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

Paying for Shared Policies: Last Affordable Overcooked Evaluations under Two Source-Asset Boundaries

Joseph Hayes[1] · David Clark[2] · Matthew Fisher[3]

[1] Croatia · [2] Estonia · [3] Latvia

Published: April 30, 2026 | Article 839

Abstract

A saved source policy has an acquisition history even when its next use requires no retraining. We examine how allocating that history changes the published evaluation affordable under a fixed interaction budget. The MEDoE Overcooked archive supplies 49,776 records and already includes source training in its transfer-method horizontal axis. After removing that existing offset, we apply two explicit capital-cost boundaries and five counterfactual sharing rules. At a ten-million-unit budget, selected-pair accounting yields MEDoE means of 0.9935–0.9983, whereas charging the complete source library without sufficient sharing leaves no affordable evaluation. Across the fixed grid, 118 MEDoE–QMIX comparisons are available: 114 differences are positive and four negative. Their scope is a finite archive, with deterministic record ranges rather than independent-seed confidence intervals. Cost allocation changes accessible evidence; it supplies neither new policy outcomes nor observed downstream reuse. Keywords: multi-agent reinforcement learning; source-policy reuse; budget allocation; published evaluations; archive support

Paying for Shared Policies: Last Affordable Overcooked Evaluations under Two Source-Asset Boundaries

Abstract

A saved source policy has an acquisition history even when its next use requires no retraining. We examine how allocating that history changes the published evaluation affordable under a fixed interaction budget. The MEDoE Overcooked archive supplies 49,776 records and already includes source training in its transfer-method horizontal axis. After removing that existing offset, we apply two explicit capital-cost boundaries and five counterfactual sharing rules. At a ten-million-unit budget, selected-pair accounting yields MEDoE means of 0.9935-0.9983, whereas charging the complete source library without sufficient sharing leaves no affordable evaluation. Across the fixed grid, 118 MEDoE-QMIX comparisons are available: 114 differences are positive and four negative. Their scope is a finite archive, with deterministic record ranges rather than independent-seed confidence intervals. Cost allocation changes accessible evidence; it supplies neither new policy outcomes nor observed downstream reuse.

Keywords: multi-agent reinforcement learning; source-policy reuse; budget allocation; published evaluations; archive support

1. The acquisition history of a reusable policy

A pretrained policy can enter another training job without repeating the interactions that produced it. The accounting question then becomes which acquisition history belongs to that job. Charging a particular pair of source teams, charging an entire policy library, and dividing either investment among several uses yield different resource coordinates for the same recorded return. These rules can be specified without changing a policy, but their consequences for an empirical comparison still require actual performance records and defensible source-cost evidence.

MEDoE provides a concrete setting. Fosong et al. (2024a) train subteams on simpler source tasks and combine their policies for target-task fine-tuning. Their public data release includes prepared Overcooked results and saved source-agent metadata (Fosong et al., 2024b). Crucially, the original paper already defines source-plus-target training steps, and its Figure 3 shifts transfer-method curves to include source training. Our question concerns a different allocation scenario: **which last published evaluation is affordable at a fixed budget when a declared source asset is shared among a specified number of cost-bearing uses?** There is no alleged omission in the original figure.

Transfer-learning research distinguishes what knowledge is moved and how source and target tasks relate (Taylor and Stone, 2009; Pan and Yang, 2010). Curriculum-learning frameworks additionally make the sequence or graph of tasks part of the design (Narvekar et al., 2020). Policy reuse is an established learning mechanism, including probabilistic guidance from previously learned policies (Fernández and Veloso, 2006). These ideas motivate retaining the cost history of a source asset. They do not determine a unique way to distribute that cost among uses.

The reusable object also matters. Progressive neural networks retain previously learned features through lateral connections (Rusu et al., 2016), while behavior-prior methods model reusable patterns across related contexts (Tirumala et al., 2022). MEDoE uses its own source-policy and expertise machinery. Another organization appears in evolutionary multi-agent reinforcement learning: the formal flocking chapter by Guo et al. (2023) combines cooperation and competition, including stochastic inheritance of senior-agent parameters by junior agents. That mechanism gives a substantive example of learned parameters being resources within a team-learning process. Our accounting study implements no

evolutionary selection and makes no inference about flocking performance; it asks how an already acquired policy resource is charged.

Resource definitions need similar care. Green AI treats computational demands as evaluation concerns alongside predictive performance (Schwartz et al., 2020). The Efficiency Misnomer shows why different cost indicators cannot safely substitute for one another (Dehghani et al., 2022). Reporting experimental search and resource choices also helps readers understand what a result required (Dodge et al., 2019). We consequently use a narrowly defined nominal interaction account. Neither an environmental transition nor an optimizer update is automatically a unit of elapsed time, energy or floating-point computation.

2. Published records and saved source counters

The empirical input is the official Edinburgh DataShare release, available on 8 May 2024. Its complete ZIP has 1,355,908,730 bytes. A range-readable central directory allowed selective acquisition of the Overcooked results, three source indices, plotting text and small metadata files. Selected ZIP members passed their stored CRC32 checks, and local SHA256 hashes identify the exact inputs. Complete model weights and experience buffers were not acquired or deserialized. Dataset terms are CC BY 4.0; the associated code repository uses MIT terms.

The analysis preserves all 49,776 rows of the released cooking results table. Its numerical columns are the published training-step coordinate and the published team-return summary. The five method labels are IPPO, MAPPO, MEDoE, PS and QMIX. There are 816 records over 51 steps for each of MEDoE and PS, 96 records over six steps for each of IPPO and MAPPO, and 47,952 records over 2,997 steps for QMIX. Every method-specific step contains 16 records. All input numerical values are finite, and none is removed because its score is unfavorable.

The labels identify different algorithmic contexts. PPO uses a policy-gradient surrogate (Schulman et al., 2017), following work on trust-region policy optimization (Schulman et al., 2015). The MAPPO literature examines PPO in cooperative multi-agent games (Yu et al., 2022). QMIX instead uses a monotonic factorization of joint action values to support decentralized policies (Rashid et al., 2018). The numerical outcomes here come from the MEDoE authors' released Overcooked table, rather than from those algorithm papers' original experiments. PS is retained as its released label because an explicit machine-readable alias to the paper's named naive curriculum baseline was not established.

The released return scale is preserved unchanged. The plotting script directly displays the stored values. Applying another division by team size would therefore alter the published measurement. Return summarizes rewards accumulated through interaction (Sutton and Barto, 2018); a value near one here is not relabeled a success probability or win rate. More generally, scale interpretation determines which numerical statements are meaningful (Stevens, 1946). Slight floating-point values above one remain in the analysis and are rounded only for display.

Source costs are supported by saved metadata rather than inferred from intended training caps. All 16 Overcooked actor files record a source task, seed, original training partner, 3,000,000 training steps and 93,744 completed actor updates. Reciprocal partner links identify eight distinct co-trained source teams, four for each source task. Two actors trained together share one environmental training history, so counting each actor as another three-million-step run would double the investment.

Read-only inspection of the pinned author code explains the counters. The saved training-step value comes from a runtime vector-step counter multiplied by the number of parallel environments. The optimizer-update value is a separate accumulated counter. With 32 parallel environments, three million recorded transitions correspond to 93,750 vector steps. Sixteen-step rollouts and sixteen optimization

epochs yield 93,744 completed updates after the incomplete final rollout is accounted for. Adam supplies the gradient-based optimizer in this implementation (Kingma and Ba, 2015). Multiplying the actor-update count by 32 would conflate its iterations with environmental sampling.

We nevertheless call these author/metadata nominal interaction units. The release does not supply a complete account of hardware time, classifier fitting, evaluation computation or hyperparameter search for this calculation. The two capital boundaries are six million units for the selected left/right source-team pair and 24 million units for building all eight source teams. The second boundary represents a complete-library investment scenario; it does not claim that each target actually consumes all eight teams.

3. A rule for the last affordable evaluation

The protocol was frozen after source, format and support inspection, before calculating budget-grid comparisons. Some raw and endpoint values were visible during decoder and scale checks, so the preparation is not described as outcome-blinded. The fixed decisions are two capital boundaries, sharing counts of 1, 2, 4, 8 and 16, and fourteen budgets from two to 28 million in increments of two million. The ten-million budget was designated for the main numerical table before the contrasts were calculated.

Let x denote the released training-step coordinate. For MEDoE and PS, target fine-tuning time is t=x-6,000,000 because the source offset is already present. For source capital C and declared number of equal shares R, the alternative cost coordinate is

$$a_{C,R}(t) = t + C/R.$$

For each budget B, we choose the largest released target step satisfying a≤B. From-scratch methods retain their released step coordinate as cost. This is a last-observation rule, not a search for the largest recorded return. Selection by favorable observed performance can itself bias evaluation (Cawley and Talbot, 2010); our rule has no return value in its eligibility criterion.

R is an analyst-defined counterfactual allocation. It does not describe measured new tasks, claim that the source asset was actually reused R times, or assume that target performance would generalize to those tasks. Selected-pair R=1 reproduces the original six-million source charge. Complete-library R=16 allocates 24 million units over sixteen shares, charging 1.5 million per share. This is an explicit portfolio cost rule rather than a recovered deployment history.

Three propositions organize the calculation. P1 is definitional: increasing R weakly expands affordable target-step support at a fixed budget and capital boundary. P2 is empirical: selecting a later affordable recorded step need not increase its published mean return. P3 concerns allocation scope: charging the complete library can change available evidence and budget-matched contrasts relative to charging the selected pair. The model diagram in Figure 1 maps these operations to their measurements.

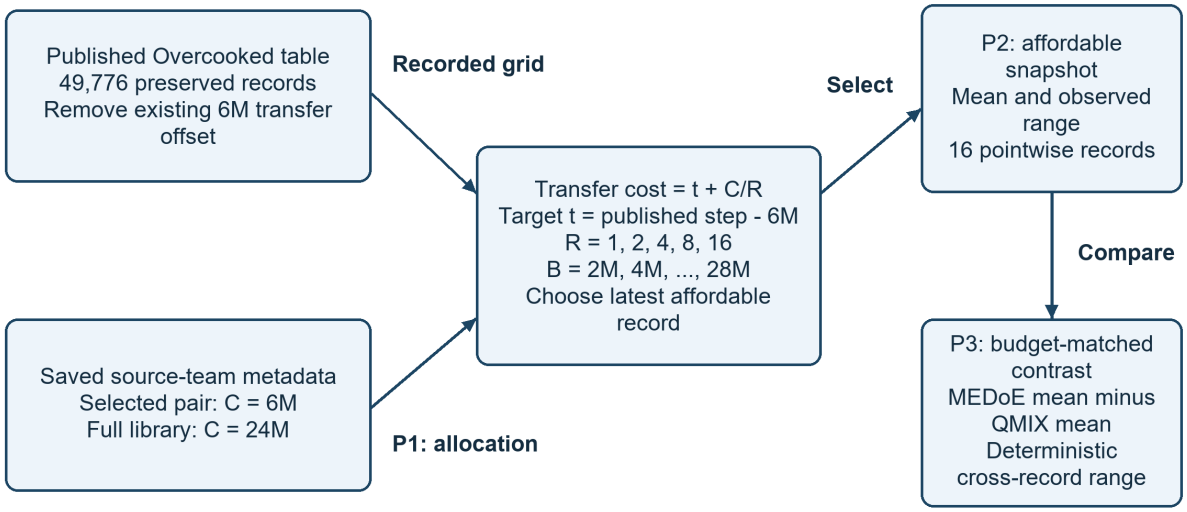


Figure 1. Recorded performance and source metadata enter an explicit cost-allocation and affordability rule. P1 concerns accessible recorded steps, P2 their actual return summaries, and P3 the resulting budget-matched contrast. The arrows show accounting operations, not causal changes in policies.

A budget below every eligible recorded cost produces an unavailable cell with zero selected records and missing performance. An observed zero return remains an observed value. When the budget exceeds the cost of the final archived point, the rule retains that actual point and flags archive exhaustion. Budget minus selected cost is reported as unallocated budget within this archive comparison. It combines evaluation-grid gaps with any budget extending beyond the archived endpoint; it is not measured wasted computation.

The source table limits statistical interpretation. Its ID index is unique per row within each method, rather than repeating as an authenticated seed or partner key. Fifteen step resets in source order suggest ordered segments but do not establish their identities. Reproducibility work and RL evaluation guidance emphasize the need to state experimental variation and sampling units (Henderson et al., 2018; Agarwal et al., 2021; Patterson et al., 2024). Benchmark variation can arise from several stages of a learning pipeline (Bouthillier et al., 2021). Those sources cannot be reconstructed from row order alone.

Accordingly, each affordable snapshot reports the mean and observed minimum-maximum range of its 16 released records. The principal contrast subtracts the QMIX mean at the same budget from the MEDoE mean. Its deterministic range is [minimum MEDoE-maximum QMIX, maximum MEDoE-minimum QMIX], spanning all 256 cross-record differences. It is not a confidence interval or an authenticated paired effect. Reusing a record across allocation scenarios supplies no additional independent experiment.

4. Affordable evidence across the fixed grid

The calculation produces 700 method-by-budget-by-allocation summaries. Of these, 656 have an affordable snapshot and 44 are unavailable. Each available snapshot uses exactly 16 released records. All 700 eligibility checks verified that the selected point was affordable and that no later released point was affordable. In addition, 210 checks confirmed identical selections when different capital/share combinations produced the same charge.

Table 1 presents all ten allocation conditions at the prespecified ten-million budget. The QMIX reference uses the published point at 9,995,727 units, with mean 0.418920 and observed range 0-0.996239. Its unallocated budget is 4,273 units. MEDoE comparisons remain in the published return scale.

Table 1. Last affordable MEDoE evaluations at B=10M. Means use 16 released pointwise records per available method; unavailable MEDoE cells use none. Brackets in the MEDoE column are observed record ranges. Brackets beside the difference are deterministic all-cross-record ranges, not confidence intervals. Capital and unallocated budget are nominal interaction units.

Capital	R	Target step	MEDoE mean [range]	Difference [range]	Unallocated
Selected pair,6M	1	4.0M	0.9935 [0.9300,1.0000]	0.5746 [-0.0662,1.0000]	0.000M
Selected pair,6M	2	7.0M	0.9983 [0.9900,1.0000]	0.5794 [-0.0062,1.0000]	0.000M
Selected pair,6M	4	8.4M	0.9971 [0.9853,1.0000]	0.5782 [-0.0109,1.0000]	0.100M
Selected pair,6M	8	9.2M	0.9937 [0.9680,1.0000]	0.5748 [-0.0282,1.0000]	0.050M
Selected pair,6M	16	9.6M	0.9978 [0.9900,1.0000]	0.5789 [-0.0062,1.0000]	0.025M
Full library,24M	1	unavailable	unavailable	unavailable	unavailable
Full library,24M	2	unavailable	unavailable	unavailable	unavailable
Full library,24M	4	4.0M	0.9935 [0.9300,1.0000]	0.5746 [-0.0662,1.0000]	0.000M
Full library,24M	8	7.0M	0.9983 [0.9900,1.0000]	0.5794 [-0.0062,1.0000]	0.000M
Full library,24M	16	8.4M	0.9971 [0.9853,1.0000]	0.5782 [-0.0109,1.0000]	0.100M

Under the selected-pair boundary, all five sharing rules have an affordable MEDoE point. Its target coordinate moves from four million with one share to 9.6 million with sixteen shares. The mean return, however, follows 0.993542, 0.998333, 0.997125, 0.993708 and 0.997833. Thus later affordable observations do not produce a monotonic improvement in the recorded mean. This is a comparison of selected historical observations, not evidence that changing an accounting rule changes learning.

The eight available main-table differences are positive, approximately 0.5746-0.5794. Every corresponding deterministic cross-record range extends below zero. For example, the one-share selected-pair difference is 0.574622, with range [-0.066239,1.000000]. The positive difference between means and overlap among individual record values describe different features of the finite table. The range neither tests a population null hypothesis nor establishes equivalence when it contains zero.

Charging the whole library changes whether this evidence is accessible. At ten million units, R=1 and R=2 cannot pay even the initial transfer-point charge of 24 million and twelve million respectively. These are unavailable evaluations, not failures with a score of zero. At R=4, the complete-library share equals the selected-pair R=1 charge; both therefore select the same four-million target point. Equivalent charge identities explain the other repeated rows without creating additional outcome observations.

The remaining labels remain visible in the analysis. At ten million under the selected-pair one-share rule, IPPO and MAPPO each select their six-million point, both with mean zero, leaving four million units unallocated by the archive. PS selects its four-million target point with mean 0.123125 and range 0-1. Its recorded evidence is retained even though the main contrast is fixed to MEDoE and QMIX. The different evaluation densities are substantive: a sparse from-scratch grid leaves a much larger observation gap than the dense QMIX grid at this budget.

Table 2. MEDoE support across all fourteen fixed budgets. Counts describe budget cells, not independent trials. The archive-ended count is a subset of available cells and flags budgets strictly above the last recorded cost.

Capital	R	Available /14	Unavailable /14	Archive-ended /14
Selected pair,6M	1	12	2	6
Selected pair,6M	2	13	1	8
Selected pair,6M	4	14	0	9
Selected pair,6M	8	14	0	9
Selected pair,6M	16	14	0	9
Full library,24M	1	3	11	0
Full library,24M	2	9	5	3
Full library,24M	4	12	2	6
Full library,24M	8	13	1	8
Full library,24M	16	14	0	9

There are 67 available MEDoE budget cells under selected-pair allocation and 51 under complete-library allocation. Across both boundaries, 118 of 140 MEDoE-QMIX contrasts are available. Of these, 114 differences are positive and four negative. These counts are descriptive coverage of overlapping scenarios; dividing them into a probability of algorithm superiority would misstate the design.

The negative cells also clarify the missingness distinction. At C=24M, R=1 and B=24M, the selected MEDoE record is the actual released t=0 point, whose mean is zero. QMIX has mean 0.873597 at its last affordable published point, yielding difference -0.873597 and cross-record range [-1,0]. The preceding eleven smaller budgets in that allocation are unavailable. A recorded initial zero is therefore directly distinguishable from an unaffordable observation.

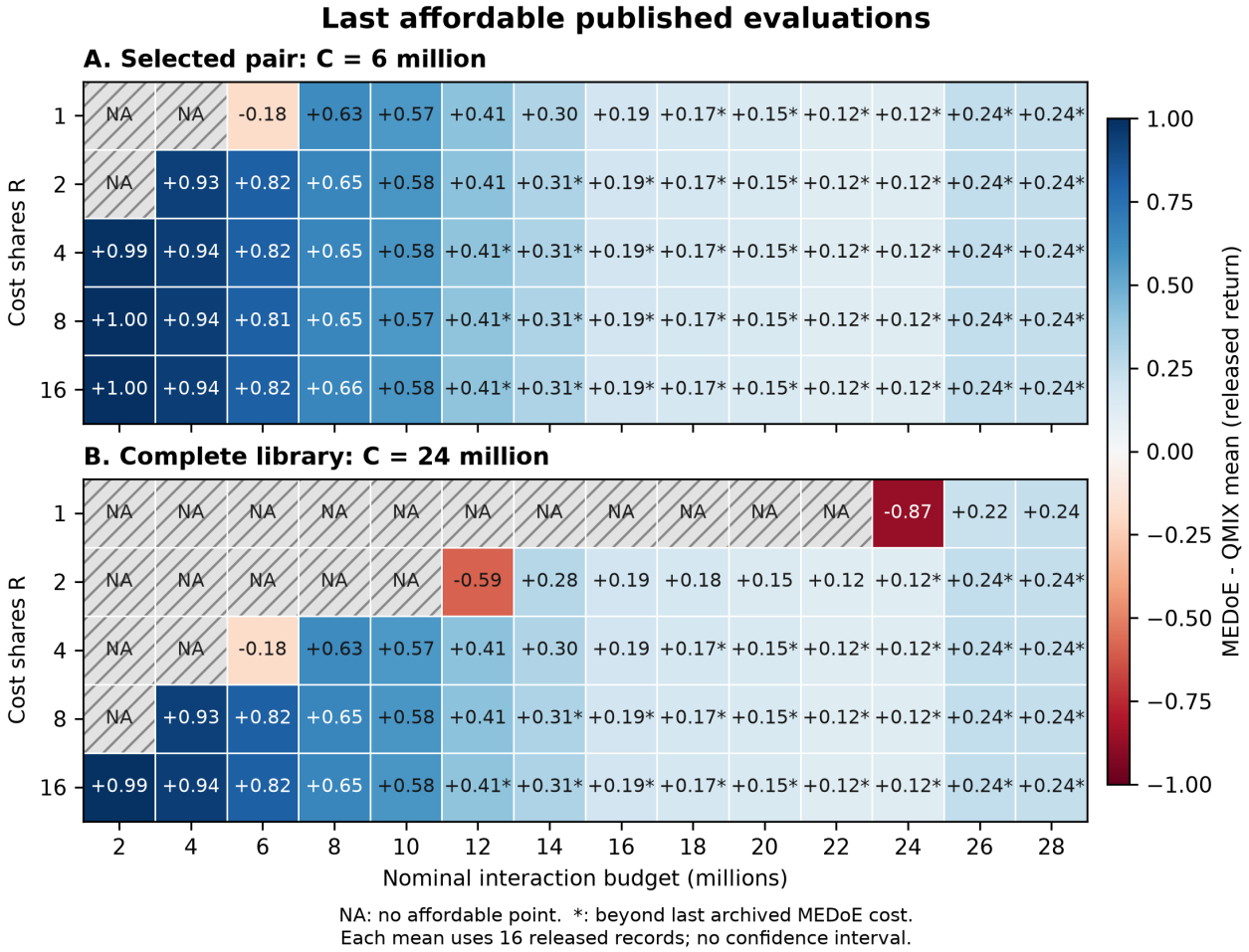


Figure 2. MEDoE mean minus QMIX mean at every frozen budget and allocation. NA and hatching identify no affordable MEDoE point. An asterisk marks a budget beyond MEDoE's last archived cost, rather than statistical significance. Cell values are rounded differences; no uncertainty interval is encoded by the color.

Among 90 adjacent-sharing comparisons where both MEDoE snapshots exist, 24 recorded mean changes are positive, 14 negative and 52 exactly zero. Affordable target steps never move backward. Many zero changes occur once the archive endpoint has been reached. Across the ten allocation conditions, 67 budget cells exceed the final MEDoE recorded cost. Their repeated terminal value is an archive boundary, not observed training through the larger budget. At the largest selected-pair, sixteen-share budget, 17.625 million units remain beyond the selected record's cost.

5. What changes when an asset is shared

The calculation separates two operations that can otherwise appear as one learning curve. A lower allocated source charge expands which existing evaluations fit a budget. The selected return then depends on the actual archived trajectory and its observation grid. The former relationship is monotonic by construction; the latter need not be. A line connecting budget-conditioned means would conceal this distinction if readers treated every horizontal position as a newly executed training job.

The result also depends on the investment boundary. A selected pair and a complete library are different resources. Charging the latter can be reasonable for a declared library-building portfolio, but it is not a factual correction to a target run that used only a selected pair. Conversely, claiming zero acquisition cost because files already exist would represent another allocation policy. Our study quantifies the two declared boundaries without presenting either as universally fair or optimal.

Algorithm proposals address a separate level of explanation. Quadratic Exponential Decrease Roll-Back proposes a modification to the PPO surrogate/gradient update and evaluates continuous-control tasks (Zhao et al., 2023). Its update-level efficiency question is relevant when interpreting what a nominal interaction budget measures. Our allocation arithmetic changes neither the PPO update nor a gradient bound, and the archive contains no execution of that proposal. Implementation studies likewise show why performance cannot be attributed to an algorithm name without its surrounding choices (Engstrom et al., 2020; Andrychowicz et al., 2021). This study holds the published implementations fixed and varies only how their recorded evidence is charged.

Nor does sharing a cost establish coordination with new partners. Other-play explicitly studies policies coordinating with previously unseen partners (Hu et al., 2020). That question requires identified partner assignments and suitable outcomes. The present row index does not preserve those assignments, and the sharing count does not generate them. The available evidence supports neither a new partner-generalization estimate nor a decomposition of source-team interactions.

The main limitations are consequently specific. We have one prepared Overcooked table, not a representative sample of environments. Its release-time interpolation or smoothing is not documented by the available plotting script. Nominal environmental counters omit several forms of computation, and saved metadata were checked without acquiring complete weight contents. Dense and sparse evaluation schedules produce unequal observation gaps. Finally, deterministic ranges describe the values in this release; authenticated run-level uncertainty would require provenance that the table no longer contains.

Documentation remains useful despite these bounds. Dataset documentation frameworks motivate recording provenance, structure and intended use (Gebru et al., 2021). Reproducible computational research requires retaining the transformations connecting inputs and claims (Peng, 2011). Here that chain consists of official member hashes, source-counter metadata, a frozen allocation protocol, selected record identifiers and a standard table with explicit unavailable states. It permits recalculation without asking an unobserved cost or missing run identity to behave like measured data.

6. Conclusion

In this released Overcooked case, charging shared source assets changes which already-published evaluation a fixed budget can afford. Selected-pair accounting permits high MEDoE mean returns at the ten-million budget under every declared sharing rule, while complete-library accounting leaves two of those conditions unavailable. The full grid retains both negative contrasts and budgets beyond the recorded endpoint. Later affordable steps also include decreases in recorded mean return. These findings describe a transparent allocation scenario applied to preserved observations. They do not add downstream tasks, turn row indices into independent seeds, or replace the original authors' source-inclusive accounting.

Data and computation availability

The official data DOI is 10.7488/ds/7731. Article-local source inventory, frozen protocol, static result-table decoder, analysis script, condition-summary CSV, result JSON and figure sources are retained with this manuscript. No complete model or experience-buffer member is included in the working evidence set. The saved CSV is identified by its actual byte SHA256 in the result JSON.

References

1. Sutton, Richard S., and Andrew G. Barto. *Reinforcement Learning: An Introduction*. 2nd ed., MIT Press, 2018. http://incompleteideas.net/book/the-book-2nd.html.

2. Gebru, Timnit; Morgenstern, Jamie; Vecchione, Briana; Vaughan, Jennifer Wortman; Wallach, Hanna; III, Hal Daumé; Crawford, Kate. (2021). Datasheets for datasets. Communications of the ACM. https://doi.org/10.1145/3458723

3. Henderson, Peter; Islam, Riashat; Bachman, Philip; Pineau, Joelle; Precup, Doina; Meger, David. (2018). Deep Reinforcement Learning That Matters. Proceedings of the AAAI Conference on Artificial Intelligence. https://doi.org/10.1609/aaai.v32i1.11694

4. Andrychowicz, Marcin; Raichuk, Anton; Stańczyk, Piotr; Orsini, Manu; Girgin, Sertan; Marinier, Raphael; Hussenot, Léonard; Geist, Matthieu; Pietquin, Olivier; Michalski, Marcin; Gelly, Sylvain; Bachem, Olivier. (2021). What Matters In On-Policy Reinforcement Learning? A Large-Scale Empirical Study. International Conference on Learning Representations. https://openreview.net/forum?id=nIAxjsniDzg

5. Rashid, Tabish; Samvelyan, Mikayel; Schroeder de Witt, Christian; Farquhar, Gregory; Foerster, Jakob; Whiteson, Shimon. (2018). QMIX: Monotonic Value Function Factorisation for Deep Multi-Agent Reinforcement Learning. International Conference on Machine Learning. https://proceedings.mlr.press/v80/rashid18a.html

6. Tirumala, Dhruva; Galashov, Alexandre; Noh, Hyeonwoo; Hasenclever, Leonard; Pascanu, Razvan; Schwarz, Jonathan; Desjardins, Guillaume; Czarnecki, Wojciech Marian; Ahuja, Arun; Teh, Yee Whye; Heess, Nicolas. (2022). Behavior Priors for Efficient Reinforcement Learning. Journal of Machine Learning Research. https://www.jmlr.org/papers/v23/20-1038.html

7. Taylor, Matthew E.; Stone, Peter. (2009). Transfer Learning for Reinforcement Learning Domains: A Survey. Journal of Machine Learning Research. https://www.jmlr.org/papers/v10/taylor09a.html

8. Fernández, Fernando; Veloso, Manuela. (2006). Probabilistic policy reuse in a reinforcement learning agent. Proceedings of the fifth international joint conference on Autonomous agents and multiagent systems. https://doi.org/10.1145/1160633.1160762

9. Dodge, Jesse; Gururangan, Suchin; Card, Dallas; Schwartz, Roy; Smith, Noah A.. (2019). Show Your Work: Improved Reporting of Experimental Results. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). https://doi.org/10.18653/v1/d19-1224

10. Peng, Roger D.. (2011). Reproducible Research in Computational Science. Science. https://doi.org/10.1126/science.1213847

11. Pan, Sinno Jialin; Yang, Qiang. (2010). A Survey on Transfer Learning. IEEE Transactions on Knowledge and Data Engineering. https://doi.org/10.1109/tkde.2009.191

12. Guo, Yunxiao; Xie, Xinjia; Zhao, Runhao; Zhu, Chenglan; Yin, Jiangting; Long, Han. (2023). Cooperation and Competition: Flocking with Evolutionary Multi-Agent Reinforcement Learning. Lecture Notes in Computer Science. https://doi.org/10.1007/978-3-031-30105-6_23

13. Fosong, Elliot; Rahman, Arrasy; Carlucho, Ignacio; Albrecht, Stefano V.. (2024a). Learning Complex Teamwork Tasks Using a Given Sub-task Decomposition. Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems, pp. 598-606. https://www.ifaamas.org/Proceedings/aamas2024/pdfs/p598.pdf

14. Bouthillier, Xavier; Delaunay, Pierre; Bronzi, Mirko; Trofimov, Assya; Nichyporuk, Brennan; Szeto, Justin; Mohammadi Sepahvand, Nazanin; Raff, Edward; Madan, Kanika; Voleti, Vikram; Ebrahimi Kahou, Samira; Michalski, Vincent; Serdyuk, Dmitriy; Arbel, Tal; Pal, Chris; Varoquaux, Gael; Vincent, Pascal. (2021). Accounting for Variance in Machine Learning Benchmarks. Proceedings of Machine Learning and Systems. https://proceedings.mlsys.org/paper_files/paper/2021/file/0184b0cd3cfb185989f858a1d9f5c1eb-Paper.pdf

15. Schulman, John; Wolski, Filip; Dhariwal, Prafulla; Radford, Alec; Klimov, Oleg. (2017). Proximal Policy Optimization Algorithms. arXiv preprint. https://arxiv.org/abs/1707.06347

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

17. Schwartz, Roy; Dodge, Jesse; Smith, Noah A.; Etzioni, Oren. (2020). Green AI. Communications of the ACM. https://doi.org/10.1145/3381831

18. Schulman, John; Levine, Sergey; Abbeel, Pieter; Jordan, Michael; Moritz, Philipp. (2015). Trust Region Policy Optimization. International Conference on Machine Learning. https://proceedings.mlr.press/v37/schulman15.html

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. Narvekar, Sanmit; Peng, Bei; Leonetti, Matteo; Sinapov, Jivko; Taylor, Matthew E.; Stone, Peter. (2020). Curriculum Learning for Reinforcement Learning Domains: A Framework and Survey. Journal of Machine Learning Research. https://www.jmlr.org/papers/v21/20-212.html

21. Yu, Chao; Velu, Akash; Vinitsky, Eugene; Gao, Jiaxuan; Wang, Yu; Bayen, Alexandre; Wu, Yi. (2022). The Surprising Effectiveness of PPO in Cooperative Multi-Agent Games. Advances in Neural Information Processing Systems. https://papers.nips.cc/paper/2022/hash/9c1535a02f0ce079433344e14d910597-Abstract-Datasets_and_Benchmarks.html

22. Hu, Hengyuan; Lerer, Adam; Peysakhovich, Alex; Foerster, Jakob. (2020). “Other-Play” for Zero-Shot Coordination. International Conference on Machine Learning. https://proceedings.mlr.press/v119/hu20a.html

23. Cawley, Gavin C.; Talbot, Nicola L. C.. (2010). On Over-fitting in Model Selection and Subsequent Selection Bias in Performance Evaluation. Journal of Machine Learning Research. https://jmlr.org/papers/v11/cawley10a.html

24. Rusu, Andrei A.; Rabinowitz, Neil C.; Desjardins, Guillaume; Soyer, Hubert; Kirkpatrick, James; Kavukcuoglu, Koray; Pascanu, Razvan; Hadsell, Raia. (2016). Progressive Neural Networks. arXiv preprint. https://arxiv.org/abs/1606.04671

25. Dehghani, Mostafa; Arnab, Anurag; Beyer, Lucas; Vaswani, Ashish; Tay, Yi. (2022). The Efficiency Misnomer. International Conference on Learning Representations. https://openreview.net/forum?id=uQShkCh8C97

26. Zhao, Runhao; Xu, Dan; Jian, Shuoqu; Tan, Tianle; Sun, Xiaohui; Zhang, Wanchao. (2023). Quadratic Exponential Decrease Roll-Back: An Efficient Gradient Update Mechanism in Proximal Policy Optimization. 2023 2nd International Conference on Machine Learning, Cloud Computing and Intelligent Mining (MLCCIM). https://doi.org/10.1109/mlccim60412.2023.00015

27. Kingma, Diederik P.; Ba, Jimmy. (2015). Adam: A Method for Stochastic Optimization. International Conference on Learning Representations. https://arxiv.org/abs/1412.6980

28. Engstrom, Logan; Ilyas, Andrew; Santurkar, Shibani; Tsipras, Dimitris; Janoos, Firdaus; Rudolph, Larry; Madry, Aleksander. (2020). Implementation Matters in Deep Policy Gradients: A Case Study on PPO and TRPO. arXiv preprint. https://arxiv.org/abs/2005.12729

29. Agarwal, Rishabh; Schwarzer, Max; Castro, Pablo Samuel; Courville, Aaron; Bellemare, Marc G.. (2021). Deep Reinforcement Learning at the Edge of the Statistical Precipice. arXiv preprint. https://arxiv.org/abs/2108.13264

30. Fosong, Elliot; Rahman, Arrasy; Carlucho, Ignacio; Albrecht, Stefano V.. (2024b). Learning Complex Teamwork Tasks Using a Given Sub-task Decomposition [dataset]. University of Edinburgh. School of Informatics. https://doi.org/10.7488/ds/7731