2292 (2014). https://doi.org/10.1016/j.patcog.2014.01.005

[18] Saltelli, A., et al.: Five ways to ensure that models serve society: a manifesto. *Nature* **582**, 482–484 (2020). https://doi.org/10.1038/d41586-020-01812-9

Statements and Declarations

Funding. The authors declare that no funds, grants, or other financial support were received for this work.

Competing interests. The authors declare that they have no relevant financial or non-financial competing interests.

Author contributions. Aarav Jalan developed the software and code, simulator implementation, computational experiments, and statistical analysis. Darsh Zaveri led the hardware design and construction, mechanical and electronics integration, instrumentation, and physical testing.

Data and code availability. Source code and processed research data are available at https://github.com/Aarav-J-10/Starlight-Ranking-Certificates. Released discordant-event counts enable recalculation of the principal statistical intervals. The current public release does not include every raw physical sensor stream or an independently verified end-to-end regeneration of all simulated missions.

Ethics approval and consent. The reported experiments involved computational models and robotic equipment, without human participants or animals. Research ethics approval, consent to participate, and consent for publication of participant data are therefore not applicable.