# Constructing Meaningful Robot Assurances in the Age of Large Models

David Snyder
*University of Pennsylvania*

## I. Introduction

A critical, necessary condition of truly generalist robot policies is the capacity to operate robustly in environments that do not possess *a priori* known structure. Thus, general and high-dimensional exteroceptive perception is required to allow the robot to adapt to a wide array of possible contexts (e.g., variations in task, lighting, objects, and background). Filtering these observations to extract *relevant* information is a major challenge [40], with recent state-of-the-art (SOTA) methods growing rapidly in complexity [12, 20, 13]. As a result, perceptual generality closes off standard avenues for constructive analytical guarantees for robot policies.

Alternatively, characterizing improvements in robot performance via *empirical validation*, while more generally feasible, faces significant challenges in its own right. Data is expensive to collect for embodied systems [21], and proxy signals like physics-based simulators [39, 26] imperfectly model complicated phenomena (i.e., contact dynamics) [37, 36]. Additionally, at present there is not a general framework for complex validation tasks like policy ranking [6], active data collection [4], and analysis of within-trajectory behavioral signals [42].

The preceding considerations are a candidate basis for a practical theory of large models in robotics. Such theory can meaningfully accelerate the "design feedback loop" of empirical investigation (see Figure 1) and generalize to future, novel architectures. **My research directly addresses this problem by developing efficient procedures for evaluating robot behavior that provide rigorous, *policy-agnostic* guarantees under precise and practical models of uncertainty in the robot environment or evaluation domain.**

## II. Past and Current Work

The guarantees we henceforth consider are defined to be practical for robotics according to the following metrics: they are non-asymptotic (are valid in finite samples), explicitly account for and, where possible, minimize sample complexity, and make minimal assumptions about the structure of the perception-action pipeline.

**Time-Uniform Robust, Online Behavioral Guarantees.** The most robust guarantee of a robot's behavior gives a uniform statement of its efficacy for any realization of uncertainty in the environment. As such, these methods must generally consider worst-case notions of uncertainty, for example building from foundational results in robust control [16] to set-based methods like control-barrier functions [2, 3] and motion-planning results from Hamilton-Jacobi (HJ) reachability [8, 27]. Such

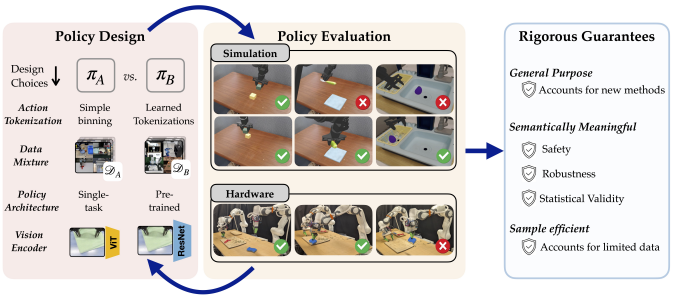


Fig. 1: **Design Feedback Loop**. Increasing model complexity promotes policy-agnostic and evaluation-centered assurances. My work accelerates the empirical design feedback loop by constructing rigorous evaluation methods.

a pipeline yields strong assurances – for example that the robot will remain safe under *any* (plausible) realization of uncertainty – at the cost of usually assuming highly accurate perception to map observations to a low-dimensional representation of the state. Further, necessary conservatism in the uncertainty model can require increased computational burden [14, 41, 31] and limit capacity to evaluate counterfactual robot policy choices. My work extends results in this domain to provide guarantees which: (1) hold counterfactually over the choice of policy [18]; (2) adapt to the nature of uncertainty in the environment [34]; and (3) explicitly account for unbounded perceptual noise [30].

These results are primarily obtained via the frameworks of online learning and regret minimization [19] and safe, anytime-valid inference (SAVI) [29]. Online learning introduces a novel flexibility to analytical methods in the low-dimensional state regime by constructing decision rules which can automatically adapt to the functional structures realized in the environment. My work in [18] constructs an adaptive, regret-minimizing evaluation test to synthesize worst-case perturbations for control tasks. This computationally efficient, policy-agnostic test provably recovers known optima, such as the game-theoretic $H_\infty$ control law [44], while accurately characterizing the robustness of general, possibly-learned controllers which deviate from these equilibria, potentially due to noisy perception. We extended the regret framework to motion planning in [34]. Here, we construct an optimization-based planner that adaptively interpolates robust and optimistic motion plans based upon the adversariality in the underlying dynamics and environment obstacles. In [30] (under review), we extend this to (semantic) perception-agnostic settings,

using SAVI procedures to give probabilistic guarantees for avoiding unsafe regions under noisy, long-horizon perception.

**Expected Performance Evaluation Guarantees.** As in [30], higher-level robotic tasks often require richer internal representations [32] possessing unknown dynamics. Such representations have limited interpretability, causing analytical guarantees to break down [10]. Therefore, constructing interpretable filters – e.g., for failure prediction [42] or out-of-distribution detection – generally requires additional learned components, with costly empirical evaluation necessary to evaluate the component-wise efficacy [21, 1].

My work addresses this challenge in two ways. First, it constructs data-driven probabilistic guarantees for policy-agnostic failure predictors [17] and sensor-agnostic signal detectors [23] that carefully reuse the training data to synthesize assurances, saving valuable resources in data collection. Second, it provides methods to incorporate proxy variables [7] and actively collect new training data [45] to accelerate estimation and learning while ensuring statistical robustness to overfitting. In each instance, the methods are general and adapt to arbitrarily complex policy architectures.

**Sequential Performance Evaluation.** Recently, there has been a profusion of challenging, dexterous, safety-critical, and long-horizon benchmark tasks across robotics disciplines [12, 38], reflecting the increased capabilities of novel complex and SOTA robot policies. In these cases, proxy methods like simulation can be harder to develop [38] and accumulate additional sources of noise. Given the analytic opacity of complex policy architectures, hardware evaluation thus remains the *gold-standard* metric to assess the quality of research innovations [21]. However, the time and equipment costs of hardware evaluation are amplified in this context due to long rollout times and complicated environment reset pipelines.

My work addresses this challenge via the introduction of novel *sequential* evaluation procedures to the robotics context. Indeed, despite the aforementioned costs of evaluation, the standard practice has been to choose a batch size (often arbitrarily), collect the full batch of (expensive) evaluations, and only then run statistical analyses. Put simply: when the batch is too big, effort is wasted; when it is too small, significance is lost. In [33], we constructed near-optimal sequential tests for comparing robot policy performance (e.g., between a novel policy and a baseline) under binary (success / failure) metrics. By explicitly optimizing for the expected sample size, the evaluator could save significant effort without invalidating statistical assurances. In [35] (under review), we extend these tests to arbitrary performance metrics, allowing for rigorous testing of partial credit measures, reinforcement learning (RL) episodic rewards, and behavioral measures of performance like jitter [9]. In each case, we achieve SOTA sample complexity for rigorous policy comparisons, accelerating the evaluator's capacity to obtain feedback signals for novel design innovations in the robot policy architecture. Importantly, these methods are again entirely agnostic to the (complex) robot policy, ensuring that they will generalize to subsequent improvements in the policy design feedback.

## III. Research Agenda

In each preceding instance, a core theme of our synthesized guarantees is ensuring minimal reliance on the internal structure of the robot policy. Instead, structure is absorbed, where possible, through environment dynamics [18, 34, 23, 30], evaluation procedure design [33, 35, 7, 45], or through added learning modules [17]. Each path is indicative of valuable directions of future inquiry.

**Guarantees Under Low-Dimensional Latent States.** There has been substantial interest in understanding low-dimensional latent states corresponding to safe or task-relevant representations in robotics [28, 24]. These states may correspond, for example, to the interaction of a (self-)attention mechanism with the high-dimensional perception modules [40]; this is a promising avenue to utilize existing capabilities as the perception-to-state 'bridge.' I intend to investigate the temporal coherence of such attentive latent states for long-horizon tasks, with the idea to use episodic regret bounds [19] to compose fixed-latent sub-tasks in complex and long-horizon scenarios.

**Unified, Data-Efficient Frameworks for Evaluation.** Empirical evaluation has received significant, renewed attention across robotics, machine learning, and RL [21, 1]; however, it remains an underexplored family of practical problems. As a concrete example, our previous work has advanced the SOTA in terms of policy comparison [33, 35] and the rigorous integration of proxy data [7]. However, as described in the introduction, problems of ranking [6], active evaluation [4], and detecting execution failure [42] and distribution shift, remain underexplored. Continuing our work on developing sequential procedures adapted to richer validation problems would unify the empirical evaluation literature as tailored for the highly data-constrained robotics context.

**Fundamental Limits for Robots and Evaluators.** In control theory, computational complexity theory, and statistics, frameworks to compare and classify problem difficulty can be rigorously defined and understood [5, 11]. However, such frameworks are at present strongly limited in robotics and adjacent fields [43, 25, 22]. The impacts of developing such frameworks are manifold, from better-standardized benchmarks to rigorously understanding the sample complexity of learning robotic tasks. For complex robot policies, I seek to develop bounds that specify task difficulty through the functional behavior and stability of the perception-to-state latent compression. In evaluation, I envision constructing guarantees of optimal sample efficiency (akin to results in learning for control [15]) which guide evaluation effort and set lower bounds on the trials necessary to establish rigorous decisions at provable levels of confidence. These results would additionally guide principled robot evaluation by determining when certain evaluation tasks are infeasible in the available evaluation budget. **Ultimately, my research agenda seeks to develop rich, usefully predictive theoretical frameworks which simultaneously illuminate functional behaviors of complex, learned policies and accelerate empirical developments within the field.**

## References


[1] Rishabh Agarwal, Max Schwarzer, Pablo Samuel Castro, Aaron C Courville, and Marc Bellemare. Deep reinforcement learning at the edge of the statistical precipice. volume 34, pages 29304–29320, 2021. doi: 10.48550/arXiv.2108.13264.

[2] Aaron D. Ames, Xiangru Xu, Jessy W. Grizzle, and Paulo Tabuada. Control Barrier Function Based Quadratic Programs for Safety Critical Systems. *IEEE Transactions on Automatic Control*, 62(8):3861–3876, August 2017. ISSN 1558-2523. doi: 10.1109/TAC.2016.2638961. URL https://ieeexplore.ieee.org/abstract/document/7782377.

[3] Aaron D. Ames, Samuel Coogan, Magnus Egerstedt, Gennaro Notomista, Koushil Sreenath, and Paulo Tabuada. Control Barrier Functions: Theory and Applications. In *2019 18th European Control Conference (ECC)*, pages 3420–3431, June 2019. doi: 10.23919/ECC.2019.8796030. URL https://ieeexplore.ieee.org/document/8796030.

[4] Abrar Anwar, Rohan Gupta, Zain Merchant, Sayan Ghosh, Willie Neiswanger, and Jesse Thomason. Efficient Evaluation of Multi-Task Robot Policies With Active Experiment Selection, February 2025. URL http://arxiv.org/abs/2502.09829. arXiv:2502.09829 [cs].

[5] Sanjeev Arora and Boaz Barak. *Computational Complexity: A Modern Approach*. Cambridge University Press, 2009.

[6] Pranav Atreya, Karl Pertsch, Tony Lee, Moo Jin Kim, Arhan Jain, Artur Kuramshin, Clemens Eppner, Cyrus Neary, Edward Hu, Fabio Ramos, et al. Roboarena: Distributed real-world evaluation of generalist robot policies. *arXiv preprint arXiv:2506.18123*, 2025.

[7] Apurva Badithela, David Snyder, Lihan Zha, Joseph Mikhail, Matthew O'Kelly, Anushri Dixit, and Anirudha Majumdar. Reliable and Scalable Robot Policy Evaluation with Imperfect Simulators, October 2025. URL http://arxiv.org/abs/2510.04354. arXiv:2510.04354 [cs].

[8] Somil Bansal, Mo Chen, Sylvia Herbert, and Claire J. Tomlin. Hamilton-Jacobi Reachability: A Brief Overview and Recent Advances, September 2017. URL http://arxiv.org/abs/1709.07523. arXiv:1709.07523 [eess].

[9] Yoav Beck, Talia Herman, Marina Brozgol, Nir Giladi, Anat Mirelman, and Jeffrey M. Hausdorff. SPARC: a new approach to quantifying gait smoothness in patients with Parkinson's disease. *Journal of NeuroEngineering and Rehabilitation*, 15(1):49, June 2018. ISSN 1743-0003. doi: 10.1186/s12984-018-0398-3.

[10] Leonard Bereska and Efstratios Gavves. Mechanistic Interpretability for AI Safety — A Review. *TMLR*, August 2024. URL https://openreview.net/forum?id=ePUVetPKu6.

[11] Peter J. Bickel and Kjell A. Doksum. *Mathematical Statistics: Basic Ideas and Selected Topics, Volumes I-II Package*. Chapman and Hall/CRC, New York, December 2015. ISBN 978-1-315-36926-6. doi: 10.1201/9781315369266.

[12] Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Lachy Groom, Karol Hausman, Brian Ichter, et al. $\pi_0$: A Vision-Language-Action Flow Model for General Robot Control. *arXiv preprint arXiv:2410.24164*, 2024. doi: 10.48550/arXiv.2410.24164.

[13] Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Xi Chen, Krzysztof Choromanski, Tianli Ding, Danny Driess, Avinava Dubey, Chelsea Finn, et al. Rt-2: Vision-language-action models transfer web knowledge to robotic control. *arXiv preprint arXiv:2307.15818*, 2023. doi: 10.48550/arXiv.2307.15818.

[14] Jérôme Darbon and Stanley Osher. Algorithms for overcoming the curse of dimensionality for certain Hamilton–Jacobi equations arising in control theory and elsewhere. *Research in the Mathematical Sciences*, 3 (1):19, September 2016. ISSN 2197-9847. doi: 10.1186/s40687-016-0068-7. URL https://doi.org/10.1186/s40687-016-0068-7.

[15] Sarah Dean, Horia Mania, Nikolai Matni, Benjamin Recht, and Stephen Tu. On the Sample Complexity of the Linear Quadratic Regulator, December 2018. URL http://arxiv.org/abs/1710.01688. arXiv:1710.01688 [math].

[16] Geir E. Dullerud and Fernando Paganini. *A Course in Robust Control Theory*, volume 36 of *Texts in Applied Mathematics*. Springer, New York, NY, 2000. ISBN 978-1-4419-3189-4 978-1-4757-3290-0. doi: 10.1007/978-1-4757-3290-0. URL http://link.springer.com/10.1007/978-1-4757-3290-0.

[17] Alec Farid, David Snyder, Allen Z. Ren, and Anirudha Majumdar. Failure Prediction with Statistical Guarantees for Vision-Based Robot Control. volume 18, June 2022. ISBN 978-0-9923747-8-5. URL https://www.roboticsproceedings.org/rss18/p042.html.

[18] Udaya Ghai, David Snyder, Anirudha Majumdar, and Elad Hazan. Generating Adversarial Disturbances for Controller Verification. In *Proceedings of the 3rd Conference on Learning for Dynamics and Control*, pages 1192–1204. PMLR, May 2021. URL https://proceedings.mlr.press/v144/ghai21a.html.

[19] Elad Hazan. *Introduction to Online Convex Optimization*. Adaptive Computation and Machine Learning series. MIT Press, London, 2nd edition, December 2021. URL http://arxiv.org/abs/1909.05207. arXiv: 1909.05207.

[20] Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Foster, Grace Lam, Pannag Sanketi, et al. OpenVLA: An Open-Source Vision-Language-Action Model. *arXiv preprint arXiv:2406.09246*, 2024. doi: 10.48550/arXiv.2406.09246.

[21] Hadas Kress-Gazit, Kunimatsu Hashimoto, Naveen Kuppuswamy, Paarth Shah, Phoebe Horgan, Gordon Richardson, Siyuan Feng, and Benjamin Burchfiel. Robot Learning as an Empirical Science: Best Practices for Policy Evaluation. In *Robotics: Science and Systems*, 2024. doi:

10.48550/arXiv.2409.09491.

[22] Arjun Krishna, Edward Hu, and Dinesh Jayaraman. The Value of Sensory Information to a Robot. *International Conference on Learning Representations*, 2025:90348–90376, May 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/e1126028d9f1f69c13571ec462084d31-Abstract-Conference.html.

[23] Eric Lepowsky, David Snyder, Alexander Glaser, and Anirudha Majumdar. Privacy-Preserving Map-Free Exploration for Confirming the Absence of a Radioactive Source. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 10073–10080, October 2024. doi: 10.1109/IROS58592.2024.10802428. URL https://ieeexplore.ieee.org/document/10802428.

[24] Paul Lutkus, Kaiyuan Wang, Lars Lindemann, and Stephen Tu. Latent Representations for Control Design with Provable Stability and Safety Guarantees, May 2025. URL http://arxiv.org/abs/2505.23210. arXiv:2505.23210 [eess].

[25] Anirudha Majumdar, Zhiting Mei, and Vincent Pacelli. Fundamental limits for sensor-based robot control. *The International Journal of Robotics Research*, 42(12):1051–1069, October 2023. ISSN 0278-3649. doi: 10.1177/02783649231190947. URL https://doi.org/10.1177/02783649231190947.

[26] Viktor Makoviychuk, Lukasz Wawrzyniak, Yunrong Guo, Michelle Lu, Kier Storey, Miles Macklin, David Hoeller, Nikita Rudin, Arthur Allshire, Ankur Handa, and Gavriel State. Isaac Gym: High Performance GPU-Based Physics Simulation For Robot Learning, August 2021. URL http://arxiv.org/abs/2108.10470. arXiv:2108.10470 [cs].

[27] Ian M Mitchell, Alexandre M Bayen, and Claire J Tomlin. https://ieeexplore.ieee.org/document/1463302A time-dependent Hamilton-Jacobi formulation of reachable sets for continuous dynamic games. In *IEEE Transactions on automatic control*, volume 50, pages 947–957. IEEE, 2005.

[28] Kensuke Nakamura, Lasse Peters, and Andrea Bajcsy. Generalizing Safety Beyond Collision-Avoidance via Latent-Space Reachability Analysis. volume 21, June 2025. ISBN 979-8-9902848-1-4. URL https://www.roboticsproceedings.org/rss21/p113.html.

[29] Aaditya Ramdas, Peter Grünwald, Vladimir Vovk, and Glenn Shafer. Game-theoretic statistics and safe anytime-valid inference. *Statistical Science*, 38(4):576–601, 2023.

[30] Zachary Ravichandran, David Snyder, Alexander Robey, Hamed Hassani, Vijay Kumar, and George J. Pappas. Contextual Safety Reasoning and Grounding for Open-World Robots, February 2026. URL http://arxiv.org/abs/2602.19983. arXiv:2602.19983 [cs].

[31] Alexander Robey, Haimin Hu, Lars Lindemann, Hanwen Zhang, Dimos V. Dimarogonas, Stephen Tu, and Nikolai Matni. Learning Control Barrier Functions from Expert Demonstrations. In *2020 59th IEEE Conference on Decision and Control (CDC)*, pages 3717–3724, December 2020. doi: 10.1109/CDC42340.2020.9303785. URL https://ieeexplore.ieee.org/document/9303785/. ISSN: 2576-2370.

[32] Ranjan Sapkota, Yang Cao, Konstantinos I. Roumeliotis, and Manoj Karkee. Vision-language-action models: Concepts, progress, applications and challenges, 2025. URL https://arxiv.org/abs/2505.04769.

[33] David Snyder, Asher James Hancock, Apurva Badithela, Emma Dixon, Patrick Miller, Rares Andrei Ambrus, Anirudha Majumdar, Masha Itkina, and Haruki Nishimura. Is Your Imitation Learning Policy Better than Mine? Policy Comparison with Near-Optimal Stopping. volume 21, Los Angeles, California. URL https://roboticsconference.org/program/papers/77/.

[34] David Snyder, Meghan Booker, Nathaniel Simon, Wenhan Xia, Daniel Suo, Elad Hazan, and Anirudha Majumdar. Online Learning for Obstacle Avoidance. In *Proceedings of The 7th Conference on Robot Learning*, pages 2926–2954. PMLR, December 2023. URL https://proceedings.mlr.press/v229/snyder23a.html.

[35] David Snyder, Apurva Badithela, Nikolai Matni, George Pappas, Anirudha Majumdar, Masha Itkina, and Haruki Nishimura. Beyond Binary Success: Sample-Efficient and Statistically Rigorous Robot Policy Comparison, March 2026. URL http://arxiv.org/abs/2603.13616. arXiv:2603.13616 [cs].

[36] Hyung Ju Suh, Max Simchowitz, Kaiqing Zhang, and Russ Tedrake. Do Differentiable Simulators Give Better Policy Gradients? In *Proceedings of the 39th International Conference on Machine Learning*, pages 20668–20696. PMLR, June 2022. URL https://proceedings.mlr.press/v162/suh22b.html.

[37] Hyung Ju Terry Suh, Tao Pang, and Russ Tedrake. Bundled Gradients Through Contact Via Randomized Smoothing. *IEEE Robotics and Automation Letters*, 7(2):4000–4007, April 2022. ISSN 2377-3766. doi: 10.1109/LRA.2022.3146931. URL https://ieeexplore.ieee.org/abstract/document/9697337.

[38] TRI LBM Team, Jose Barreiros, and et. al. A Careful Examination of Large Behavior Models for Multitask Dexterous Manipulation, July 2025. URL http://arxiv.org/abs/2507.05331. arXiv:2507.05331 [cs].

[39] Emanuel Todorov, Tom Erez, and Yuval Tassa. Mujoco: A physics engine for model-based control. In *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 5026–5033. IEEE, 2012. doi: 10.1109/IROS.2012.6386109.

[40] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is All you Need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://papers.nips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

[41] Kim P. Wabersich, Andrew J. Taylor, Jason J. Choi, Koushil Sreenath, Claire J. Tomlin, Aaron D. Ames, and Melanie N. Zeilinger. Data-Driven Safety Filters: Hamilton-Jacobi Reachability, Control Barrier Functions, and Predictive Methods for Uncertain Systems. *IEEE Control Systems Magazine*, 43(5):137–177, October 2023. ISSN 1941-000X. doi: 10.1109/MCS.2023.3291885. URL https://ieeexplore.ieee.org/document/10266799.

[42] Chen Xu, Tony Khuong Nguyen, Emma Dixon, Christopher Rodriguez, Patrick Miller, Robert Lee, Paarth Shah, Rares Ambrus, Haruki Nishimura, and Masha Itkina. Can We Detect Failures Without Failure Data? Uncertainty-Aware Runtime Failure Detection for Imitation Learning Policies, April 2025. URL http://arxiv.org/abs/2503.08558. arXiv:2503.08558 [cs].

[43] Yilun Xu, Shengjia Zhao, Jiaming Song, Russell Stewart, and Stefano Ermon. A THEORY OF USABLE INFORMATION UNDER COMPUTATIONAL CONSTRAINTS. page 24, 2020.

[44] G. Zames. Feedback and optimal sensitivity: Model reference transformations, multiplicative seminorms, and approximate inverses. *IEEE Transactions on Automatic Control*, 26(2):301–320, April 1981. ISSN 1558-2523. doi: 10.1109/TAC.1981.1102603. URL https://ieeexplore.ieee.org/document/1102603. Conference Name: IEEE Transactions on Automatic Control.

[45] Lihan Zha, Apurva Badithela, Michael Zhang, Justin Lidard, Jeremy Bao, Emily Zhou, David Snyder, Allen Z. Ren, Dhruv Shah, and Anirudha Majumdar. Guiding Data Collection via Factored Scaling Curves, May 2025. URL http://arxiv.org/abs/2505.07728. arXiv:2505.07728 [cs].