

Foundation World Models for Agents that Learn, Verify, and Adapt Reliably Beyond Static Environments

Blue Sky Ideas Track

Florent Delgrange

AI Lab, Vrije Universiteit Brussel & Flanders Make
Brussels, Belgium

ABSTRACT

The next generation of autonomous agents must not only learn efficiently but also act reliably and adapt their behavior in open worlds. Standard approaches typically assume fixed tasks and environments with little or no novelty, which limits world models' ability to support agents that must evolve their policies as conditions change. This paper outlines a vision for **foundation world models**: persistent, compositional representations that unify reinforcement learning, reactive/program synthesis, and abstraction mechanisms. We propose an agenda built around four components: (i) learnable reward models from specifications to support optimization with clear objectives; (ii) adaptive formal verification integrated throughout learning; (iii) online abstraction calibration to quantify the reliability of the model's predictions; and (iv) test-time synthesis and world-model generation guided by verifiers. Together, these components enable agents to synthesize verifiable programs, derive new policies from a small number of interactions, and maintain correctness while adapting to novelty. The resulting framework positions foundation world models as a substrate for learning, reasoning, and adaptation, laying the groundwork for agents that not only act well but can explain and justify the behavior they adopt.

KEYWORDS

Reinforcement Learning, Reactive Synthesis, Formal Verification, Foundation Models, World Models, Abstraction, Adaptive Agents

ACM Reference Format:

Florent Delgrange. 2026. Foundation World Models for Agents that Learn, Verify, and Adapt Reliably Beyond Static Environments : Blue Sky Ideas Track. In *Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), Paphos, Cyprus, May 25 – 29, 2026*, IFAAMAS, 7 pages. <https://doi.org/10.65109/WCEI7331>

1 INTRODUCTION

In the last decade, *reinforcement learning* (RL [61]) has achieved remarkable empirical progress, from surpassing human performance in high-dimensional games [36, 37, 50] to mastering complex real-world control tasks such as energy management, robotics, epidemics mitigation [24, 35, 57, 68]. Beyond control and decision-making, RL has also become central to the post-training of large language

models (LLMs). Techniques such as RL from human feedback, verifiable rewards, and other variants [21, 46, 48, 51] have turned RL into a general framework for aligning agents with qualitative goals such as coherence, safety, or reasoning depth. These advances, however, have largely relied on reward maximization and massive data collection, with little understanding of the structure underlying the learned agent's behaviors. Aligning agents' intended behaviors with designed or learned reward models remains deeply challenging: it is prone to reward hacking, fragile to misspecification, and often demands extensive engineering. In addition, modern approaches couple RL with deep neural networks to scale to large domains. As a result, learning guarantees no longer hold, policies may drift from the designer's intent, and fail to respect safety constraints. Despite numerous efforts to regularize or constrain learning [4, 10, 23, 27, 41–43, 70, 76, 77], the ability to guarantee correctness remains fundamentally elusive. The question is no longer whether agents learn efficiently, but whether they learn reliably.

In contrast, *reactive synthesis* [11, 55] offers a complementary paradigm rooted in formal methods. Given a model of the environment (e.g., a *Markov process* [56]) and a logical specification of the desired behavior, synthesis algorithms construct control policies that are correct by design. These methods guarantee that the induced agent's behaviors satisfy the specification, providing the transparency and guarantees missing from most learning-based systems. Yet, their scope remains limited: the need for explicit, finite models of the environment and the cost of exhaustively exploring large state spaces to construct guaranteed policies render them impractical for the open-ended, general, uncertain domains where learning excels. The two fields thus stand as mirror images: one efficient, scalable but ungrounded, the other reliable but rigid.

Bridging these paradigms is a defining challenge for the future of the (Multi-)Agent Systems community: in many multi-agent domains, ranging from marketplaces to autonomous mobility, agents must adapt to uncertainty while still satisfying global coordination and safety constraints. Purely reward-driven multi-agent RL often yields unstable or undesired behaviors, while purely synthesized policies rely on fixed models and rigid interaction assumptions [20, 75]. Recent integrations of multi-agent learning with temporal-logic objectives [38, 39] highlight the need for methods that ensure adaptation and correctness. A unified perspective would support flexible learning of interaction patterns while maintaining verifiable guarantees over joint behavior in open multi-agent systems.

Beyond theoretical convergence, current AI trends make guaranteed policy synthesis urgent. The rise of **foundation models** and agentic AI frameworks has renewed interest in systems that reason over goals, tools, and environments expressed in natural language. First, adaptive and unsupervised RL highlight the need



This work is licensed under a Creative Commons Attribution International 4.0 License.

Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), C. Amato, L. Dennis, V. Mascardi, J. Thangarajah (eds.), May 25 – 29, 2026, Paphos, Cyprus. © 2026 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org). <https://doi.org/10.65109/WCEI7331>

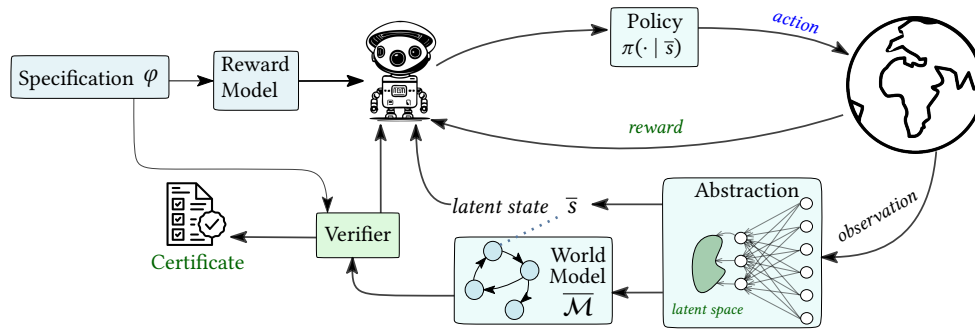


Figure 1: Given a *specification* φ formalizing the intended agent’s behavior,¹ a *reward model* is automatically converted/generated from φ . Simultaneously with optimizing the return, the agent learns a representation of the observations through the state space of a *verifiable world model*. Consequently, the policy is learned directly on this representation. To provide the guarantees, the world model is processed through a verifier (e.g., a model-checker). The verifier’s output *guides* the agent or *corrects* the policy as needed. At any time, the verifier can return a *certificate* on both specification satisfaction and the abstraction quality of the world model. Note that we do not assume to have access to the explicit environment dynamics (simulating it is sufficient).

for representations that persist across tasks and agent populations, rather than being retrained from scratch [1, 2, 52, 59, 63, 65]. In this context, the “train-then-verify” paradigm [8, 9, 25, 28, 29], in which a policy is first learned and later verified for compliance, remains conceptually unsatisfactory and computationally brittle in *dynamic environments, evolving goals, and distributed settings* where **agents must continuously adapt while preserving constraints**. Second, LLMs already exhibit rudimentary forms of reactive behavior: they interpret instructions, generate sequences of actions, and adapt decisions from interaction feedback [71]. Yet, they lack the formal backbone needed for correctness and long-term reliability. Integrating LLM-based reasoning with formal verification — by treating language as a medium for generating, refining, and testing specifications — offers a route toward interpretable and verifiable intelligence. For instance, LLMs can act as specification refiners, translating high-level objectives into temporal-logic constraints or decomposing tasks into verifiable subtasks (e.g., [15]).

The vision: learning verifiable world models. We argue that the next generation of intelligent agents should **learn verifiable world models**: internal representations that are not only **predictive** but also **formally analyzable**. Learning and verification become intertwined: as agents interact with their environment, they refine a predictive model while building a symbolic abstraction on which reasoning and verification can operate. This joint structure enables efficient planning, online guarantees, and the ability to **detect, quantify, and correct** deviations from intended specifications. In this sense, it bridges empirical learning and formal reasoning and points toward reliable, adaptive, and self-improving (multi-)agent systems.

This vision builds upon emerging evidence that such integration is feasible. Recent research demonstrates that world models can be learned with measurable abstraction error [28, 29], that **deep** model-based RL can exploit learned, local dynamics to achieve guaranteed safe policy improvement [26], and that learned representations can be lifted to the symbolic level for compositional synthesis [15, 27, 53, 77]. These developments suggest that RL and reactive synthesis

¹ φ can be provided by the user or be *generated* from experiences by an LLM.

need not be antagonistic paradigms but can form the two halves of a coherent framework for reliable autonomy.

2 LEARNING VERIFIABLE WORLD MODELS

Achieving reliability and adaptability in autonomous agents demands that learning, verification, and reasoning become parts of a single closed loop. This direction is promising and aligns with recent agendas on safe AI [22] and with industrial interest in runtime verification for machine learning systems [30]. We envision agents that continuously **learn, verify, and refine** their own **internal models of the world**. These self-verifying agents will not rely on externally imposed models or post-hoc safety checks; instead, they will build world models that are formally analyzable as they learn them, preserving guarantees of correctness and safety throughout their evolution. *The general framework is depicted in Fig. 1.*

At the core of this vision lies the concept of *RL-Synthesizers*: agents that treat learning itself as a synthesis problem. Each interaction cycle with the environment will jointly improve a predictive model, strengthen a symbolic abstraction of that model, and test compliance with a formal specification. The boundary between learning to act and reasoning about action will vanish: logical reasoning becomes part of the learning process itself, guiding policy updates rather than evaluating them afterward.

As a concrete example, consider a package-delivery agent in a dynamic warehouse: it must deliver a package while avoiding collisions with workers and robots. In the following, we will use this thread to illustrate the components illustrated in Fig. 1.

Learning from Formalized Rewards. Reward design remains a persistent source of brittleness in RL. Translating human intent into scalar feedback often requires extensive tuning and rarely ensures that the resulting policy reflects the intended objective. A promising path is to replace manual shaping with formally specified *reward models*: functions derived from logical or programmatic task descriptions that retain semantic clarity and verifiability.

Temporal-logic–based translations, such as those introduced in [14] and other approaches incorporating LTL reasoning [54] into RL [5, 13, 58, 64], already point in this direction. That way, in the

package delivery thread, a property such as “eventually deliver while always avoiding collisions” can be translated into a reward model, aligning optimization and specification satisfaction. Yet most approaches emphasize expressiveness over statistical learnability: in general, learning with traditional LTL objective is intractable [12, 69]. Learnable alternatives leverage reward machines but lead to extremely large structures [6], impeding formal verification. Recent studies [25] have shown that certain logics, for example, the discounted logic of De Alfaro et al. [3], retain much of LTL’s expressiveness while being computably efficient and *PAC-learnable*. Despite these advantages, such learnable fragments have been largely overlooked in contemporary RL research. Revisiting them is timely: when coupled with world-model abstraction, they enable differentiable and compact reward formulations whose optimization directly aligns with satisfaction of the underlying specification.

Together, these developments suggest that future agents should treat reward generation itself as a formal synthesis problem: combining learnable logics, programmatic representations, and language-based refinement to ensure that optimization and verification operate on the same semantic ground.

Verification During Learning. Recent advances have increasingly blurred the boundary between deep learning and verification. Techniques such as *probabilistic shielding* [4, 23, 42, 45, 70] and the learning of *neural certificates* [10, 34, 41, 49, 73, 76, 77] already embed formal reasoning into the training process. They aim to ensure safety or specification satisfaction as the agent learns, rather than verifying a static policy post hoc. Yet these methods still rest on strong assumptions: they either presuppose access to an accurate, explicit model of the safety aspects of the environment, require the full dynamics to be known and differentiable, or do not scale to large domains. In open-ended settings, such assumptions collapse, limiting applicability precisely where guarantees are most needed.

A promising direction is to treat verification not as a one-time property of a model but as a continuous process of **calibration** between the learned world model and the real environment. As the agent updates its internal model, it simultaneously evaluates how the satisfaction of the specification evolves under model uncertainty. This allows the agent to **track and quantify** the confidence in its own guarantees, yielding a form of **runtime formal reasoning**. When the satisfaction margin falls below a safe threshold, the verifier can steer exploration toward critical regions, trigger additional data collection, or refine the world model, effectively transforming guarantees into **adaptive control signals**. In the delivery example, if traffic patterns shift and the safety margin drops, verifier feedback can reject risky updates and redirect exploration.

A concrete route to verification during learning may build on the theory of *Safe Policy Improvement* (SPI [16, 33, 47, 60, 62, 67]), which guarantees that new policies do not degrade performance relative to a baseline. Intuitively, SPI maintains a set of plausible world models consistent with the collected data and accepts a policy update only when improvement can be certified under *all* such models. This makes SPI a canonical instance of learning and formal reasoning co-evolving. While effective in offline and tabular settings, these methods require extensive data collection and do not scale to general spaces. Extending this to online, more general settings, deep RL agents could maintain such safety guarantees

as their models evolve. While this extension is non-trivial, recent works enable such a deep analogue of SPI theorems [26], but do not offer practical algorithms explicitly leveraging the world model to act; doing so would allow the agent to reason beyond the distribution of experiences on which the model was trained, which is a clear direction toward broader adaptability.

Such mechanisms point toward self-calibrating verification, where reasoning adapts jointly with learning. The verifier is no longer an external auditor but a co-evolving module that constrains policy updates according to the learned model’s reliability. This turns verification into an active feedback mechanism: a continuous dialogue between learning, uncertainty, and formal correctness.

Abstraction and World-Model Calibration. At the foundation of any verifiable learning process lies the question of representation: *how to model the environment compactly without losing the behavioral properties that matter for reasoning*. For agents that learn and verify, abstraction is not a preprocessing step but a continuously maintained contract between the learned world model and the true environment. The reliability of any verification or synthesis component ultimately depends on how well this contract is calibrated.

Traditional abstraction mechanisms, such as *bisimulation metrics* [31], quantify *behavioral equivalence* by bounding discrepancies between transition probabilities and rewards. In modern RL, these metrics have been extended to learned latent representations [7, 17, 32, 66, 74], enabling approximate verification of neural policies and symbolic reasoning over compressed dynamics [28, 29]. Yet current estimators of abstraction error are conservative and often static, assuming fixed data distributions or fully explored state spaces. As policies evolve, such assumptions no longer hold: the abstraction must remain mathematically valid while drifting away from the parts of the environment the agent actually visits. For example, after training on the main corridor, the RL agent may identify a rarely visited shortcut as a high-reward choice. The world model predicts it as safe, but unseen forklift traffic can violate collision avoidance unless the verifier marks that route as uncertified.

A crucial step forward is to make abstraction certificates **adaptive**: updated online as the agent collects data and revises its world model. This involves estimating local abstraction error, uncertainty, or divergence and feeding this information back to the verifier and planner. The resulting coverage-calibrated world models would provide a principled basis for determining when model-based reasoning can be trusted. Within this framework, abstraction bounds act as dynamic confidence radii that govern rollout budgets, planning horizons, or the activation of verification routines.

3 FOUNDATION WORLD MODELS

Adaptive abstraction opens the door to a broader goal: *building persistent, general-purpose world models for open environments*:

We see **foundation world models** as **persistent, structured** representations of open environments that unify **dynamics, abstractions, and semantic knowledge**. They act as **reusable** substrates that can be queried **zero- or few-shot** to **synthesize or adapt policies, verify** specifications, and **guide behavior in unseen situations**.

Such models would capture not only predictive dynamics but also the logical and relational structure of the environment: how local behaviors compose, what invariants hold, and under which abstractions guarantees remain valid. The objective is to move from isolated models learned per task toward persistent, verifiable priors that generalize across tasks and domains.

A foundation world model differs from conventional predictive models in three ways. First, *it integrates abstraction and calibration* as first-class components: each learned representation carries an explicit reliability measure, linking model accuracy to formal guarantees. Second, *it encodes compositional structure*, allowing previously learned fragments (such as verified local controllers or programmatic dynamics) to be assembled into novel behaviors. Third, *it supports semantic querying*: by conditioning on formal or linguistic task descriptions, the model can generate consistent world abstractions or policy priors for unseen objectives.

Recent research in *unsupervised RL* [52, 59, 65] and *compositional synthesis* [15, 27, 53, 77] provide evidence that these properties are attainable. Metric-aware representation learning suggests that temporal distance and behavioral similarity can guide the formation of transferable abstractions; modular or programmatic world-model frameworks show that local dynamics can be encoded as reusable components; and hierarchical control architectures demonstrate how verified local policies can be synthesized into globally correct strategies. Together, these advances draw a path toward foundation world models that unify generalization, structure, and verifiability.

Such models underpin the integration of learning and verification. They provide an analyzable substrate on which logical specifications can be generated, verified, and composed, enabling agents to acquire new competencies without retraining by reasoning over the verifiable components already contained in their world model.

LLMs as Specification Refiners. Even the most robust foundation world model cannot anticipate every new configuration of an open world. When confronted with novel dynamics or goals, agents must rapidly form structured hypotheses about how the world behaves and how their objectives should be reformulated [18, 19, 72]. Both classical RL and LLM-based agents struggle in this setting: fast **adaptation** often requires retraining, and newly encountered constraints must be rediscovered from scratch. What is missing is a mechanism for learning world models and specifications **at test time**, grounded in formal reasoning and prior knowledge.

We propose an interactive refinement loop in which an LLM collaborates with a verifier to construct and validate new world models on the fly. This loop enables agents to synthesize verifiable programs and derive new policies after only a few interactions in an unseen environment.

- (1) *Specification and decomposition.* When the agent encounters unseen environment’s regions, the LLM proposes candidate task specifications or decompositions (e.g., temporal-logic formulas or automata) derived from human instructions or from the agent’s initial exploratory trajectories.
- (2) *Program generation.* The observations then serve as a basis for the LLM to generate *verifiable* programs in a formal modeling language (e.g., PRISM [44]) that can be processed by

probabilistic model checkers (e.g., STORM [40]). **These programs act as world models**, capturing hypothesized causal or temporal relationships observed in the new environment.

- (3) *Formal verification.* The verifier checks the validity of the generated programs against its specifications and optionally reports counterexamples or structural inconsistencies.
- (4) *Revision and synthesis.* The LLM integrates this feedback to refine the program or propose new sub-tasks. The world model-programs and subtasks can then be used for planning, RL, or reactive synthesis to derive “low-level” policies or exploratory behaviors in the newly encountered domain.
- (5) *Repeat.* Execute a low-level policy for a specific sub-task, collect new experiences, and repeat from step (1).

This process mirrors human reasoning: the agent alternates between hypothesis generation, verification, and empirical testing. Successive refinements would yield world models that are both compact and verifiable, enabling fast adaptation without retraining. In the package-delivery example, a blocked corridor invalidates part of the original specification; the LLM thus refines that specification accordingly, regenerates the corresponding program, and submits it for formal verification. Then, the resulting program can be used as a world model to guide the policy to avoid the blocked corridor.

Importantly, the verifier’s feedback can also serve as a *formally verified reward signal*. Thus, this reward may be used to *reinforce the LLM’s capacity* to produce consistent, valid, and executable programs in RL post-training. This closes the learning loop: successful verifications provide positive reinforcement, while counterexamples penalize inconsistent hypotheses.

This mechanism enhances adaptability by letting agents synthesize new abstractions and policies in situ, guided by logical structure and formal feedback. Yet, this interactive loop remains challenging: generated programs may be inconsistent, verification can bottleneck repeated refinement, and abstraction or specification drift can amplify errors; robust deployment therefore requires stable, efficient inconsistency detection and feedback integration.

4 CONCLUSION

The trajectory of AI agents has long oscillated between empirical efficiency and formal correctness. RL yields increasingly competent agents, yet their behavior can be difficult to trust and interpret, whereas formal methods offer guarantees but depend on fixed models. Our vision reconciles these traditions by shifting what agents should learn: not only a policy, but a verifiable understanding of the world.

Our agenda rests on four components: formalized rewards tie optimization to intended task semantics; verification during learning turns correctness into an adaptive control signal; abstraction and world-model calibration quantify when model-based reasoning can be trusted; and LLM-based refinement supports test-time synthesis of specifications and programs in open environments.

Together, these components define foundation world models: persistent, compositional, and verifiable representations that unify perception, reasoning, and formal analysis. For the (Multi-)Agent Systems community, this shifts the goal from optimizing behavior in fixed domains to designing architectures that construct and certify their own domains of competence.

ACKNOWLEDGMENTS

This work was realized under F. Delgrange’s VUB OZR mandate (VUB-OZR4417) and was supported by the “DESCARTES” iBOF project. We are grateful to Ann Nowé, Guillermo A. Pérez, Willem Röpke, and Pieter Libin for insightful discussions and thoughtful feedback that helped shape this vision and research agenda.

REFERENCES

- [1] Siddhant Agarwal, Harshit Sikchi, Peter Stone, and Amy Zhang. 2025. Proto Successor Measure: Representing the Behavior Space of an RL Agent. In *Forty-second International Conference on Machine Learning, ICML 2025, Vancouver, BC, Canada, July 13-19, 2025*. OpenReview.net. <https://openreview.net/forum?id=mUDnPzopZF>
- [2] Lucas Nunes Alegre, Ana L. C. Bazzan, and Bruno C. da Silva. 2022. Optimistic Linear Support and Successor Features as a Basis for Optimal Policy Transfer. In *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA (Proceedings of Machine Learning Research, Vol. 162)*, Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvári, Gang Niu, and Sivan Sabato (Eds.). PMLR, 394–413. <https://proceedings.mlr.press/v162/alegre22a.html>
- [3] Luca de Alfaro, Thomas A. Henzinger, and Rupak Majumdar. 2003. Discounting the Future in Systems Theory. In *Automata, Languages and Programming, 30th International Colloquium, ICALP 2003, Eindhoven, The Netherlands, June 30 - July 4, 2003. Proceedings (Lecture Notes in Computer Science, Vol. 2719)*, Jos C. M. Baeten, Jan Karel Lenstra, Joachim Parrow, and Gerhard J. Woeginger (Eds.). Springer, 1022–1037. https://doi.org/10.1007/3-540-45061-0_79
- [4] Mohammed Alshiekh, Roderick Bloem, Rüdiger Ehlers, Bettina Könighofer, Scott Niekum, and Ufuk Topcu. 2018. Safe Reinforcement Learning via Shielding. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, Sheila A. McIlraith and Kilian Q. Weinberger (Eds.). AAAI Press, 2669–2678. <https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/17211>
- [5] Rajeev Alur, Suguman Bansal, Osbert Bastani, and Kishor Jothimurugan. 2022. A Framework for Transforming Specifications in Reinforcement Learning. In *Principles of Systems Design - Essays Dedicated to Thomas A. Henzinger on the Occasion of His 60th Birthday*, Vol. 13660. Springer, 604–624. https://doi.org/10.1007/978-3-031-22337-2_29
- [6] Rajeev Alur, Osbert Bastani, Kishor Jothimurugan, Mateo Perez, Fabio Somenzi, and Ashutosh Trivedi. 2023. Policy Synthesis and Reinforcement Learning for Discounted LTL. In *Computer Aided Verification - 35th International Conference, CAV 2023, Paris, France, July 17-22, 2023, Proceedings, Part I (Lecture Notes in Computer Science, Vol. 13964)*, Constantin Enea and Akash Lal (Eds.). Springer, 415–435. https://doi.org/10.1007/978-3-031-37706-8_21
- [7] Raphaël Avalos, Florent Delgrange, Ann Nowe, Guillermo A. Perez, and Diederik M. Roijers. 2024. The Wasserstein Believer: Learning Belief Updates for Partially Observable Environments through Reliable Latent Space Models. In *The Twelfth International Conference on Learning Representations*. <https://openreview.net/forum?id=KrtGTGaGe>
- [8] Edoardo Bacci and David Parker. 2020. Probabilistic Guarantees for Safe Deep Reinforcement Learning. In *Formal Modeling and Analysis of Timed Systems - 18th International Conference, FORMATS 2020, Vienna, Austria, September 1-3, 2020, Proceedings (LNCS, Vol. 12288)*, Nathalie Bertrand and Nils Jansen (Eds.). Springer, 231–248. https://doi.org/10.1007/978-3-030-57628-8_14
- [9] Edoardo Bacci and David Parker. 2022. Verified Probabilistic Policies for Deep Reinforcement Learning. In *NASA Formal Methods, Jyotirmoy V. Deshmukh, Klaus Havelund, and Ivan Perez (Eds.)*. Springer International Publishing, Cham, 193–212.
- [10] Thom S. Badings, Wietze Koops, Sebastian Junges, and Nils Jansen. 2025. Policy Verification in Stochastic Dynamical Systems Using Logarithmic Neural Certificates. In *Computer Aided Verification - 37th International Conference, CAV 2025, Zagreb, Croatia, July 23-25, 2025, Proceedings, Part II (Lecture Notes in Computer Science, Vol. 15932)*, Ruzica Piskac and Zvonimir Rakamaric (Eds.). Springer, 349–375. https://doi.org/10.1007/978-3-031-98679-6_16
- [11] Christel Baier, Luca de Alfaro, Vojtech Forejt, and Marta Kwiatkowska. 2018. Model Checking Probabilistic Systems. In *Handbook of Model Checking*, Edmund M. Clarke, Thomas A. Henzinger, Helmut Veith, and Roderick Bloem (Eds.). Springer, 963–999. https://doi.org/10.1007/978-3-319-10575-8_28
- [12] Hugo Bazille, Blaise Genest, Cyrille Jégourel, and Jun Sun. 2020. Global PAC Bounds for Learning Discrete Time Markov Chains. In *Computer Aided Verification - 32nd International Conference, CAV 2020, Los Angeles, CA, USA, July 21-24, 2020, Proceedings, Part II (Lecture Notes in Computer Science, Vol. 12225)*, Shuvendu K. Lahiri and Chao Wang (Eds.). Springer, 304–326. https://doi.org/10.1007/978-3-030-53291-8_17
- [13] Alper Kamil Bozkurt, Yu Wang, Michael M. Zavlanos, and Miroslav Pajic. 2020. Control Synthesis from Linear Temporal Logic Specifications using Model-Free Reinforcement Learning. In *2020 IEEE International Conference on Robotics and Automation, ICRA 2020, Paris, France, May 31 - August 31, 2020*. IEEE, 10349–10355. <https://doi.org/10.1109/ICRA40945.2020.9196796>
- [14] Alberto Camacho, Rodrigo Toro Icarte, Toryn Q. Klassen, Richard Anthony Valenzano, and Sheila A. McIlraith. 2019. LTL and Beyond: Formal Languages for Reward Function Specification in Reinforcement Learning. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019*, Sarit Kraus (Ed.). ijcai.org, 6065–6073. <https://doi.org/10.24963/IJCAI.2019/840>
- [15] Roger Creus Castanyer, Faisal Mohamed, Pablo Samuel Castro, Cyrus Neary, and Glen Berseth. 2025. ARM-FM: Automated Reward Machines via Foundation Models for Compositional Reinforcement Learning. <https://arxiv.org/abs/2510.14176>
- [16] Alberto Castellini, Federico Bianchi, Edoardo Zorzi, Thiago D. Simão, Alessandro Farinelli, and Matthijs T. J. Spaan. 2023. Scalable Safe Policy Improvement via Monte Carlo Tree Search. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA (Proceedings of Machine Learning Research, Vol. 202)*, Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (Eds.). PMLR, 3732–3756. <https://proceedings.mlr.press/v202/castellini23a.html>
- [17] Pablo Samuel Castro, Tyler Kastner, Prakash Panangaden, and Mark Rowland. 2021. MICO: Improved representations via sampling-based state similarity for Markov decision processes. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, Marc’Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan (Eds.). 30113–30126. <https://proceedings.neurips.cc/paper/2021/hash/fd06b8ea02fe5b1c2496fe1700e9d16c-Abstract.html>
- [18] François Chollet. 2019. On the Measure of Intelligence. *CoRR* abs/1911.01547 (2019). <http://arxiv.org/abs/1911.01547>
- [19] François Chollet, Mike Knoop, Gregory Kamrath, Bryan Landers, and Henry Pinkard. 2025. ARC-AGI-2: A New Challenge for Frontier AI Reasoning Systems. *CoRR* abs/2505.11831 (2025). <https://doi.org/10.48550/ARXIV.2505.11831> arXiv: 2505.11831
- [20] Amit K. Chopra, Leendert van der Torre, H. Verhagen, and Serena Villata. 2018. Handbook of Normative Multiagent Systems. <https://api.semanticscholar.org/CorpusID:203708095>
- [21] Paul F. Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep Reinforcement Learning from Human Preferences. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett (Eds.). 4299–4307. <https://proceedings.neurips.cc/paper/2017/hash/d5e2c0adad503c91f91df24d0cd4e49-Abstract.html>
- [22] David Dalrymple, Joar Skalse, Yoshua Bengio, Stuart Russell, Max Tegmark, Sanjit Seshia, Steve Omohundro, Christian Szegedy, Ben Goldhaber, Nora Ammann, Alessandro Abate, Joe Halpern, Clark W. Barrett, Ding Zhao, Tan Zhi-Xuan, Jeannette Wing, and Joshua B. Tenenbaum. 2024. Towards Guaranteed Safe AI: A Framework for Ensuring Robust and Reliable AI Systems. *CoRR* abs/2405.06624 (2024). <https://doi.org/10.48550/ARXIV.2405.06624> arXiv: 2405.06624
- [23] David Debot, Gabriele Venturato, Giuseppe Marra, and Luc De Raedt. 2025. Neurosymbolic Reinforcement Learning: Playing MiniHack with Probabilistic Logic Shields. In *AAAI-25, Sponsored by the Association for the Advancement of Artificial Intelligence, February 25 - March 4, 2025, Philadelphia, PA, USA*, Toby Walsh, Julie Shah, and Zico Kolter (Eds.). AAAI Press, 29631–29633. <https://doi.org/10.1609/AAAI.V39I28.35349>
- [24] Jonas Degraeve, Federico Felici, Jonas Buchli, Michael Neunert, Brendan D. Tracey, Francesco Carpanese, Timo Ewalds, Roland Hafner, Abbas Abdolmaleki, Diego de Las Casas, Craig Donner, Leslie Fritz, Cristian Galperti, Andrea Huber, James Keeling, Maria Tsimpoukelli, Jackie Kay, Antoine Merle, Jean-Marc Moret, Seb Noury, Federico Pesamosca, David Pfau, Olivier Sauter, Cristian Sommariva, Stefano Coda, Basil Duval, Ambrogio Fasoli, Pushmeet Kohli, Koray Kavukcuoglu, Demis Hassabis, and Martin A. Riedmiller. 2022. Magnetic control of tokamak plasmas through deep reinforcement learning. *Nat.* 602, 7897 (2022), 414–419. <https://doi.org/10.1038/S41586-021-04301-9>
- [25] F. Delgrange. 2024. *Activating Formal Verification of Deep Reinforcement Learning Policies by Model Checking Bisimilar Latent Space Models: Dissertation*. <https://books.google.be/books?id=PwHv0AEACAAJ>
- [26] Florent Delgrange, Raphaël Avalos, and Willem Röpke. 2026. Deep SPI: Safe Policy Improvement via World Models. In *The Fourteenth International Conference on Learning Representations*. <https://openreview.net/forum?id=24C3bSaH3F>
- [27] Florent Delgrange, Guy Avni, Anna Lukina, Christian Schilling, Ann Nowe, and Guillermo Perez. 2025. Composing Reinforcement Learning Policies, with Formal Guarantees. In *Proceedings of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2025), Detroit, Michigan, USA, May 19-23, 2025, IFAAMAS*.

- [28] Florent Delgrange, Ann Nowé, and Guillermo A. Pérez. 2022. Distillation of RL Policies with Formal Guarantees via Variational Abstraction of Markov Decision Processes. *Proceedings of the AAAI Conference on Artificial Intelligence* 36, 6 (June 2022), 6497–6505. <https://doi.org/10.1609/aaai.v36i6.20602>
- [29] Florent Delgrange, Ann Nowé, and Guillermo A. Pérez. 2023. Wasserstein Auto-encoded MDPs: Formal Verification of Efficiently Distilled RL Policies with Many-sided Guarantees. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net. <https://openreview.net/pdf?id=jLLTtEdh1ZY>
- [30] Hervé Delseny, Christophe Gabreau, Adrien Gaufriau, Bernard Beaudouin, Ludovic Ponsolle, Lucian Alecu, Hugues Bonnin, Brice Beltran, Didier Duchel, Jean-Brice Ginestet, Alexandre Hervieu, Ghilaine Martinez, Sylvain Pasquet, Kevin Delmas, Claire Pagetti, Jean-Marc Gabriel, Camille Chapdelaine, Sylvaine Picard, Mathieu Damour, Cyril Cappi, Laurent Gardès, Florence de Grancey, Eric Jenn, Baptiste Lefèvre, Grégory Flandin, Sébastien Gerchinovitz, Franck Mamalet, and Alexandre Albore. 2021. White Paper Machine Learning in Certified Systems. *CoRR* abs/2103.10529 (2021). <http://arxiv.org/abs/2103.10529> arXiv: 2103.10529.
- [31] José Desharnais, Vineet Gupta, Radha Jagadeesan, and Prakash Panangaden. 2004. Metrics for labelled Markov processes. *Theor. Comput. Sci.* 318, 3 (2004), 323–354. <https://doi.org/10.1016/J.TCS.2003.09.013>
- [32] Carles Gelada, Saurabh Kumar, Jacob Buckman, Ofir Nachum, and Marc G. Bellemare. 2019. DeepMDP: Learning Continuous Latent Space Models for Representation Learning. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA (Proceedings of Machine Learning Research, Vol. 97)*, Kamalika Chaudhuri and Ruslan Salakhutdinov (Eds.). PMLR, 2170–2179. <http://proceedings.mlr.press/v97/gelada19a.html>
- [33] Mohammad Ghavamzadeh, Marek Petrik, and Yinlam Chow. 2016. Safe Policy Improvement by Minimizing Robust Baseline Regret. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, Daniel D. Lee, Masashi Sugiyama, Ulrike von Luxburg, Isabelle Guyon, and Roman Garnett (Eds.), 2298–2306. <https://proceedings.neurips.cc/paper/2016/hash/9a3d458322d70046f63dfd8b0153ece4-Abstract.html>
- [34] Mirco Giacobbe, Daniel Kroening, Abhinandan Pal, and Michael Tautschnig. 2024. Neural Model Checking. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang (Eds.). http://papers.nips.cc/paper_files/paper/2024/hash/9d0947107ea92d6ce369dce7749180dd-Abstract-Conference.html
- [35] Tuomas Haarnoja, Ben Moran, Guy Lever, Sandy H. Huang, Dhruva Tirumala, Jan Humplik, Markus Wulfmeier, Saran Tunyasuvunakool, Noah Y. Siegel, Roland Hafner, Michael Bloesch, Kristian Hartikainen, Arunkumar Byravan, Leonard Hasenclever, Yuval Tassa, Fereshth Sadeghi, Nathan Batchelor, Federico Casarini, Stefano Saliceti, Charles Gao, Neil Sreendra, Kushal Patel, Marlon Gwira, Andrea Huber, Nicole Hurley, Francesco Nori, Raia Hadsell, and Nicolas Heess. 2024. Learning agile soccer skills for a bipedal robot with deep reinforcement learning. *Science Robotics* 9, 89 (2024), eadi8022. <https://doi.org/10.1126/scirobotics.adi8022> eprint: <https://www.science.org/doi/pdf/10.1126/scirobotics.adi8022>
- [36] Danijar Hafner, Timothy P. Lillicrap, Mohammad Norouzi, and Jimmy Ba. 2021. Mastering Atari with Discrete World Models. In *International Conference on Learning Representations*. <https://openreview.net/forum?id=0oabwyZbOu>
- [37] Danijar Hafner, Wilson Yan, and Timothy Lillicrap. 2025. Training Agents Inside of Scalable World Models. <https://arxiv.org/abs/2509.24527> eprint: 2509.24527.
- [38] Ernst Moritz Hahn, Mateo Perez, Sven Schewe, Fabio Somenzi, Ashutosh Trivedi, and Dominik Wojtczak. 2024. Multi-Agent Reinforcement Learning for Alternating-Time Logic. In *ECAI 2024 - 27th European Conference on Artificial Intelligence, 19-24 October 2024, Santiago de Compostela, Spain - Including 13th Conference on Prestigious Applications of Intelligent Systems (PAIS 2024) (Frontiers in Artificial Intelligence and Applications, Vol. 392)*, Ulle Endriss, Francisco S. Melo, Kerstin Bach, Alberto José Bugarin Diz, Jose Maria Alonso-Moral, Senén Barro, and Fredrik Heintz (Eds.). IOS Press, 1680–1687. <https://doi.org/10.3233/FAIA240676>
- [39] Lewis Hammond, Alessandro Abate, Julian Gutierrez, and Michael J. Wooldridge. 2021. Multi-Agent Reinforcement Learning with Temporal Logic Specifications. In *AAMAS '21: 20th International Conference on Autonomous Agents and Multiagent Systems, Virtual Event, United Kingdom, May 3-7, 2021*, Frank Dignum, Alessio Lomuscio, Ulle Endriss, and Ann Nowé (Eds.). ACM, 583–592. <https://doi.org/10.5555/3463952.3464024>
- [40] Christian Hensel, Sebastian Junges, Joost-Pieter Katoen, Tim Quatmann, and Matthias Volk. 2022. The probabilistic model checker Storm. *Int. J. Softw. Technol. Transf.* 24, 4 (2022), 589–610. <https://doi.org/10.1007/S10009-021-00633-Z>
- [41] Thomas A. Henzinger, Kaushik Mallik, Pouya Sadeghi, and Dordé Zikelic. 2025. Supermartingale Certificates for Quantitative Omega-Regular Verification and Control. In *Computer Aided Verification - 37th International Conference, CAV 2025, Zagreb, Croatia, July 23-25, 2025, Proceedings, Part II (Lecture Notes in Computer Science, Vol. 15932)*, Ruzica Piskac and Zvonimir Rakamaric (Eds.). Springer, 29–55. https://doi.org/10.1007/978-3-031-98679-6_2
- [42] Nils Jansen, Bettina Könighofer, Sebastian Junges, Alex Serban, and Roderick Bloem. 2020. Safe Reinforcement Learning Using Probabilistic Shields (Invited Paper). In *31st International Conference on Concurrency Theory (CONCUR 2020) (Leibniz International Proceedings in Informatics (LIPIcs), Vol. 171)*, Igor Konnov and Laura Kovács (Eds.). Schloss Dagstuhl–Leibniz-Zentrum für Informatik, Dagstuhl, Germany, 3:1–3:16. <https://doi.org/10.4230/LIPIcs.CONCUR.2020.3> ISSN: 1868-8969.
- [43] Sebastian Junges, Nils Jansen, Christian Dehnert, Ufuk Topcu, and Joost-Pieter Katoen. 2016. Safety-Constrained Reinforcement Learning for MDPs. In *Tools and Algorithms for the Construction and Analysis of Systems - 22nd International Conference, TACAS 2016, Eindhoven, The Netherlands, April 2-8, 2016, Proceedings (LNCS, Vol. 9636)*, Marsha Chechik and Jean-François Raskin (Eds.). Springer, 130–146. https://doi.org/10.1007/978-3-662-49674-9_8
- [44] M. Kwiatkowska, G. Norman, and D. Parker. 2005. Quantitative analysis with the probabilistic model checker PRISM. *Electronic Notes in Theoretical Computer Science* 153, 2 (2005), 5–31.
- [45] Bettina Könighofer, Roderick Bloem, Nils Jansen, Sebastian Junges, and Stefan Pranger. 2025. Shields for Safe Reinforcement Learning. *Commun. ACM* 68, 11 (Oct. 2025), 80–90. <https://doi.org/10.1145/3715958> Place: New York, NY, USA Publisher: Association for Computing Machinery.
- [46] Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahma, Lester James V. Miranda, Alisa Liu, Nouha Dziri, Shane Lyu, Yuling Gu, Saumya Malik, Victoria Graf, Jena D. Hwang, Jiangjiang Yang, Ronan Le Bras, Oyvind Tafjord, Chris Wilhelm, Luca Soldaini, Noah A. Smith, Yizhong Wang, Pradeep Dasigi, and Hannaneh Hajishirzi. 2024. TULU 3: Pushing Frontiers in Open Language Model Post-Training. *CoRR* abs/2411.15124 (2024). <https://doi.org/10.48550/ARXIV.2411.15124> arXiv: 2411.15124.
- [47] Romain Laroche, Paul Trichelair, and Remi Tachet des Combes. 2019. Safe Policy Improvement with Baseline Bootstrapping. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA (Proceedings of Machine Learning Research, Vol. 97)*, Kamalika Chaudhuri and Ruslan Salakhutdinov (Eds.). PMLR, 3652–3661. <http://proceedings.mlr.press/v97/laroche19a.html>
- [48] Harrison Lee, Samrat Phatale, Hassan Mansoor, Thomas Mesnard, Johan Ferret, Kellie Lu, Colton Bishop, Ethan Hall, Victor Carbone, Abhinav Rastogi, and Sushant Prakash. 2024. RLAI vs. RLHF: Scaling Reinforcement Learning from Human Feedback with AI Feedback. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net. <https://openreview.net/forum?id=uydQ2W41KO>
- [49] Udayan Mandal, Guy Amir, Haoze Wu, Ieva Daukantas, Fletcher Lee Newell, Umberto J. Ravaoli, Baoluo Meng, Michael Durling, Milan Ganai, Tobey Shim, Guy Katz, and Clark W. Barrett. 2024. Formally Verifying Deep Reinforcement Learning Controllers with Lyapunov Barrier Certificates. In *Formal Methods in Computer-Aided Design, FMCAD 2024, Prague, Czech Republic, October 15-18, 2024*, Nina Narodytska and Philipp Rümmer (Eds.). IEEE, 95–106. https://doi.org/10.34727/2024/ISBN.978-3-85448-065-5_15
- [50] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A. Rusu, Joel Veness, Marc G. Bellemare, Alex Graves, Martin A. Riedmiller, Andreas Fidjeland, Georg Ostrovski, Stig Petersen, Charles Beattie, Amir Sadik, Ioannis Antonoglou, Helen King, Dharmash Kumar, Daan Wierstra, Shane Legg, and Demis Hassabis. 2015. Human-level control through deep reinforcement learning. *Nature* 518, 7540 (2015), 529–533. <https://doi.org/10.1038/nature14236>
- [51] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh (Eds.). http://papers.nips.cc/paper_files/paper/2022/hash/b1efde53be364a73914f58805a001731-Abstract-Conference.html
- [52] Seohong Park, Oleh Rybkin, and Sergey Levine. 2024. METRA: Scalable Unsupervised RL with Metric-Aware Abstraction. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net. <https://openreview.net/forum?id=c5pwL0Soay>
- [53] Wasu Top Piriyaikulij, Yichao Liang, Hao Tang, Adrian Weller, Marta Kryven, and Kevin Ellis. 2025. PoE-World: Compositional World Modeling with Products of Programmatic Experts. *NeurIPS 2025* (2025).
- [54] Amir Pnueli. 1977. The Temporal Logic of Programs. In *18th Annual Symposium on Foundations of Computer Science, Providence, Rhode Island, USA, 31 October - 1 November 1977*. IEEE Computer Society, 46–57. <https://doi.org/10.1109/SFCS.1977.32>
- [55] Amir Pnueli and Roni Rosner. 1989. On the Synthesis of a Reactive Module. In *Conference Record of the Sixteenth Annual ACM Symposium on Principles of Programming Languages, Austin, Texas, USA, January 11-13, 1989*. ACM Press, 179–190. <https://doi.org/10.1145/75277.75293>

- [56] Martin L. Puterman. 1994. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. Wiley. <https://doi.org/10.1002/9780470316887>
- [57] Mathieu Reymond, Conor F. Hayes, Lander Willem, Roxana Rădulescu, Steven Abrams, Diederik M. Roijers, Enda Howley, Patrick Mannion, Niel Hens, Ann Nowé, and Pieter Libin. 2024. Exploring the Pareto front of multi-objective COVID-19 mitigation policies using reinforcement learning. *Expert Systems with Applications* 249 (2024), 123686. <https://doi.org/10.1016/j.eswa.2024.123686>
- [58] Daqian Shao and Marta Kwiatkowska. 2023. Sample Efficient Model-free Reinforcement Learning from LTL Specifications with Optimality Guarantees. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI 2023, 19th-25th August 2023, Macao, SAR, China*. ijcai.org, 4180–4189. <https://doi.org/10.24963/IJCAI.2023/465>
- [59] Harshit Sikchi, Andrea Tirinzoni, Ahmed Touati, Yingchen Xu, Anssi Kanervisto, Scott Niekum, Amy Zhang, Alessandro Lazaric, and Matteo Pirodda. 2025. Fast Adaptation with Behavioral Foundation Models. *RLC 2025* (2025). <https://doi.org/10.48550/ARXIV.2504.07896> arXiv: 2504.07896.
- [60] Thiago D. Simão, Romain Laroche, and Rémi Tachet des Combes. 2020. Safe Policy Improvement with an Estimated Baseline Policy. In *Proceedings of the 19th International Conference on Autonomous Agents and Multiagent Systems, AAMAS '20, Auckland, New Zealand, May 9-13, 2020, Amal El Fallah Seghrouchni, Gita Sukthankar, Bo An, and Neil Yorke-Smith (Eds.)*. International Foundation for Autonomous Agents and Multiagent Systems, 1269–1277. <https://doi.org/10.5555/3398761.3398908>
- [61] Richard S. Sutton and Andrew G. Barto. 2018. *Reinforcement Learning: An Introduction*. A Bradford Book, Cambridge, MA, USA.
- [62] Philip S. Thomas, Georgios Theodorou, and Mohammad Ghavamzadeh. 2015. High Confidence Policy Improvement. In *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015 (JMLR Workshop and Conference Proceedings, Vol. 37)*, Francis R. Bach and David M. Blei (Eds.). JMLR.org, 2380–2388. <http://proceedings.mlr.press/v37/thomas15.html>
- [63] Andrea Tirinzoni, Ahmed Touati, Jesse Farebrother, Mateusz Guzek, Anssi Kanervisto, Yingchen Xu, Alessandro Lazaric, and Matteo Pirodda. 2025. Zero-Shot Whole-Body Humanoid Control via Behavioral Foundation Models. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net. <https://openreview.net/forum?id=9sOR0nYLTz>
- [64] Rodrigo Toro Icarte, Toryn Q. Klassen, Richard Valenzano, and Sheila A. McIlraith. 2018. Teaching Multiple Tasks to an RL Agent using LTL. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems (AAMAS '18)*. International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 452–461. event-place: Stockholm, Sweden.
- [65] Ahmed Touati and Yann Ollivier. 2021. Learning One Representation to Optimize All Rewards. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan (Eds.). 13–23. <https://proceedings.neurips.cc/paper/2021/hash/003dd617c12d444ff9c80f717c3fa982-Abstract.html>
- [66] Elise van der Pol, Thomas Kipf, Frans A. Oliehoek, and Max Welling. 2020. Plannable Approximations to MDP Homomorphisms: Equivariance under Actions. In *Proceedings of the 19th International Conference on Autonomous Agents and MultiAgent Systems (AAMAS '20)*. International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 1431–1439.
- [67] Patrick Wienhöft, Marnix Suilen, Thiago D. Simão, Clemens Dubsloff, Christel Baier, and Nils Jansen. 2023. More for Less: Safe Policy Improvement with Stronger Performance Guarantees. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI 2023, 19th-25th August 2023, Macao, SAR, China*. ijcai.org, 4406–4415. <https://doi.org/10.24963/IJCAI.2023/490>
- [68] Xuesu Xiao, Bo Liu, Garrett Warnell, and Peter Stone. 2022. Motion planning and control for mobile robot navigation using machine learning: a survey. *Auton. Robots* 46, 5 (2022), 569–597. <https://doi.org/10.1007/S10514-022-10039-8>
- [69] Cambridge Yang, Michael L. Littman, and Michael Carbin. 2022. On the (In)Tractability of Reinforcement Learning for LTL Objectives. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI 2022, Vienna, Austria, 23-29 July 2022, Luc De Raedt (Ed.)*. ijcai.org, 3650–3658. <https://doi.org/10.24963/IJCAI.2022/507>
- [70] Wen-Chi Yang, Giuseppe Marra, Gavin Rens, and Luc De Raedt. 2023. Safe Reinforcement Learning via Probabilistic Logic Shields. In *IJCAI*. ijcai.org, 5739–5749. <https://doi.org/10.24963/ijcai.2023/637>
- [71] Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R. Narasimhan, and Yuan Cao. 2023. ReAct: Synergizing Reasoning and Acting in Language Models. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net. https://openreview.net/forum?id=WE_vluYUL-X
- [72] Lance Ying, Katherine M. Collins, Prafull Sharma, Cédric Colas, Kaiya Ivy Zhao, Adrian Weller, Zenna Tavares, Phillip Isola, Samuel J. Gershman, Jacob Andreas, Thomas L. Griffiths, François Chollet, Kelsey R. Allen, and Joshua B. Tenenbaum. 2025. Assessing Adaptive World Models in Machines with Novel Games. *CoRR* abs/2507.12821 (2025). <https://doi.org/10.48550/ARXIV.2507.12821> arXiv: 2507.12821.
- [73] Emily Yu, Dorde Zikelic, and Thomas A. Henzinger. 2025. Neural Control and Certificate Repair via Runtime Monitoring. In *AAAI-25, Sponsored by the Association for the Advancement of Artificial Intelligence, February 25 - March 4, 2025, Philadelphia, PA, USA, Toby Walsh, Julie Shah, and Zico Kolter (Eds.)*. AAAI Press, 26409–26417. <https://doi.org/10.1609/AAAI.V39I25.34840>
- [74] Amy Zhang, Rowan Thomas McAllister, Roberto Calandra, Yarin Gal, and Sergey Levine. 2021. Learning Invariant Representations for Reinforcement Learning without Reconstruction. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net. <https://openreview.net/forum?id=-2FCwDKRREu>
- [75] Changxi Zhu, Mehdi Dastani, and Shihan Wang. 2024. A survey of multi-agent deep reinforcement learning with communication. *Auton. Agents Multi Agent Syst.* 38, 1 (2024), 4. <https://doi.org/10.1007/S10458-023-09633-6>
- [76] Dorde Zikelic, Mathias Lechner, Thomas A. Henzinger, and Krishnendu Chatterjee. 2023. Learning Control Policies for Stochastic Systems with Reach-Avoid Guarantees. In *Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI 2023, Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence, IAAI 2023, Thirteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2023, Washington, DC, USA, February 7-14, 2023*, Brian Williams, Yiling Chen, and Jennifer Neville (Eds.). AAAI Press, 11926–11935. <https://doi.org/10.1609/AAAI.V37I10.26407>
- [77] Djordje Žikelić, Mathias Lechner, Abhinav Verma, Krishnendu Chatterjee, and Thomas A Henzinger. 2023. Compositional Policy Learning in Stochastic Control Systems with Formal Guarantees. In *NeurIPS*.