Observable State and Hidden Failure in Neural Game Rendering

Arthur Douglas (BH), Vincent Brooks (BW), Liam Gardner (SM)

Abstract

A generated frame may be compatible with several underlying game states. This conceptual analysis asks how an evaluator can detect a hidden state error when current appearance remains plausible. It proposes action sequences that reveal earlier decisions, separate visual and behavioral observations, and record which state variables have an external reference. Make a game motivates the interactive setting. No state-recovery experiment is performed. The central claim is methodological: an evaluation should distinguish what is directly observed, what can be probed through consequences, and what remains an untested assumption about the renderer's internal representation.

1. The current image can be insufficient evidence

Two histories can produce a similar view while requiring different later behavior. An object may have been activated, a resource may have been consumed, or a route may have become unavailable outside the visible region. A renderer that produces a convincing current image can still have lost the earlier distinction. The error becomes visible only when another action depends on it. A frame-level comparison therefore need not be a complete test of the state required for interaction.

Make a game treats rendered content as part of an interactive process (Meng et al., 2026a). That motivates examining the relation between visible output and task state, without establishing a particular hidden-state implementation from the available source description. This article offers a conceptual test design. It does not inspect internal representations, run the proposed system, or measure a new game environment. The aim is to identify which observations would make a state-preservation claim falsifiable rather than infer it from visual plausibility.

A benchmark interface supplies a common language for launching tasks, but it does not eliminate differences among the tasks. The Arcade Learning Environment was designed to evaluate general agents across games (Bellemare et al., 2012). Gym offers a common interaction interface (Brockman et al., 2016), while Procgen makes procedural variation an explicit part of the assessment population (Cobbe et al., 2019). These contributions help distinguish software compatibility from evidence of transfer. The BBF release supplies actual repeated-run scores that can be reanalyzed (Schwarzer et al., 2023), whereas the statistical-precipice analysis explains why uncertainty and aggregation choices deserve attention (Agarwal et al., 2021). Together, these sources motivate preserving task-level results. They do not imply that an Atari return is a visual-quality score or that a renderer supports the mechanics of every game available through the same interface. A proposed evaluation should identify its games, initial conditions, and action histories, then state which population the aggregate represents. Additional runs can improve knowledge of a task's outcome distribution without increasing the number of different tasks observed.

Replication has several levels: a second measurement, a second trajectory, and a second independently acquired setting are different additions to evidence. Statistical calculations should retain those distinctions. Heteroskedasticity-robust regression inference concerns unequal conditional variances (White, 1980); cluster-robust inference concerns dependencies shared within groups (Cameron and Miller, 2015). The group definition should be justified through the data collection and evaluation process, and not chosen because it produces a preferred significance result. Unequal group sizes also influence what a pooled average represents. Reporting both the group count and observation count permits a reader

to see whether the calculation describes common cases, heavily recorded cases, or an explicitly balanced group population. No variance estimator alone resolves that choice of estimand.

2. A sequence designed to reveal the earlier distinction

A proposed experiment could construct two histories that end in comparable visible conditions but differ in one task-relevant earlier event. A later probe action would have different required consequences under the two histories. The evaluator would retain the complete sequence and the rule defining the distinction. The relevant outcome would be whether the final behavior remains compatible with the earlier event, not merely whether the two final images differ. An arbitrary visual difference would not establish the intended state dependence.

The design needs an external reference for the task rule. A conventional environment may provide a state log. A generative interface may provide only commands and images. In the latter case, a behavioral probe can support a limited observational claim, but it cannot prove that a specific latent variable was continuously maintained. An evaluator should say which state quantities are directly recorded, which are inferred through consequences, and which cannot be assessed. These categories should not be collapsed into a general statement that the system understands the world.

The probe must also be sensitive. If the selected action produces the same required consequence under both histories, a successful-looking output cannot discriminate between correct state preservation and a memoryless continuation. A useful protocol should therefore include a task-level check that the alternative histories actually imply distinguishable outcomes. This prevents a nominally sophisticated test from becoming another comparison of images that does not exercise the claimed capability.

Learned world models connect visual representation to behavior in several ways. World Models explores compressed representations of environment experience (Ha and Schmidhuber, 2018). Dreamer uses latent imagination to support behavior learning (Hafner et al., 2019), IRIS studies transformer-based world modeling (Micheli et al., 2022), and DIAMOND investigates diffusion-based representations in which visual detail can matter to control (Alonso et al., 2024). These works establish a useful set of questions rather than a single interchangeable implementation. What information is retained, which histories condition prediction, and how is a predicted observation connected to an action consequence? A rendering study should state which of these questions it actually tests. Visually plausible outputs alone do not establish an accurate transition model. Conversely, an agent can obtain a favorable return without reproducing all details relevant to a human viewer. The measurement design should preserve this distinction and avoid translating a result from one endpoint into an unobserved improvement on another.

The scientific interpretation should specify what would change under the proposed intervention. Holding an observed covariate fixed statistically is not equivalent to manipulating it in an environment. A causal graph helps separate confounders, mediators, and selection mechanisms (Greenland et al., 1999). It can also reveal why controlling for every recorded variable is not automatically defensible. Meanwhile, an evaluation with multiple conditions and endpoints creates a family of opportunities for apparently striking findings. Multiplicity methods require a declared set of comparisons and an appropriate inferential purpose (Benjamini and Hochberg, 1995). Descriptive tables may be useful without testing every cell. When the data do not identify a mechanism, the appropriate contribution can be a sharper question and a transparent comparison rather than an unsupported causal answer.

A validation split defines which form of transfer has been observed. Related observations can share subjects, scenes, prompts, or acquisition conditions, so randomly splitting rows may leave important information on both sides of the boundary. Structured cross-validation addresses these dependencies

(Roberts et al., 2017). Model selection introduces another boundary: choosing a representation, threshold, or tuning parameter using the final assessment data makes the resulting error estimate optimistic (Varma and Simon, 2006). The appropriate held-out unit follows the intended future use. Generalization to a new frame, a new trajectory, and a new subject are not equivalent claims. A study should record those units before fitting or choosing a scoring rule, and preserve an assessment set that has not become part of iterative development.

3. Interpreting a failed or ambiguous probe

A failed probe can arise from several stages. The system may not have retained the relevant history, may have misunderstood the later command, or may have produced an output that the evaluator cannot measure. The result should preserve these observations without assigning an internal cause prematurely. Intermediate representations, when available, can help formulate hypotheses, but agreement between them does not itself establish correct external behavior. The task consequence remains the principal evidence.

An ambiguous output should be recorded as ambiguous. Forcing every case into correct or incorrect can conceal a measurement limitation, while dropping ambiguous cases can inflate apparent capability. A report should give both the rate of valid assessment and the result among assessable cases. Independent judgments may help if observers receive the relevant history and task instructions. A judge who sees only the final image cannot be expected to recover an earlier hidden distinction that the image does not reveal.

The test population should vary the delay, intervening events, and visual similarity of the paired histories. A result on one short sequence does not establish that every task variable survives arbitrary delay or distraction. The design can instead characterize the supported range of conditions. That limited profile is scientifically valuable because it identifies where further experiments are needed and what kind of state demand those experiments should impose. Broad claims should not be obtained by omitting the circumstances in which the probe loses discriminating power.

Generative interaction introduces measurement requirements beyond ordinary video synthesis. GameNGen studies neural generation supporting real-time game interaction (Valevski et al., 2024), while Genie investigates action-controllable environments learned from unlabeled videos (Bruce et al., 2024). Their objectives help distinguish response to control from the production of plausible motion. Structural similarity evaluates a defined image comparison (Wang et al., 2004), learned perceptual similarity uses a different visual representation (Zhang et al., 2018), and video-distribution metrics add temporal features to generative assessment (Unterthiner et al., 2018). None of these scores automatically establishes that an action had its intended effect. A practical study should preserve the action, its timing, and the relevant output history alongside visual measurements. It should also record cases where the system produces attractive imagery but violates a state or control requirement. Such disagreement can identify an important limitation that a combined average would hide. A multi-endpoint report is therefore more informative than an unexplained total labeled interactive quality.

An automated measurement should not inherit the authority of its training labels without a direct assessment of use. The relevant checks include systematic disagreement with the intended procedure (Bland and Altman, 1986) and calibration of any probability claims (Guo et al., 2017). Ranking performance, absolute accuracy, and confidence reliability are distinct outcomes. A study can improve one while leaving the others unchanged. Keeping the measurements separate permits examination of those tradeoffs and avoids implying that one favorable statistic establishes all aspects of validity. When independent judgments are unavailable, the limitation should remain part of the result. The absence of a target observation cannot be repaired by combining several correlated proxies into a more elaborate total score.

4. Controls, histories, and the evaluation boundary

An action-dependent observation has meaning only within a specified interaction process. Reinforcement learning provides a formal distinction among states, actions, rewards, and policies (Sutton and Barto, 2018). Offline learning further emphasizes that a fixed collection of histories has limited behavioral support (Levine et al., 2020). Conservative value learning addresses a particular problem arising when actions are evaluated beyond that support (Kumar et al., 2020). These sources do not validate a rendering architecture, but they clarify why familiar trajectories can be an incomplete test. A renderer may look reliable when replaying common histories while responding poorly to interventions that the archive rarely contains. A future evaluation should state which commands and transitions are supported by its data, which are deliberately held out, and which remain untested. Reweighting observed histories can alter their importance but cannot create an observation of a missing control consequence. That boundary should stay visible in the interpretation of any broad capability claim.

Difficult cases can disappear at several stages: data eligibility, generation, measurement, and final scoring. Each exclusion changes the population represented by the reported result. Coverage-aware evaluation makes those decisions visible (Geifman and El-Yaniv, 2017). A confidence score should also be checked when the input distribution changes, because uncertainty under shift is an empirical property rather than a guarantee of the model interface (Ovadia et al., 2019). An informative record preserves failures to produce an output and failures to obtain a valid measurement. Treating either as absent data without explanation can inflate apparent reliability. A proposed method should state its unsupported conditions and test the boundary between accepted and rejected cases as carefully as the quality of its successful outputs.

Reproducibility begins with preserving which scientific objects are being linked. Stable identifiers and reusable metadata help connect inputs, reference observations, model versions, and reported outcomes (Wilkinson et al., 2016). The scoring pipeline also has dependencies that can change the interpretation of a result: preprocessing, representation checkpoints, thresholds, and aggregation code can all matter (Sculley et al., 2015). A compact release should retain those decisions rather than only the final mean. Where source data cannot be redistributed, an access description and hashes can still identify the analyzed inputs. This record allows later users to distinguish a change in the underlying system from a change in how the system was measured. Reproducibility is therefore part of interpretation, not simply a final packaging step.

5. Conclusion

A current image can be plausible while the interaction has lost an important state distinction. The proposed design tests that distinction through later consequences and states clearly which quantities are externally observed. It also separates a failed task outcome from an unverified explanation of the failure. Applying the design would require actual histories and outputs. Its contribution is a way to make hidden-state claims in neural rendering testable without granting visual realism more evidentiary force than it possesses.

References

Meng, Zijie, et al. "Make a game: A novel paradigm for interactive game rendering." *ICASSP 2026-2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2026.

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.

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.

Greenland, Sander, Judea Pearl, and James M. Robins. “Causal Diagrams for Epidemiologic Research.” *Epidemiology*, vol. 10, no. 1, 1999, pp. 37-48. https://doi.org/10.1097/00001648-199901000-00008.

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.

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.

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.

Wilkinson, Mark D., et al. "The FAIR Guiding Principles for Scientific Data Management and Stewardship." *Scientific Data*, vol. 3, 2016, article 160018. https://doi.org/10.1038/sdata.2016.18.

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

Bland, J. Martin, and Douglas G. Altman. "Statistical Methods for Assessing Agreement between Two Methods of Clinical Measurement." *The Lancet*, vol. 327, no. 8476, 1986, pp. 307-310. https://doi.org/10.1016/S0140-6736(86)90837-8.

Guo, Chuan, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. “On Calibration of Modern Neural Networks.” *Proceedings of the 34th International Conference on Machine Learning*, PMLR, vol. 70, 2017, pp. 1321-1330. https://proceedings.mlr.press/v70/guo17a.html.

Geifman, Yonatan, and Ran El-Yaniv. “Selective Classification for Deep Neural Networks.” *Advances in Neural Information Processing Systems*, vol. 30, 2017. https://arxiv.org/abs/1705.08500.

Ovadia, Yaniv, et al. “Can You Trust Your Model’s Uncertainty? Evaluating Predictive Uncertainty Under Dataset Shift.” *Advances in Neural Information Processing Systems*, vol. 32, 2019. https://arxiv.org/abs/1906.02530.

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.

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.

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

Brockman, Greg, et al.. "OpenAI Gym" *arXiv preprint arXiv:1606.01540*, 2016. https://arxiv.org/abs/1606.01540.

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

Hafner, Danijar, et al.. "Dream to Control: Learning Behaviors by Latent Imagination" *arXiv preprint arXiv:1912.01603*, 2019. https://arxiv.org/abs/1912.01603.

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

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.

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

Bruce, Jake, et al.. "Genie: Generative Interactive Environments" *arXiv preprint arXiv:2402.15391*, 2024. https://arxiv.org/abs/2402.15391.

Zhang, Richard, et al.. "The Unreasonable Effectiveness of Deep Features as a Perceptual Metric" *arXiv preprint arXiv:1801.03924*, 2018. https://arxiv.org/abs/1801.03924.

Unterthiner, Thomas, et al.. "Towards Accurate Generative Models of Video: A New Metric & Challenges" *arXiv preprint arXiv:1812.01717*, 2018. https://arxiv.org/abs/1812.01717.

Wang, Zhou, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. "Image Quality Assessment: From Error Visibility to Structural Similarity." *IEEE Transactions on Image Processing*, vol. 13, no. 4, 2004, pp. 600-612. https://doi.org/10.1109/TIP.2003.819861.

Schwarzer, Max, et al.. "Bigger, Better, Faster: Human-level Atari with human-level efficiency" *arXiv preprint arXiv:2305.19452*, 2023. https://arxiv.org/abs/2305.19452.

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

Levine, Sergey, Aviral Kumar, George Tucker, and Justin Fu. "Offline Reinforcement Learning: Tutorial, Review, and Perspectives on Open Problems." *arXiv*, 2020, arXiv:2005.01643. https://arxiv.org/abs/2005.01643.

Kumar, Aviral, Aurick Zhou, George Tucker, and Sergey Levine. "Conservative Q-Learning for Offline Reinforcement Learning." *Advances in Neural Information Processing Systems*, vol. 33, 2020, pp. 1179-1191. https://arxiv.org/abs/2006.04779.