Benchmarking and Comparative Evaluation
Volume 1 · Number 1 · 2026

Choosing a Game Policy by Its Mean or Its Variability: An Audit of Public Atari Checkpoint Reports

Brandon Peterson[1] · Trevor Clark[2] · Matthew Fisher[3]

[1] Belgium · [2] Austria · [3] Switzerland

Original publication: October 6, 2026 | Article 565

Abstract

The best reported average game reward need not identify the checkpoint preferred by a variability-sensitive rule. In a public table covering four algorithms on ten Atari games, subtracting half the reported episode standard deviation changed the selected checkpoint for one game. For Asteroids, the choice moved from QR-DQN to PPO. The same table supplied between five and 258 evaluation episodes per condition. A game- and algorithm-adjusted association between episode count and normalized reward spread was negative but uncertain. The audit separates an explicit selection preference from statistical confidence: a half-standard-deviation penalty is neither a confidence bound nor evidence about variability across independently trained agents.

Analysis of captured public evaluation summaries. No reinforcement-learning agent or interactive renderer was run for this study.

Choosing a Game Policy by Its Mean or Its Variability: An Audit of Public Atari Checkpoint Reports

Analysis of captured public evaluation summaries. No reinforcement-learning agent or interactive renderer was run for this study.

Abstract

The best reported average game reward need not identify the checkpoint preferred by a variability-sensitive rule. In a public table covering four algorithms on ten Atari games, subtracting half the reported episode standard deviation changed the selected checkpoint for one game. For Asteroids, the choice moved from QR-DQN to PPO. The same table supplied between five and 258 evaluation episodes per condition. A game- and algorithm-adjusted association between episode count and normalized reward spread was negative but uncertain. The audit separates an explicit selection preference from statistical confidence: a half-standard-deviation penalty is neither a confidence bound nor evidence about variability across independently trained agents.

1. What does a reported game score recommend?

A table of average rewards invites a simple choice: select the largest number in a game's column. That rule is appropriate only for a decision maker who values the reported mean without an additional preference about variability. Another reader may prefer a checkpoint with slightly lower mean reward and fewer highly variable episodes. These readers can make different choices from exactly the same observations without either changing an algorithm or rerunning the environment.

This article asks whether a declared variability penalty changes checkpoint selection in a public Atari evaluation table. It also examines how the reported spread relates to the number of evaluation episodes after accounting for game and algorithm labels. These are questions about the interpretation of a captured report. They do not compare newly trained agents or estimate how a user would actually trade reward against variability in an interactive product.

Interactive-game generation, including Meng et al., 2026, makes evaluation criteria relevant beyond a static image score. Our analysis concerns game-policy reward reports, not the rendering behavior of that framework. The relationship is contextual: a game system can be evaluated through different outcomes, and the choice among them should be explicit. The present data cannot establish visual fidelity, action responsiveness, or success of a particular game-generation architecture.

2. The captured source makes an important qualification

The source is the public RL Baselines3 Zoo performance table, cited as RL Baselines3 Zoo contributors, 2026. Its documentation states that each listed condition corresponds to one training run and qualifies the table as a way to expose maximal performance, identify possible problems, and provide access to pretrained agents. It should not be silently converted into a repeated-seed algorithm benchmark. We retain that qualification throughout the empirical interpretation.

The implementation setting described by Raffin et al., 2021 supplies context for the repository, while the platform in Bellemare et al., 2012 explains the Atari evaluation environment. Neither makes the forty selected rows independent training replications. The reproducibility concerns discussed by Henderson et al., 2017 and the statistical-evaluation arguments of Agarwal et al., 2021 are particularly relevant here:

variability among evaluation episodes and variability across independently trained agents answer different questions.

Environment interfaces in Towers et al., 2024 and implementation documentation in Huang et al., 2021 further illustrate the importance of recording what a run actually means. Generalization benchmarks such as Cobbe et al., 2019 address still other questions about environments and transfer. We do not join their scores to the Zoo table, assume identical observation preprocessing, or use their task counts as additional sample size. The acquired table alone defines the data reanalyzed below.

3. Forty checkpoint conditions, ten separate reward scales

We parse the rows whose environment identifier contains `NoFrameskip`. The captured table has forty such rows: four algorithm labels for each of ten games. The labels are A2C, DQN, PPO, and QR-DQN. Each row supplies mean reward, reward standard deviation, listed training steps, evaluation steps, and completed evaluation episodes. There is exactly one row for each game-algorithm combination in the selected subset.

All forty selected conditions list ten million training steps. A log-training-step regressor would therefore be constant and inseparable from the model intercept. It is excluded from the fitted design. Equal listed step counts nevertheless do not imply matched computational cost, identical tuning effort, or equivalent checkpoint selection histories. The table is an observational report of available trained agents, rather than a randomized intervention assigning algorithms under identical development conditions.

Reward units differ across games. A change of one hundred points in Asteroids does not have the same interpretation as that change in Pong. We therefore make selections within a game. For the descriptive spread regression, each row's reward standard deviation is divided by the observed range of the four policy means in that game. Every game's range is positive in the captured subset, so the resulting normalized quantity is well-defined.

This range normalization expresses spread relative to the separation among the four supplied means. It is not a human-normalized score, a percentage of an optimal policy's return, or a property of the game alone. Adding another policy could change the range and therefore change normalized spread for existing policies. The normalization is useful for this bounded cross-row description, but its dependence on the observed comparison set must remain visible.

4. A transparent preference rule rather than a confidence bound

The first selection rule chooses the checkpoint with the highest reported mean reward within each game. The second chooses the checkpoint with the highest mean reward minus one half of the reported reward standard deviation. The penalty coefficient is fixed at 0.5 before calculating selections. Ties, if present, are resolved using the algorithm label so that the procedure is reproducible rather than dependent on incidental row order.

The half-standard-deviation penalty describes an explicit preference for lower episode spread. It is not estimated from observed user behavior. It also is not a lower confidence limit: a standard deviation describes dispersion, while a standard error for an estimated mean would require a sampling design and relevant dependence assumptions. We do not divide by the square root of the episode count and present the result as established inferential uncertainty.

The distinction matters because the reported episode counts vary from five to 258. Episodes may differ in length, and a time-limited evaluation can produce very different counts across games and policies. The source does not establish that all episodes are independently and identically distributed in a way sufficient

for a conventional confidence calculation. We preserve the source spread as a descriptive input to the preference rule rather than constructing unsupported precision weights.

World-model research in Ha and Schmidhuber, 2018 and game-oriented generation work in Valevski et al., 2024, Bruce et al., 2024, Micheli et al., 2022, and Alonso et al., 2024 address interactive representations or generation. The policy-optimization setting in Schulman et al., 2017 is another distinct part of the learning pipeline. None is compared experimentally in this source-table audit. Choosing among available reward summaries therefore does not establish a training-algorithm advantage, nor measure the visual quality or consistency of generated frames.

5. One game changes its selected checkpoint

The mean rule and the penalized rule agree on nine of the ten games. They disagree on Asteroids: QR-DQN has the largest supplied mean, while PPO has the largest supplied mean after subtracting half the corresponding standard deviation. This is an observed consequence of a declared criterion applied to the captured values. It does not establish that PPO would outperform QR-DQN in a fresh evaluation or satisfy a measured user's preferences.

For the mean criterion, QR-DQN is selected on Asteroids, BeamRider, Enduro, RoadRunner, Seaquest, and SpaceInvaders. PPO is selected on Breakout, Pong, and Qbert, while DQN is selected on MsPacman. Under the fixed variability penalty, only the Asteroids entry changes. These selections are descriptive checkpoint choices. They are not an aggregate algorithm ranking across a population of games or a statement about how frequently an algorithm wins across training seeds.

The observed policy-mean range varies substantially by game: it is approximately 3.697 in Pong, 118.240 in Breakout, and 14,232.643 in BeamRider. These values help explain why raw reward standard deviations should not be pooled directly across tasks. They also show why a fixed within-game selection penalty and a cross-game normalization answer related but distinct questions. The penalty uses original game units; the regression examines spread relative to the captured mean separation.

Scoring principles in Gneiting and Raftery, 2007 encourage stating the objective before interpreting the resulting choice. Forecast-accuracy discussions such as Hyndman and Koehler, 2006 similarly warn against comparing numerical losses without attention to their scale. We do not infer that the half-standard-deviation rule is optimal. Its value in the audit is that it makes a preference concrete enough to reveal whether the supplied report supports a different decision under that preference.

6. A descriptive model of reported episode spread

The second analysis regresses normalized reward spread on the logarithm of completed evaluation episodes, algorithm indicators, and game indicators. The forty rows remain the observations. Game indicators absorb mean differences between the ten environments, while algorithm indicators absorb average differences between the four labels. The design is checked for full rank after excluding the constant training-budget variable. It is a linear conditional-mean description, not a model of how evaluation episodes are generated.

We cluster covariance by game, allowing the four checkpoint rows within an environment to share errors. The procedure follows dependence concerns in Cameron and Miller, 2015, includes the recorded small-sample correction, and uses a t reference with nine degrees of freedom. Ten clusters provide limited support for broad sampling approximations. Heteroskedasticity-only covariance in White, 1980 would address another variance issue without representing this within-game grouping.

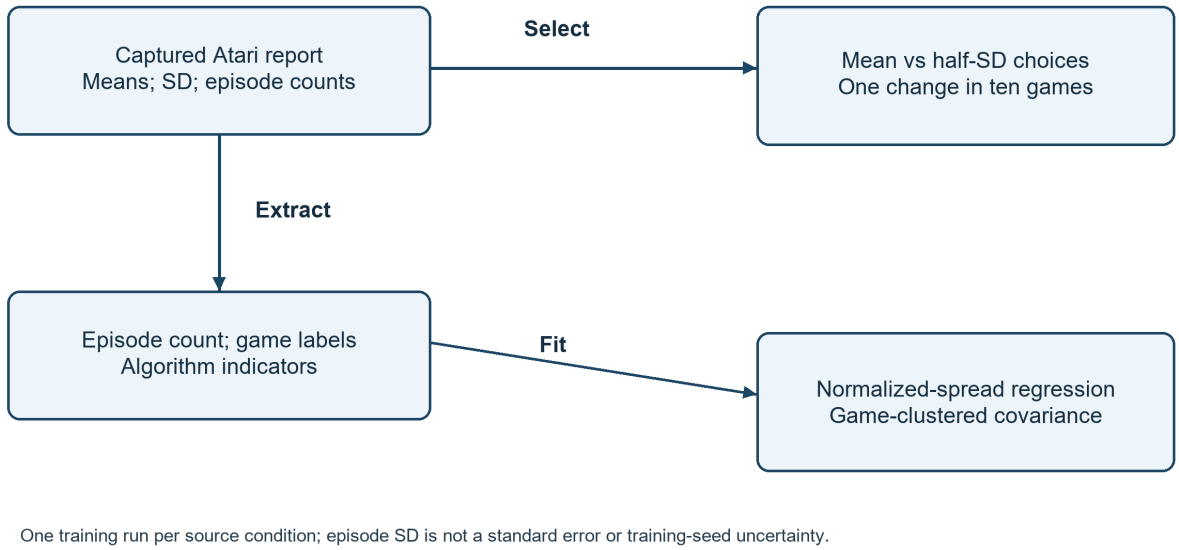

Figure 1. A captured reward report supports a declared within-game selection contrast and a separate normalized-spread regression.

Table 1. Evaluation-episode association with normalized reported reward spread, adjusted for game and algorithm indicators.

Model	Term	Estimate (SE)	95% CI	N
M1	Log evaluation episodes	-0.2913 (0.1619)	[-0.6576, 0.0750]	40
M1: fit	R-squared = 0.5542	-	-	40

Model labels: M1 = Reported normalized episode-spread association.

The log-episode coefficient is -0.2913, with standard error 0.1619 and interval -0.6576 to 0.0750. Its point direction is negative, but the interval includes zero. We therefore do not claim a reliably lower spread for conditions with more reported episodes. The model R-squared is 0.5542, reflecting the full fitted specification. It is not the share of training uncertainty explained or a validation score for predicting future game-policy behavior.

7. Episode count is an outcome of the evaluation process too

A condition can complete more episodes because its episodes terminate sooner. A different policy can survive longer and therefore supply fewer completed episodes during a similar evaluation-step budget. Consequently, episode count is not a randomly assigned measurement budget in this source. Regressing spread on that count does not estimate the causal benefit of collecting another episode. The association may combine task dynamics, policy behavior, and evaluation bookkeeping.

The normalization adds another dependence: all four rows within a game share a denominator estimated from the same four reported means. Game clustering acknowledges some shared variation but does not create independent information about that denominator. A larger release with repeated independently trained agents and raw episode returns would allow a richer analysis. The current table does not supply those distributions, so we do not fabricate them by sampling normal rewards around the published means.

System-reporting concerns in Sculley et al., 2015 are relevant because a table can lose important development and evaluation context as it is reused. The documentation's one-run warning is therefore

treated as part of the evidence. Removing it would make the table appear better suited to a formal algorithm comparison than its creators claim. The audit preserves the warning while still extracting a useful, limited decision consequence from the values that are actually present.

Robustness traditions in Huber, 1964, Hampel, 1974, and Rousseeuw, 1984 also should not be confused with the preference penalty. Subtracting a standard deviation does not make the supplied mean a robust estimator or demonstrate resistance to corrupted episodes. Influence reasoning in Cook, 1977 could identify games with unusual leverage in the spread model. No game has been removed because it weakens the observed log-episode association or makes its interval cross zero.

8. What a stronger report would need

Validation guidance in Roberts et al., 2017 and selection-bias analysis in Varma and Simon, 2006 becomes important if a criterion is tuned on the same games used to advertise its advantage. Our penalty is a fixed illustration, not the winner of a search over risk parameters. Multiplicity considerations in Benjamini and Hochberg, 1995 would apply to a broader exploration of penalties, normalizations, and subgroups. We retain both the agreement on nine games and the single changed selection, without emphasizing only whichever fact seems more dramatic.

A stronger decision study would define an actual user's objective, provide multiple training seeds, retain episode returns, and separate policy development from evaluation. It could then ask whether a preference-sensitive selection rule improves outcomes on new runs. None of those additional observations is inferred from a large number of evaluation steps in a single captured row. More steps within one condition do not establish independent replication of the training process.

Interpretation under Wasserstein and Lazar, 2016 favors reporting the uncertain regression together with the exact selection count. Documentation following Gebru et al., 2021 records the captured table, Atari filter, constant training budget, range normalization, penalty, and tie rule. Wilkinson et al., 2016 motivates preserving both the source and parsed values. The empirical contribution is modest but concrete: a public mean-reward ranking contains enough variability information to support a different declared choice for one game, while the report remains insufficient for a repeated-seed algorithm verdict.

The captured checkpoint rows also fix which policies are available to choose. A preference-sensitive criterion cannot recover an unreported policy, infer the shape of an episode-return distribution from its mean and standard deviation, or estimate dependence between different policies. The observed selection comparison is therefore conditional on a finite menu as well as on the declared penalty. Those two restrictions should travel with any recommendation based on the audit.

References

1. Agarwal, Rishabh, et al. "Deep Reinforcement Learning at the Edge of the Statistical Precipice" *arXiv preprint arXiv:2108.13264*, 2021. https://arxiv.org/abs/2108.13264.
2. Wasserstein, Ronald L.; Lazar, Nicole A.. (2016). The ASA Statement on p -Values: Context, Process, and Purpose. The American Statistician. https://doi.org/10.1080/00031305.2016.1154108.
3. RL Baselines3 Zoo contributors, . (2026). RL Baselines3 Zoo: published performance of trained agents. Captured 7 November 2026; one-run qualification retained. https://github.com/DLR-RM/rl-baselines3-zoo/blob/master/benchmark.md.
4. Schulman, John; Wolski, Filip; Dhariwal, Prafulla; Radford, Alec; Klimov, Oleg. (2017). Proximal Policy Optimization Algorithms. arXiv:1707.06347. https://arxiv.org/abs/1707.06347.
5. Bruce, Jake, et al. "Genie: Generative Interactive Environments" *arXiv preprint arXiv:2402.15391*, 2024. https://arxiv.org/abs/2402.15391.

6. Sculley, D., et al. "Hidden Technical Debt in Machine Learning Systems." *Advances in Neural Information Processing Systems*, vol. 28, 2015. https://papers.nips.cc/paper/2015/hash/86df7dcfd896fcaf2674f757a2463eba-Abstract.html

7. White, Halbert. “A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity.” *Econometrica*, vol. 48, no. 4, 1980, pp. 817-838. https://doi.org/10.2307/1912934.

8. Ha, David and Jürgen Schmidhuber. "World Models" *arXiv preprint arXiv:1803.10122*, 2018. https://arxiv.org/abs/1803.10122.

9. Raffin, Antonin; Hill, Ashley; Gleave, Adam; Kanervisto, Anssi; Ernestus, Maximilian; Dormann, Noah. (2021). Stable-Baselines3: Reliable Reinforcement Learning Implementations. Journal of Machine Learning Research 22(268), 1-8. https://jmlr.org/papers/v22/20-1364.html.

10. Rousseeuw, Peter J.. (1984). Least Median of Squares Regression. Journal of the American Statistical Association. https://doi.org/10.1080/01621459.1984.10477105.

11. Bellemare, Marc G., et al. "The Arcade Learning Environment: An Evaluation Platform for General Agents" *arXiv preprint arXiv:1207.4708*, 2012. https://arxiv.org/abs/1207.4708.

12. Henderson, Peter; Islam, Riashat; Bachman, Philip; Pineau, Joelle; Precup, Doina; Meger, David. (2017). Deep Reinforcement Learning that Matters. arXiv:1709.06560. https://arxiv.org/abs/1709.06560.

13. Roberts, David R., et al. “Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure.” *Ecography*, vol. 40, no. 8, 2017, pp. 913-929. https://doi.org/10.1111/ecog.02881.

14. Cobbe, Karl, et al. "Leveraging Procedural Generation to Benchmark Reinforcement Learning" *arXiv preprint arXiv:1912.01588*, 2019. https://arxiv.org/abs/1912.01588.

15. Cook, R. Dennis. (1977). Detection of Influential Observation in Linear Regression. Technometrics. https://doi.org/10.1080/00401706.1977.10489493.

16. Benjamini, Yoav and Yosef Hochberg. “Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing.” *Journal of the Royal Statistical Society Series B: Statistical Methodology*, vol. 57, no. 1, 1995, pp. 289-300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x.

17. Gneiting, Tilmann; Raftery, Adrian E. (2007). Strictly Proper Scoring Rules, Prediction, and Estimation. Journal of the American Statistical Association. https://doi.org/10.1198/016214506000001437.

18. Micheli, Vincent, et al. "Transformers are Sample-Efficient World Models" *arXiv preprint arXiv:2209.00588*, 2022. https://arxiv.org/abs/2209.00588.

19. Hyndman, Rob J. and Anne B. Koehler. “Another look at measures of forecast accuracy.” *International Journal of Forecasting*, vol. 22, no. 4, 2006, pp. 679-688. https://doi.org/10.1016/j.ijforecast.2006.03.001.

20. Valevski, Dani, et al. "Diffusion Models Are Real-Time Game Engines" *arXiv preprint arXiv:2408.14837*, 2024. https://arxiv.org/abs/2408.14837.

21. Gebru, Timnit, et al. "Datasheets for datasets" *Communications of the ACM*, 2021. https://doi.org/10.1145/3458723.

22. 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:2407.17032. https://arxiv.org/abs/2407.17032.

23. Huang, Shengyi, et al. "CleanRL: High-quality Single-file Implementations of Deep Reinforcement Learning Algorithms." *arXiv preprint arXiv:2111.08819*, 2021. https://arxiv.org/abs/2111.08819.

24. 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.

25. Huber, Peter J.. (1964). Robust Estimation of a Location Parameter. The Annals of Mathematical Statistics. https://doi.org/10.1214/aoms/1177703732.

26. Hampel, Frank R.. (1974). The Influence Curve and its Role in Robust Estimation. Journal of the American Statistical Association. https://doi.org/10.1080/01621459.1974.10482962.

27. Alonso, Eloi, et al. "Diffusion for World Modeling: Visual Details Matter in Atari" *arXiv preprint arXiv:2405.12399*, 2024. https://arxiv.org/abs/2405.12399.

28. Cameron, A. Colin, and Douglas L. Miller. “A Practitioner’s Guide to Cluster-Robust Inference.” *Journal of Human Resources*, vol. 50, no. 2, 2015, pp. 317-372. https://doi.org/10.3368/jhr.50.2.317.

29. Meng, Zijie; Che, Jinming; Wei, Bingcai; Cao, Xixin. (2026). Make a game: A novel paradigm for interactive game rendering. ICASSP 2026, 1026-1030. https://doi.org/10.1109/ICASSP55912.2026.11461927.

30. Varma, Sudhir and Richard Simon. “Bias in error estimation when using cross-validation for model selection.” *BMC Bioinformatics*, vol. 7, no. 1, 2006. https://doi.org/10.1186/1471-2105-7-91.