Studies in Knowledge and Innovation
Vol. 2 No. 4A (2026)

The same endpoints, a different goal: reference alignment in a released continuing PointMaze collection

Eric Gardner[1] · Owen Brooks[2] · Evan Anderson[3]

[1] Netherlands · [2] Sweden · [3] Finland

Published: April 30, 2026 | Article 844

Abstract

A stored navigation segment can begin with an observation inherited from an earlier goal. We audited one locked Minari PointMaze release containing 1,000,000 transitions and 13,210 segments, then compared endpoint distance reduction relative to the initial goal and the stable post-first-transition goal. Both rules used identical physical endpoints. All 13,210 supported segments were retained. Mean approach changed from -1.082 to 1.230 simulation coordinate units, a paired difference of 2.312; ten fixed record-order block deletions gave [2.304, 2.326], a deterministic range rather than a confidence interval. Sign classifications differed for 11,915 segments. Including observation0 in an any-success summary produced 5 disagreements with post-action support. The results quantify reference-dependent measurement in this released record, without estimating policy improvement, independent trial success, or real-world robot performance. Keywords: offline trajectories; goal reference; episode alignment; PointMaze; measurement; truncation

The same endpoints, a different goal: reference alignment in a released continuing PointMaze collection

Abstract

A stored navigation segment can begin with an observation inherited from an earlier goal. We audited one locked Minari PointMaze release containing 1,000,000 transitions and 13,210 segments, then compared endpoint distance reduction relative to the initial goal and the stable post-first-transition goal. Both rules used identical physical endpoints. All 13,210 supported segments were retained. Mean approach changed from -1.082 to 1.230 simulation coordinate units, a paired difference of 2.312; ten fixed record-order block deletions gave [2.304, 2.326], a deterministic range rather than a confidence interval. Sign classifications differed for 11,915 segments. Including observation0 in an any-success summary produced 5 disagreements with post-action support. The results quantify reference-dependent measurement in this released record, without estimating policy improvement, independent trial success, or real-world robot performance.

Keywords: offline trajectories; goal reference; episode alignment; PointMaze; measurement; truncation

1. Introduction

Offline reinforcement learning makes previously collected transitions reusable, but a reusable numerical file still requires an explicit account of what each row and boundary means. D4RL deliberately includes datasets collected by hand-designed controllers and data spanning several tasks in one environment, rather than only partially trained single-task agents (Fu et al., 2020). The broader offline-learning problem concerns learning or evaluating decisions from fixed experience, with limitations arising from coverage and distribution shift (Levine et al., 2020). Before either activity, a descriptive statement such as “the agent approached its goal during this episode” needs a goal reference and a temporal unit. This paper examines that prior measurement question in a public simulated PointMaze collection. It does not train or compare policies.

Goal identity is a substantive argument of a task description. Kaelbling (1993) studies learning to achieve dynamically changing goals, while universal value function approximators explicitly condition value on both state and goal (Schaul et al., 2015). Hindsight experience replay deliberately changes the goal assigned to stored experience to support learning (Andrychowicz et al., 2017). These methods do not establish our endpoint statistic, but they make clear why a position trajectory alone is insufficient to identify a goal-directed quantity. Two analyses can read the same positions correctly and still describe different progress if they retain different goal vectors.

Sequential-decision formulations distinguish a state sequence, its actions, its rewards, and the task being optimized (Puterman, 1994; Sutton and Barto, 2018). Stored episode names do not by themselves specify independently initiated trials. Likewise, options are temporally extended policies with defined initiation and stopping structure, rather than simply arbitrary file groups (Sutton et al., 1999). Pardo et al. (2018) distinguish finite-horizon tasks from interruptions of continuing experience. Gymnasium's separate termination and truncation fields express a related distinction at the interface level (Farama Foundation, n.d.-b). Neither field can be translated into task failure without inspecting the particular collector and environment.

This need for an explicit task definition also appears in applied policy studies. PPO is an optimization method based on sampled interaction and a surrogate objective (Schulman et al., 2017). In the distinct fixed-wing formation setting, Xu et al. (2023) specify the system state and transition tuple, define

destination arrival as mission completion, and describe other stopping causes. That work provides methodological context for keeping state, task completion, and stopping reason explicit. The present data are simulated point navigation, not observations of that aircraft system or a validation of its algorithm.

We ask how much endpoint approach changes when a segment's inherited initial goal is replaced by its recorded stable goal after the first transition. The same initial and final physical positions are used in both calculations. A secondary audit checks whether success information at the inherited observation is distinguishable from success after actions within the segment. The contribution is a quantified contrast between two measurement rules, accompanied by actual boundary and support checks. It concerns neither an undocumented data defect nor a new goal-reaching policy.

2. Release identity and recorded units

The source is the official Minari release D4RL/pointmaze/umaze-v2, identified in its metadata with Rodrigo Perez-Vicente as dataset author (Perez-Vicente, n.d.). We locked the Farama repository revision a0fd465d1d9d253b075d532b270e03d8fb4f46e6 and its main_data.hdf5 asset. The complete 556,229,176-byte file matched the provider's SHA256, ccb24a753bcaf56126a40cd749b2a6ed7e0cd22d301dee4f6e1367bd9b11efe1. The source contains 13,210 numerically ordered groups and 1,000,000 action transitions. This Minari reproduction is not asserted to contain identical rows to the original D4RL Maze2D archive. We used this one asset and did not combine versions.

The collection metadata names PointMaze_UMaze-v3, continuing_task=true, reset_target=true, and a collection time limit of 1,000,000 steps. Its separate evaluation environment uses a different goal-reset setting and a 300-step limit; those evaluation settings are not assigned retrospectively to the collection. The official creation tutorial describes random-goal navigation and a callback that marks a stored truncation when success is reported (Farama Foundation, n.d.-a). A standard environment interface facilitates interoperable storage, but does not make every field application-independent (Towers et al., 2024). We therefore call the stored groups segments of a continuing collection when interpreting their dependence.

For a segment with T actions, actions, rewards, terminations, and truncations have T entries. Achieved goals, desired goals, physical-state observations, and information arrays have T+1 entries. Action and reward index k correspond to the post-action observation and information at k+1. Observation0 has no action or reward belonging to the current segment. The two-dimensional achieved goal is the recorded position; the four-dimensional observation adds two velocity components. Distance is reported in simulation coordinate units. No physical field calibration or real-robot scale is inferred from the arrays.

We inspected each HDF5 group's actual dimensions and numerical types, rather than trusting only a wrapper's episode count. HDF5 groups, datasets, dataspaces, and attributes organize stored information; their application meaning remains the producer's mapping (The HDF Group, n.d.). The inventory preserves the companion metadata and embedded attributes separately. The embedded identifier is the legacy pointmaze-umaze-v2, and its minari_version and dataset_size text differ from the companion's 0.4.3 and size text. Actual bytes and the verified provider hash identify the analyzed object. We do not infer why those metadata differ.

The metadata specifies gymnasium-robotics>=1.2.4, rather than an exact historical executable version or collection date. Current source and a tagged source version were read as explanatory text; neither was executed. The metadata's generator link resolves to a current repository, not a frozen historical run. Consequently, current code helps explain the observed convention but is not evidence that an exact

current line produced every historical row. Collection, composition, and use restrictions are recorded separately, following the documentation concerns raised by Gebru et al. (2021).

3. Fixed comparison and support

The full protocol was saved after a nonrandom five-segment schema probe and before complete-file effects were calculated. The first three and last two segments established that the proposed fields existed; they were not a sample for estimating prevalence. The analysis universe remained all 13,210 groups. No segment was selected by reward, duration, speed, sign, or the apparent size of its effect. Missing or malformed groups would remain in the inventory with an explicit unsupported reason. We did not pad arrays, replace finite negative values, interpolate positions, or repair a boundary.

Let p0 and pT denote the initial and final recorded positions, and let g0 through gT be the desired-goal vectors. The primary support requires T≥1, finite endpoints and goals, and exact equality g1=…=gT. It permits either no goal change or a change only at observation1. A change later within the segment would prevent the single stable-reference interpretation and would remain separately counted. Goal changes are detected by exact vector comparison, not by a tolerance chosen after viewing their magnitudes. This keeps support distinct from a success-radius rule.

The initial-reference approach and aligned-reference approach are defined by

$$A_{\text{initial}} = \|p_0 - g_0\|_2 - \|p_T - g_0\|_2,$$

$$A_{\text{aligned}} = \|p_0 - g_1\|_2 - \|p_T - g_1\|_2, \quad D = A_{\text{aligned}} - A_{\text{initial}}.$$

Positive approach means that the final endpoint is closer to the named, held-fixed goal. Both rules span exactly the same physical endpoints. The aligned rule is retrospective: the first post-action recorded goal is used to define a descriptive contrast on stable-goal support. It is not information available to an initial-time forecasting algorithm, and it is not a reconstruction of an unrecorded controller intention. The two rules are calculated for every supported segment, never chosen according to which gives the more favorable number.

Naming the reference is part of defining the quantity, rather than a cosmetic labeling decision. Measurement scales restrict what numerical comparisons mean (Stevens, 1946), and operational procedures determine the quantities that a numerical analysis actually measures (Hand, 2004). The distinction between a construct and its operational measurement is also central to Jacobs and Wallach (2021). Here the construct “successful navigation” is broader than either a distance reduction or a stored Boolean. We therefore retain the concrete names of our measurements rather than treating one as independently verified true success.

Straight-line endpoint distance does not measure obstacle-respecting route length or path optimality. Planning distinguishes feasible trajectories and costs that depend on their complete route (LaValle, 2006). Moreover, state and reference-tracking concepts in feedback systems make position only one component of behavior; position near a goal alone does not establish a settled state (Åström and Murray, 2008). Endpoint velocity is consequently described without imposing a hovering or stability criterion. Although fixed-goal distance resembles a potential, this paper introduces no shaped reward and invokes no policy-invariance guarantee of the kind studied by Ng et al. (1999).

The primary effect is the equal-segment mean of D over the declared support. We also retain medians, quartiles, extrema, and the complete negative/zero/positive cross-tabulation of the two approach values. Different-sign classifications use exact signs of the saved numbers. These are properties of a finite

released record, not estimates from independent training seeds. Empirical RL design requires the statistical unit and source of variation to be explicit (Patterson et al., 2024). Treating every inherited segment as a fresh independent experiment would not meet that requirement.

To describe record-order sensitivity, we assigned each segment to one of ten fixed blocks using floor((cumulative ending transition-1)/100000). We omitted each block in turn and recomputed mean D and the different-sign fraction. The range of those ten results is a deterministic deletion range. It is neither a confidence interval nor a claim that the ten blocks are independently sampled. There was no bootstrap, hypothesis test, tuned threshold, fitted prediction surface, or new policy rollout. Segment-level extrema are reported separately because a range across individual effects answers a different question from a range across aggregate deletions.

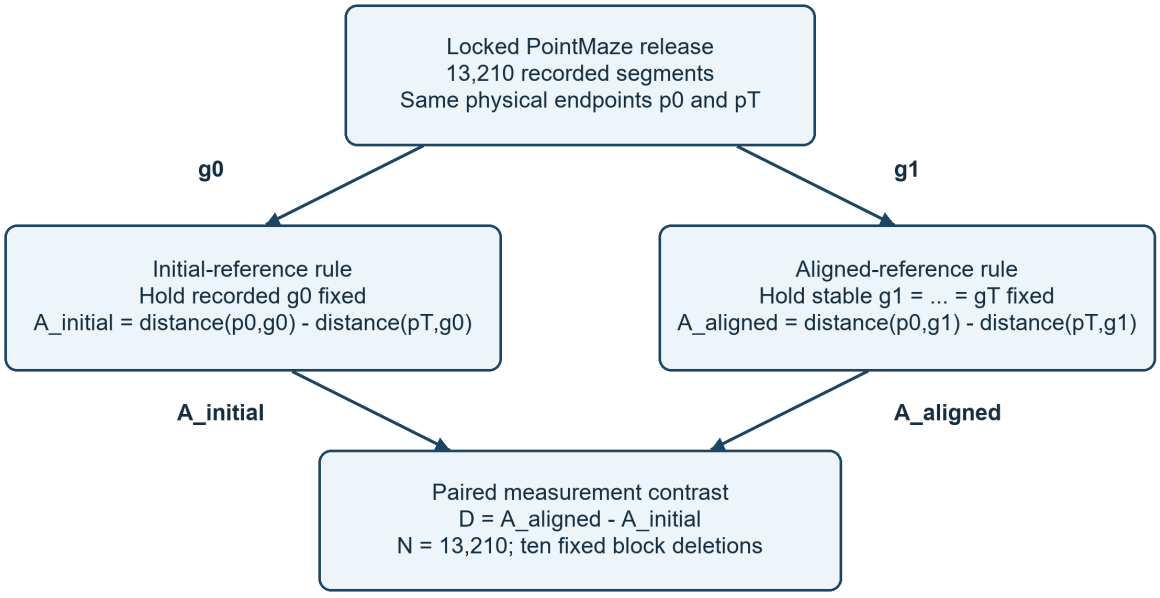


Figure 1. Research measurement model. The same endpoints enter both reference rules. Stable post-first-observation goals define support; arrows denote calculations, not causal effects. Deletion ranges describe the recorded collection only.

4. Complete-file results

All 13,210 segments passed the declared array-shape checks, and all 13,210 met primary geometric support; 0 remained unsupported. The exact goal-change inventory contained 5 no-change segments, 13,205 segments changing only at observation1, and 0 segments with other change patterns. Thus the all-file result is based on actual full-source support, rather than extrapolation from the five pilot cases. The primary columns contained no nonfinite observation, goal, action, or reward values. The recorded information goal and position/velocity arrays agreed with their corresponding observation fields in every shape-supported segment.

Mean approach under the inherited initial goal was -1.082229 coordinate units. Using the stable post-first-transition goal on the same endpoints gave 1.229766, yielding mean D=2.311995. The median difference was 2.641752, with quartiles 1.008821 and 3.444396. Individual differences ranged from -0.551092 to 5.306029, retaining changes in both directions. The effect is a change in the interpretation of existing movement, not additional movement by the controller. Table 1 separates the all-file contrast from

its goal-change strata, so a zero difference caused by identical references is not mistaken for an estimated null effect of navigation.

Table 1. Paired endpoint approach under two fixed reference rules. All values are simulation coordinate units. N is the number of supported stored segments. Observed ranges summarize segment effects and are not confidence intervals.

Support	N	Mean A_initial	Mean A_aligned	Mean D	Median D	Observed D range
All supported	13210	-1.082229	1.229766	2.311995	2.641752	[-0.551092, 5.306029]
No goal change	5	1.290137	1.290137	0.000000	0.000000	[0.000000, 0.000000]
Goal switch at observation1	13205	-1.083127	1.229743	2.312870	2.642220	[-0.551092, 5.306029]

The two measurements had different sign classes in 11,915 segments (90.1968%). The full nine-cell table retains exact zeros rather than merging them with either direction. A negative initial-reference approach and a positive aligned-reference approach can coexist because the two goal vectors differ. It would be incorrect to interpret that pair as evidence that one recorded physical path changed direction merely because the analysis changed its reference. Figure 2 shows both signed measurements with common axis limits and an identity line.

Table 2. Complete sign cross-tabulation for the 13,210 primary-supported segments. Rows use g0 and columns use g1. Counts are finite-record descriptions.

Initial-goal sign	Aligned negative	Aligned zero	Aligned positive
negative	1	0	11913
zero	0	0	0
positive	2	0	1294

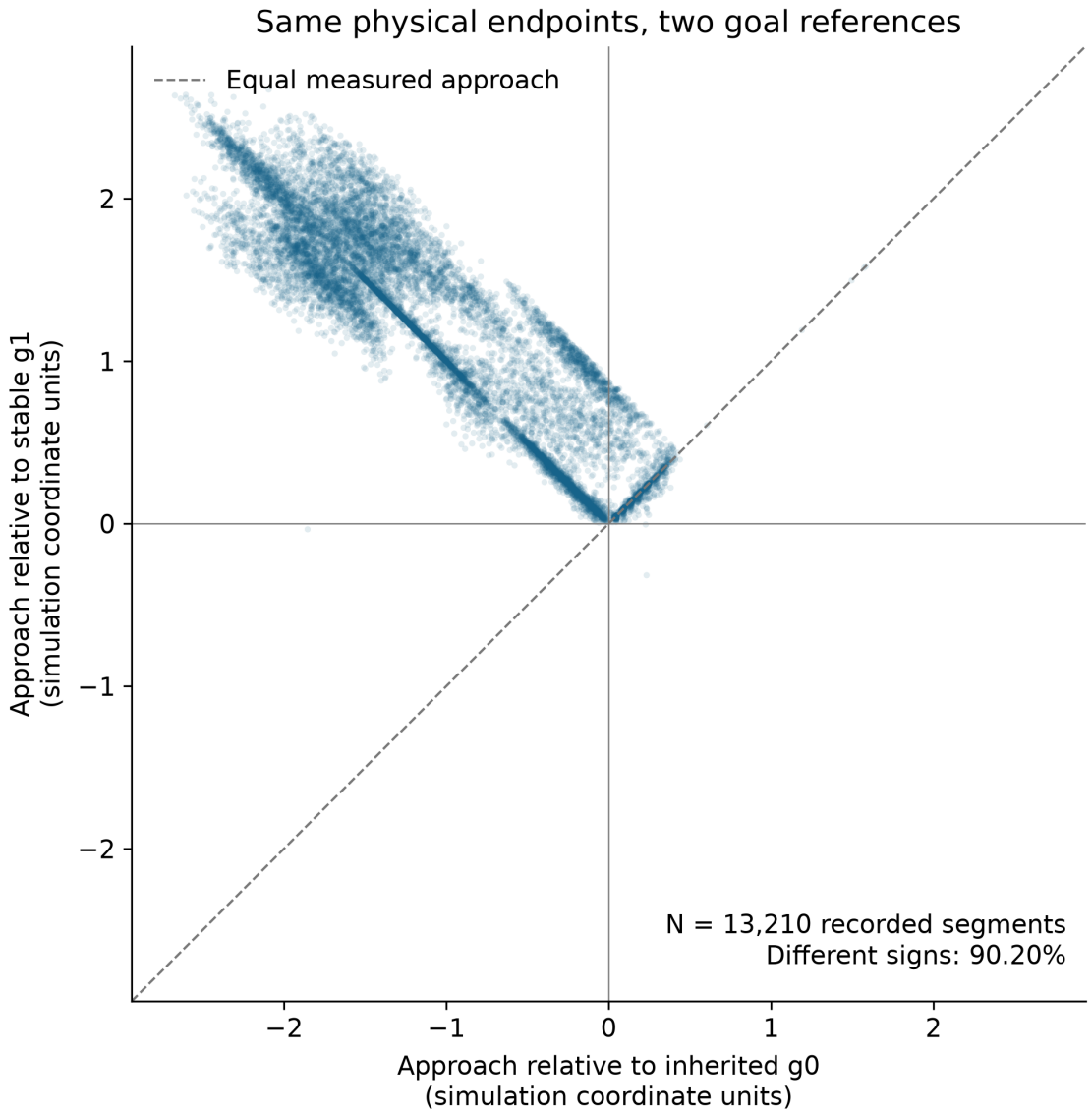


Figure 2. Every supported segment contributes the same two physical endpoints to this paired plot. Axes use simulation coordinate units and identical limits. Transparency indicates overlap; the diagonal marks equal measured approach, not an expected policy-performance benchmark.

Omitting one fixed transition block at a time gave mean differences from 2.304432 to 2.325915. The corresponding different-sign fractions ranged from 90.0579% to 90.3376%. Table 3 keeps those aggregate ranges distinct from Table 1's individual-effect range. These calculations show whether the reported aggregate is concentrated in one part of the released record. They do not establish performance variability under new collection seeds or new goal schedules.

Table 3. Actual paired effect and deterministic record-order sensitivity, N=13,210 before deletion. Each deletion removes segments whose ending transition falls in one predeclared 100,000-transition block. The ten resulting values form a range, not a sampling interval.

Measure	Full supported file	Ten block-deletion range; not CI
Mean paired difference D	2.311995	[2.304432, 2.325915]
Different-sign segment fraction	90.196821%	[90.057891%, 90.337592%]

5. Inherited information and segment boundaries

The recorded success summaries were not interchangeable at every boundary. Any success across indices0…T was true for 13,210 segments, whereas any success at indices1…T was true for 13,205. Their disagreement count was 5, with segment IDs 2654, 5312, 7902, 10584, 13209. Final information success was true for 13,205, and at least one positive reward was present in 13,205. Across all 1,000,000 transitions, positive reward and the aligned post-action success field disagreed 0 times. These are direct comparisons of stored fields, not an independently observed success outcome.

Actual final-field combinations were: 5 with final success=false, last reward=0, termination=false, truncation=true; 13,205 with final success=true, last reward=1, termination=false, truncation=true. A truncation label therefore does not, on its own, identify failure. All five segments lacking post-action arrival are retained, including the four interior cases; their cumulative ending transition indices are 200,000, 400,000, 600,000, 800,000, 1,000,000. They are unfinished goal-directed records at those stored endings, not failed policy trials. The million-step collection setting is documented, but the flags alone do not identify whether an interior ending arose from a collector flush, a collection subdivision, or another historical mechanism. A filename or boundary Boolean cannot supply that unobserved stopping cause.

All 13,209 consecutive boundaries were checked. Previous final position equaled next initial position exactly in 13,205 comparisons; the largest absolute component difference was 2.20303. Velocity and goal exact-match counts were 13,205 and 13,205; success matched in 13,209. Four boundaries were physically discontinuous: 2654→2655; 5312→5313; 7902→7903; 10584→10585. These occur after the four interior unfinished endings. They partition the file into five observed position-continuity blocks of 200,000 transitions each. Their first groups carry recorded seed attributes 123, 124, 125, 126, 127, respectively; the other 13,205 groups have no seed attribute. We preserve those labels without treating them as proof of statistical independence or a complete historical reset mechanism. The matched boundaries directly support inherited initial observations within blocks. At every segment, its post-action goal is still checked to identify the stable reference; no discontinuity is concatenated or repaired.

As source compatibility checks, the current implementation's fixed 0.45 distance comparison disagreed with the recorded success array 0 times. Some accompanying prose uses a different radius, and the precise historical executable is not supplied. We do not tune a radius or redefine success to force agreement. Neither radius enters the primary approach contrast. Endpoint distances and velocities remain supporting descriptions, and no speed threshold is used to select segments or assert stable arrival.

6. Implications and limits

The principal finding concerns reference assignment. A stored observation can simultaneously be the endpoint of one goal-directed segment and the starting physical state of the next. Keeping its old desired goal while describing subsequent motion answers an old-goal distance question. Holding the stable post-first-transition goal fixed answers a different question about the same movement. The observed paired difference quantifies that distinction in this release. It does not show that the initial observation is malformed, that the environment should reset, or that one goal definition is universally appropriate for every downstream use.

A practical reuse report should name the segment support, the goal index or goal-update rule, and whether initial information is included in a Boolean summary. If a user needs transition rewards, the post-action information index must be aligned accordingly. If a user needs task-level outcomes, a stored segment ending is insufficient without the collector's task definition. These are interpretive requirements supported

by the current record and its source documentation. They do not justify replacing every external truncation by termination, rejecting all inherited observations, or assigning an unrecorded failure label.

The paired comparison is deliberately narrow. It does not identify behavior under an alternative goal policy because no actions were regenerated. It does not evaluate an offline learner, estimate a policy value, or correct distribution shift. Nor does the aligned goal necessarily coincide with every internal waypoint remembered by the collection controller. Euclidean distance between endpoints ignores the route through the maze, collisions, action effort, and elapsed physical time. Recorded seed labels do not establish how independently the collection blocks were generated, and precise acquisition dates are not supplied. No claim about general navigation accuracy, communication quality, real-world safety, or settling performance follows from the magnitude of D.

The complete file is more informative than the five initial examples, but completeness within an asset is not representativeness beyond that asset. Equal weight per stored segment is the predeclared descriptive convention. Different group lengths mean it does not equal a transition-weighted average. We do not search alternative weightings for a preferred conclusion. The ten contiguous-block deletions retain the collection's order while revealing finite-record concentration; they supply no probability coverage. A new independent collection would be required to quantify between-run variation or to test whether the same reference effect persists under other maps and goal-reset policies.

7. Reproducibility and conclusion

The reproducibility package records the pinned source identity, complete provider-hash validation, frozen protocol, one-row-per-segment processed table, calculation scripts, figures, and actual-byte CSV hash. A separate scalar scan uses tuple comparisons and scalar hypotenuse calculations rather than importing the primary analysis, and checks every segment, field summary, sign cell, block deletion, and boundary. This supplies executable evidence for the published quantities. Reproducing a calculation from an available object is distinct from independently replicating the original collection, a distinction emphasized by Peng (2011). Sandve et al. (2013) motivate preserving the computational record, while Wilson et al. (2017) provide practical organization principles for keeping original and processed files distinguishable.

Availability also has a specific scope. Artifact access alone need not make a computational result reproducible, as the empirical work of Stodden et al. (2018) demonstrates; the environment and transformation record remain necessary. Stodden and Miguez (2014) discuss infrastructure supporting reproducible and extensible computation, motivating the retained reader-version and workflow records here. FAIR reuse includes provenance and clear data-usage terms (Wilkinson et al., 2016). We found no explicit dataset licence in the inspected release metadata and do not presume that a software licence covers the trajectory bytes. This study uses the official public research release locally and does not redistribute the original HDF5 in a new data archive. Source access dates and repository modification dates are not presented as collection dates.

For this locked PointMaze record, changing only the held-fixed goal reference changes measured endpoint approach by 2.312 coordinate units on average and changes the sign class for 90.20% of supported segments. Complete boundary and field checks explain why an inherited observation deserves a different temporal role from a post-action observation. The result supports explicit reference and indexing statements whenever continuing navigation data are reused for descriptive task analysis.

References

1. Ng, Andrew Y.; Harada, Daishi; Russell, Stuart. (1999). Policy Invariance under Reward Transformations: Theory and Application to Reward Shaping. Proceedings of the Sixteenth International Conference on Machine Learning, pp. 278-287. https://people.eecs.berkeley.edu/~pabbeel/cs287-fa09/readings/NgHaradaRussell-shaping-ICML1999.pdf
2. Gebru, Timnit; Morgenstern, Jamie; Vecchione, Briana; Vaughan, Jennifer Wortman; Wallach, Hanna; Daumé III, Hal; Crawford, Kate. (2021). Datasheets for datasets. Communications of the ACM. https://doi.org/10.1145/3458723
3. Sutton, Richard S., and Andrew G. Barto. *Reinforcement Learning: An Introduction*. 2nd ed., MIT Press, 2018. http://incompleteideas.net/book/the-book-2nd.html.
4. Åström, Karl J.; Murray, Richard M.. (2008). Feedback Systems: An Introduction for Scientists and Engineers. Princeton University Press. https://www.cds.caltech.edu/~murray/FBS/
5. Farama Foundation. (n.d.-a). Minari documentation: episode data and PointMaze dataset creation. Official software documentation. https://minari.farama.org/tutorials/dataset_creation/point_maze_dataset/
6. Xu, Dan; Guo, Yunxiao; Yu, Zhongyi; Wang, Zhenfeng; Lan, Rongze; Zhao, Runhao; Xie, Xinjia; Long, Han. (2023). PPO-Exp: Keeping Fixed-Wing UAV Formation with Deep Reinforcement Learning. Drones. https://doi.org/10.3390/drones7010028
7. Schulman, John; Wolski, Filip; Dhariwal, Prafulla; Radford, Alec; Klimov, Oleg. (2017). Proximal Policy Optimization Algorithms. arXiv preprint. https://arxiv.org/abs/1707.06347
8. Pardo, Fabio; Tavakoli, Arash; Levdik, Vitaly; Kormushev, Petar. (2018). Time Limits in Reinforcement Learning. International Conference on Machine Learning. https://proceedings.mlr.press/v80/pardo18a.html
9. Hand, David J.. (2004). Measurement Theory and Practice: The World Through Quantification. Wiley. https://www.wiley-vch.de/de/fachgebiete/mathematik-und-statistik/measurement-theory-and-practice-978-0-470-68567-9
10. The HDF Group. (n.d.). HDF5 User Guide: The HDF5 Data Model and File Structure. Official HDF5 documentation. https://support.hdfgroup.org/documentation/hdf5/latest/_h5_d_m__u_g.html
11. Jacobs, Abigail Z.; Wallach, Hanna. (2021). Measurement and Fairness. Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency. https://doi.org/10.1145/3442188.3445901
12. Towers, Mark; Kwiatkowski, Ariel; Terry, Jordan; Balis, John U.; De Cola, Gianluca; Deleu, Tristan; Goulão, Manuel; Kallinteris, Andreas; Krimmel, Markus; KG, Arjun; Perez-Vicente, Rodrigo; Pierré, Andrea; Schulhoff, Sander; Tai, Jun Jet; Tan, Hannah; Younis, Omar G.. (2024). Gymnasium: A Standard Interface for Reinforcement Learning Environments. arXiv preprint. https://arxiv.org/abs/2407.17032
13. Wilson, Greg; Bryan, Jennifer; Cranston, Karen; Kitzes, Justin; Nederbragt, Lex; Teal, Tracy K.. (2017). Good enough practices in scientific computing. PLOS Computational Biology. https://doi.org/10.1371/journal.pcbi.1005510
14. Perez-Vicente, Rodrigo. (n.d.). D4RL/pointmaze/umaze-v2 [dataset]. Farama Minari public dataset repository, pinned release a0fd465d1d9d253b075d532b270e03d8fb4f46e6. https://huggingface.co/datasets/farama-minari/D4RL/tree/a0fd465d1d9d253b075d532b270e03d8fb4f46e6/pointmaze/umaze-v2
15. Farama Foundation. (n.d.-b). Gymnasium documentation: handling time limits and the terminated/truncated step API. Official software documentation. https://gymnasium.farama.org/tutorials/gymnasium_basics/handling_time_limits/
16. Fu, Justin; Kumar, Aviral; Nachum, Ofir; Tucker, George; Levine, Sergey. (2020). D4RL: Datasets for Deep Data-Driven Reinforcement Learning. arXiv preprint. https://arxiv.org/abs/2004.07219
17. Peng, Roger D.. (2011). Reproducible Research in Computational Science. Science. https://doi.org/10.1126/science.1213847
18. Sandve, Geir Kjetil; Nekrutenko, Anton; Taylor, James; Hovig, Eivind. (2013). Ten Simple Rules for Reproducible Computational Research. PLoS Computational Biology. https://doi.org/10.1371/journal.pcbi.1003285
19. Patterson, Andrew; Neumann, Samuel; White, Martha; White, Adam. (2024). Empirical Design in Reinforcement Learning. Journal of Machine Learning Research. https://www.jmlr.org/papers/v25/23-0183.html
20. LaValle, Steven M.. (2006). Planning Algorithms. Cambridge University Press. https://lavalle.pl/planning/
21. Kaelbling, Leslie Pack. (1993). Learning to Achieve Goals. Proceedings of the Thirteenth International Joint Conference on Artificial Intelligence, pp. 1094-1099. https://www.ijcai.org/proceedings/1993/151
22. Levine, Sergey; Kumar, Aviral; Tucker, George; Fu, Justin. (2020). Offline Reinforcement Learning: Tutorial, Review, and Perspectives on Open Problems. arXiv preprint. https://arxiv.org/abs/2005.01643

23. Sutton, Richard S.; Precup, Doina; Singh, Satinder. (1999). Between MDPs and semi-MDPs: A framework for temporal abstraction in reinforcement learning. Artificial Intelligence. https://doi.org/10.1016/s0004-3702(99)00052-1

24. Puterman, Martin L.. (1994). Markov Decision Processes: Discrete Stochastic Dynamic Programming. Wiley Series in Probability and Statistics. https://doi.org/10.1002/9780470316887

25. Schaul, Tom; Horgan, Daniel; Gregor, Karol; Silver, David. (2015). Universal Value Function Approximators. International Conference on Machine Learning. https://proceedings.mlr.press/v37/schaul15.html

26. Stodden, Victoria; Seiler, Jennifer; Ma, Zhaokun. (2018). An empirical analysis of journal policy effectiveness for computational reproducibility. Proceedings of the National Academy of Sciences. https://doi.org/10.1073/pnas.1708290115

27. Wilkinson, Mark D.; Dumontier, Michel; Aalbersberg, IJsbrand Jan; Appleton, Gabrielle; Axton, Myles; Baak, Arie; Blomberg, Niklas; Boiten, Jan-Willem; da Silva Santos, Luiz Bonino; Bourne, Philip E.; Bouwman, Jildau; Brookes, Anthony J.; Clark, Tim; Crosas, Mercè; Dillo, Ingrid; Dumon, Olivier; Edmunds, Scott; Evelo, Chris T.; Finkers, Richard; Gonzalez-Beltran, Alejandra; Gray, Alasdair J.G.; Groth, Paul; Goble, Carole; Grethe, Jeffrey S.; Heringa, Jaap; 't Hoen, Peter A.C; Hooft, Rob; Kuhn, Tobias; Kok, Ruben; Kok, Joost; Lusher, Scott J.; Martone, Maryann E.; Mons, Albert; Packer, Abel L.; Persson, Bengt; Rocca-Serra, Philippe; Roos, Marco; van Schaik, Rene; Sansone, Susanna-Assunta; Schultes, Erik; Sengstag, Thierry; Slater, Ted; Strawn, George; Swertz, Morris A.; Thompson, Mark; van der Lei, Johan; van Mulligen, Erik; Velterop, Jan; Waagmeester, Andra; Wittenburg, Peter; Wolstencroft, Katherine; Zhao, Jun; Mons, Barend. (2016). The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data. https://doi.org/10.1038/sdata.2016.18

28. Stevens, S. S.. (1946). On the Theory of Scales of Measurement. Science. https://doi.org/10.1126/science.103.2684.677.

29. Andrychowicz, Marcin; Wolski, Filip; Ray, Alex; Schneider, Jonas; Fong, Rachel; Welinder, Peter; McGrew, Bob; Tobin, Josh; Abbeel, Pieter; Zaremba, Wojciech. (2017). Hindsight Experience Replay. Advances in Neural Information Processing Systems. https://proceedings.neurips.cc/paper/2017/hash/453fadbd8a1a3af50a9df4df899537b5-Abstract.html

30. Stodden, Victoria; Miguez, Sheila. (2014). Best Practices for Computational Science: Software Infrastructure and Environments for Reproducible and Extensible Research. Journal of Open Research Software. https://doi.org/10.5334/jors.ay