Contextual Intelligence The Next Leap for Reinforcement Learning

Blue Sky Ideas Track

André Biedenkapp
Albert-Ludwigs-Universität
Freiburg, Germany
biedenka@cs.uni-freiburg.de

ABSTRACT

Reinforcement learning (RL) has produced spectacular results in games, robotics, and continuous control. Yet, despite these successes, learned policies often fail to generalize beyond their training distribution, limiting real-world impact. Recent work on contextual RL (cRL) shows that exposing agents to environment characteristics – *contexts* – can improve zero-shot transfer. So far, the community has treated context as a monolithic, static observable, an approach that constrains the generalization capabilities of RL agents.

To achieve contextual intelligence we first propose a novel taxonomy of contexts that separates *allogenic* (environment-imposed) from *autogenic* (agent-driven) factors. We identify three fundamental research directions that must be addressed to promote truly contextual intelligence: (1) **Learning with heterogeneous contexts** to explicitly exploit the taxonomy levels so agents can reason about their influence on the world and vice versa; (2) **Multi-timescale modeling** to recognize that allogenic variables evolve slowly or remain static, whereas autogenic variables may change within an episode, potentially requiring different learning mechanisms; (3) **Integration of abstract, high-level contexts** to incorporate roles, resource & regulatory regimes, uncertainties, and other non-physical descriptors that crucially influence behavior.

We envision context as a first-class modeling primitive, empowering agents to reason about *who* they are, *what* the world permits, and *how* both evolve over time. By doing so, we aim to catalyze a new generation of context-aware agents that can be deployed safely and efficiently in the real world.

KEYWORDS

Context; Reinforcement Learning; Generalization

ACM Reference Format:
André Biedenkapp. 2026. Contextual Intelligence The Next Leap for Reinforcement Learning: 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, 6 pages. https://doi.org/10.65109/QNKH4630

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/QNKH4630

1 INTRODUCTION

Reinforcement learning (RL) [49] is a powerful paradigm that enables training of intelligent agents capable of solving even highly complex tasks. The simplicity of this paradigm promises great flexibility and the potential to be applicable to a large variety of target domains. However, prominent success stories of RL have largely focused on application domains with highly accurate, high-fidelity simulators. For example, the ready availability of game engines has spawned an increased interest in RL research, fueled by a string of impressive results from playing Atari Games [36] over mastering StarCraft II [51] to generally capable racing policies [25, 32, 53]. Beyond game playing, RL has been used to learn policies for magnetic control of Tokamak plasmas [16] or navigation of stratospheric balloons with difficult to predict weather conditions [5]. Despite these impressive successes, we believe RL research is largely held back by relying on having access to, or being able to design "perfect" environments. Consequently, RL agents are typically not able to be transferred to settings that are even slightly different from their training environment [8, 26, 33, 43, 57] or potentially need additional environment interactions at deployment to be able to act optimally in novel environments [4, 35].

A popular research direction in RL and robotics focused on exposing the learning agents to a wider distribution of experiences to mitigate this limitation. Domain randomization (DR) [41, 50] and procedural content generation (PCG) [15, 52] train agents on a distribution of related environments. This forces agents to learn robust behaviors rather than overfitting to particularities of a single environment. While these agents can be expected to be more transferrable across (highly) similar environments [3, 34, 50], these generalization capabilities often result in suboptimal solutions on individual environments [43, 45]. This is to be expected however, as such agents are not explicitly aware about which environment they are acting on, and thus they need to learn behaviors that work well on average. In a similar vein, robustness can be achieved by modeling this objective as a min-max problem. In this setting the goal is to learn a policy that maximizes the reward under the worst possible adversarial setting [40, 58]. This approach can mitigate worst-case outcomes but further sacrifices performance in average case scenarios as the learned policies act highly conservatively.

Contextual reinforcement learning (cRL) [8] offers a more principled approach to generalization by making environment characteristics, the so called context [29, 37], explicit in the training of agents. Contexts could, for example be physical properties of the system and consist of masses of a robot [43] or payloads [17], surface conditions [31] and decision time [11]. Such cRL works assume

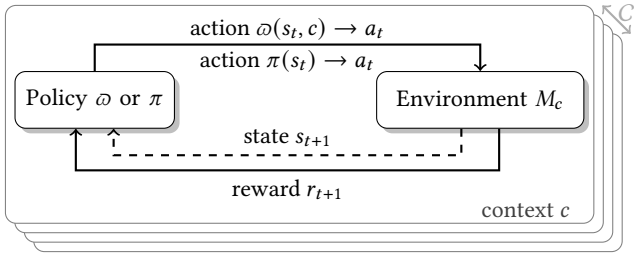

Figure 1: Schematic of a cMDP and two learning pipelines. A *context-oblivious* policy $\pi(s)$ and a *context-aware* policy $\varpi(s, c)$. Both are trained across contexts $c \sim C$.

that context is either readily available, e.g., from sensor readings [8, 25, 33, 43], or that it is unobservable and needs to be inferred [7, 39, 45, 55]. Learning in such a manner facilitates much improved zero-shot generalization capabilities [33] as agents can learn how to adapt to the environment such that they can act optimally on every environment and not just in the average case.

While cRL has proven to be effective in learning generalizable policies, especially with respect to zero-shot generalization, cRL still lacks a principled understanding and mechanisms to (1) *learning from observations and potentially heterogeneous contexts*; (2) *designing architectures and learning rules that directly leverage contextual structure*; (3) *integrate abstract, high-level contexts such as uncertainty of observations, roles of agents in a multi-agent system or resource budgets*. Solutions to these challenges will advance the field of multi-agent systems to unlocking truly contextual intelligence. The rest of the paper elaborates the contextual RL problem, provides a novel taxonomy of contexts based on which we sketch potential solution approaches to the three challenges we identified.

2 LEARNING GENERALIZABLE POLICIES

Contextual Markov Decision Processes (cMDP) [29, 37] (see Figure 1) extend the classic MDP formalism [6] to capture task generalization. An MDP $M = (\mathcal{S}, \mathcal{A}, \mathcal{T}, \mathcal{R}, \rho)$ comprises a state space $\mathcal{S}$, an action space $\mathcal{A}$, transition dynamics $\mathcal{T}$, a reward function $\mathcal{R}$ and a distribution over the initial states ρ. By introducing a context variable $c \in C$, we can *define*, *characterize*, and *parameterize* the environment's rules, thereby generating distinct task instances as contextual variations. In a cMDP, the action $\mathcal{A}$ and state spaces $\mathcal{S}$ stay the same whereas the transition dynamics $\mathcal{T}_c$, rewards $\mathcal{R}_c$ and initial state distributions ρ_c vary depending on the context $c \in C$. Consequently, the context-dependent initial distribution and altered dynamics can expose the agent to different regions of the state space across contexts. Following Benjamins et al. [8] we allow the context space C to be discrete or described by a distribution p_C. Thus, a cMDP $\mathcal{M}$ represents a family of related of MDPs $\mathcal{M} = \{M_c\}_{c \sim p_C}$ and can be seen as a sub-class of partially observable MDPs (POMDPs) [23].

When learning policies π on a cMDP, one can choose to either train context-oblivious policies $\pi \colon \mathcal{S} \rightarrow \mathcal{A}$ or context-aware ones $\varpi \colon \mathcal{S} \times C \rightarrow \mathcal{A}$. With domain randomization (DR) or procedural context generation (PCG), the goal is to learn policies that are robust to perturbations. Thus, learning agents typically can not explicitly[1] access the context during learning, leading to context-oblivious policy π that are robust to perturbations caused by sampling a new context $c \sim p_C$. To facilitate learning with either DR or PCG, it is important to carefully choose the distribution p_C to avoid exposing learning agents to highly different and potentially opposing experiences. Since the state $\mathcal{S}$ and action spaces $\mathcal{A}$ are shared in a cMDP, the same action might cause diametrically opposing outcomes (either in the reward function, the transition dynamics or both) for the same observed state. For example, assume a binary context and a binary action. Let the reward be the xor of the action and context values $r = a \oplus c$ for a given state s. Since an agent does not know the context c in this example, it is impossible to figure out which action value is the optimal choice. While this toy example exaggerates the problem, it exemplifies why the choice of context distribution is highly critical for DR and PCG and often might need some form of curriculum learning approach to provide the most stable results [see, e.g. 3, 34]. With dedicated curricula it is possible to control which experiences an agent is exposed to.

Context-aware policies ϖ on the other hand can explicitly take context information into account when choosing an action in a particular state. Thus, such policies can avoid the issue presented in the previous paragraph since the context allows agents to differentiate between outcomes. However, a new complication now arises from the design decision on how to adequately incorporate context into learning policies. Commonly, context is treated as another observable element and thus is stacked to the (observation-)state vector $\varpi(s, c) \rightarrow a$ [see, e.g., 1, 11, 42, 48, 54, 57]. While this can help in learning more general policies it is no silver bullet [8, 39]. Different algorithms seem to be more suited to this naïve treatment of context [8] and this integration of context seems to not-trivially affect the learning dynamics, requiring potentially vastly different hyperparameter settings during training [18].

A different line of work explores the use of hypernetworks [27] for contextual RL, in which such neural architectures learn to produce the weights of (or parts of) the policy networks, thereby adapting the policy's behavior to the context at hand [7, 9, 20]. Counter to the simple concatenation, this approach enables a more dedicated context feature learning as context and observations are processed in two separate representation learning streams, before being merged in downstream layers. Importantly though, counter to approaches that simply learn context specific representations before merging them with observation specific features [12, 25], hypernetwork approaches directly modify the policies behavior and do not simply condition policies on richer representations.

Opposite to the approaches that treat context largely as another observable, Prasanna et al. [43] try to exploit contextual information by more directly injecting context into latent representations of the Dreamer architecture [28]. Thereby, instead of learning how different streams of knowledge interact, this injection can be seen as a form of modulation of the latent representations that have been learned from the observations. This approach enabled learning of policies that had better out-of-distribution generalization abilities, when compared with classic concatenation an domain randomization approaches. Crucially, the analysis highlighted that contextual knowledge allowed the world model to be more robust to counterfactual observations. Similarly, Gumbsch et al. [26] aimed to learn when shifts in context occur (e.g., a door is opened/closed) to enable

[1] Relevant context might still be implicitly observable to an agent.

more proactive planning. Ultimately, this approach allowed them to learn temporal abstractions via hierarchies induced by context.

Furthermore, context-aware policies do not necessarily need explicit access to context. Instead, with system identification style approaches [21, 56, 59], as most often found in robotics [31] and meta-RL [4], attempt to estimate or recognize environment dynamics from a history of observations [39]. An important consequence of learning from a history of observations is that it enables online adaptation of context. Thus context is not treated as an unchanged quantity as it is done in the prior approaches. Essentially, context is quantified for a short snapshot and not for a whole episode or even longer period. Ndir et al. [39] for example used this fact to learn context representations that are directly tailored to the current behavior of a policy, i.e., factors that are relevant to the states a policy will traverse through, rather than ones that globally aim to estimate a contexts influence on the transition dynamics. While this work was limited to small scale simulation settings, this approach has been recently demonstrated to enable policies to generalize even on real-world robotic hardware [31].

Having formalized the learning of generalizable policies via cMDPs, we have presented two broad policy families that tackle this problem. For both we discussed lines of work that aim to address challenges of learning general policies. Across these lines of work a common thread emerges: *the way context is represented, injected, and learned dramatically shapes the resulting learning dynamics.* While recent cRL methods have achieved impressive performance, they still treat context as a static, homogeneous signal. This limitation motivates a more nuanced view of context in which we recognize its heterogeneous nature and its evolution on multiple time-scales.

In the next section we therefore propose a new taxonomy of contexts that classifies contexts by their structural properties, influence on dynamics, and temporal granularity. By making these distinctions explicit, we can design curricula, architectures, and learning mechanisms that fully exploit the rich, multi(-time)-scale character of context, moving us closer to genuine "contextual intelligence".

3 A TAXONOMY OF CONTEXT

Existing approaches treat the context variable as a monolithic, static signal. Hallak et al. [29] first formalized cMDPs only considering static contexts, i.e., a context c that is sampled once and remains unchanged while interacting with the MDP M_c it instantiates. This is fundamentally misaligned with how intelligence operates in the real world where contexts can differ dramatically in *what* they influence (dynamics, rewards, observations), *how* they are presented to the agent (explicitly observable vs. latent), and *when* they change (once per episode, intermittently within an episode, or continuously). To capture this heterogeneity we propose a taxonomy to decompose contexts into *allogenic* (environment-imposed) and *autogenic* (agent-driven) factors. A summary is provided in the appendix [10].

Allogenic Context. Such contexts are *exogenous*: they are imposed by the environment and are independent of the agent's own actions. An agent can observe or infer them, but it cannot influence their evolution. As such, this form of context provides more global knowledge about the environments reward and transition dynamics. This form of context thus aligns with the notion of context introduced by Hallak et al. [29]. Some examples of allogenic contexts include (1) physical constants (gravity [8], length of limbs of a robot [43], payload weights or atmospheric pressures); (2) hardware variations (motor torque, sensor noise levels, center of mass [17] or actuator latency); (3) environment layouts (map topology , wall placement [15], terrain type or lighting conditions).

Autogenic Context. This type of contexts are *endogenous*: they arise from the agent's own behavior, internal state, or learning process and can therefore be influenced and even deliberately controlled by the policy. Consequently, this form of context is further removed from the notion of context as introduced in [29]. Some examples of these more agent-driven factors include (1) internal states (battery level, wear-and-tear of actuators, limb failure [19], fatigue, or a hidden skill repertoire); (2) self-generated task parameters (interaction frequencies [11], goals set by a higher-level planner [22], curriculum difficulty chosen by the agent, or the current sub-task).

Whereas allogenic contexts describe factors that the environment imposes on the agent and the thereby resulting behavior, autogenic contexts describe how agent behavior can shape the environment or the interaction with it. This observation highlights a fundamental open problem: **how can we enable agents to reason jointly about allogenic and autogenic factors that exhibit fundamentally different influences and dependencies?** Existing cRL methods assume monolithic context and thus are not setup to exploit the heterogeneous structure revealed by our taxonomy. Addressing this gap requires new algorithmic primitives that (1) identify which aspects of the current situation are allogenic versus autogenic, (2) condition policies appropriately on each type, and (3) dynamically blend the two streams of information during learning and execution together with the regular observations.

4 TEMPORAL HIERARCHY OF CONTEXT

While leveraging heterogeneous sources of contextual information already brings us closer to contextual intelligence, we believe that we can further exploit the structure of context by paying additional attention to the temporal nature of context.

Allogenic context represents global information about the transition dynamics. In most episodic settings these factors are approximately stationary: they remain essentially constant throughout an episode, exhibit only minor stochastic fluctuations, and only rarely undergo abrupt, large-scale shifts [see, e.g., 26]. For instance, consider a robot that must leave a paved footpath to let a pedestrian pass. The friction and compliance of the grass beside the path differ dramatically from those of the pavement; however, once the robot has switched surfaces, the relevant properties of the new surface stay roughly unchanged until another transition occurs.

Autogenic context however evolves as a direct consequence of the agents own actions and internal state. Such variables change more smoothly and more frequently within an episode, yet their evolution is still slower than that of raw observations. A concrete example is a robot's battery charge: as the charge depletes, the robot preferentially selects low-energy maneuvers, causing the battery level to drift gradually rather than jump abruptly.

Allogenic contexts are largely piecewise-stationary and may experience sudden jumps, whereas autogenic contexts vary continuously and at a finer time-scale. Recognising and modelling these distinct temporal signatures is crucial for building agents that can

both anticipate broad environmental changes and adapt fluidly to their own evolving internal state. This observation highlights our second fundamental open problem: **how can we learn with sources of information that evolve at various different frequencies?** The typical cRL assumption of single monolithic context again will likely prohibit many approaches to directly exploit the notion of temporal dynamics discussed here. Notable exceptions here are the work by Gumbsch et al. [26] and the work on clockwork VAEs [47] which have not yet been explored for cRL. To close this gap we need new algorithmic primitives that (1) exploit the temporal dynamics of allogenic and autogenic contexts (e.g., via multi-timescale representation learning [47] or change-point detection); (2) balance exploration and exploitation with respect to both context streams (probing the environment to detect abrupt allogenic shifts versus exploiting the current autogenic context [24]); (3) integrate the two streams with the raw observation stream, adhering to the temporal structures without drowning out dynamic information with static ones (e.g., via dedicated encoder branches whose representations are dynamically blended during execution).

5 BEYOND PHYSICAL QUANTITIES

In recent developments, the focus of cRL has predominantly been directed towards physical quantities. While this is particularly enticing with the outlook on embodied AI, we believe that cRL research should broaden its focus to include more abstract context types [e.g., 13, 38, 44]. Our taxonomy admits *abstract* contexts that might not be directly measurable but are crucial for many MAS applications.

Team Roles. In various multi-agent reinforcement learning (MARL) settings [2], agents may adopt different functional roles within an environment. This is particularly evident in cooperative settings, where a team of agents might have distinct roles that must be filled to achieve a common goal. In a soccer team, for example, one can distinguish between defenders, midfielders, and attackers. While all agents share the same objective, they must exhibit different behaviors to ensure success. These roles, however, need not remain static throughout a game. If, for example, a defender possesses the ball in front of an empty goal, it should recognize that, momentarily, offensive behavior is required beyond what its normal role would permit. Thus, via dedicated communication protocols, roles might be exchanged when agents fulfill the necessary requirements.

Furthermore, in MARL settings, some agents might be human and thus require different coordination strategies than artificial agents. When approaching a human agent during a package delivery scenario, an artificial agent should prioritize the human's safety. However, when approaching another artificial agent in the same scenario, these safety considerations might not apply, allowing for a different approach procedure.

Resource Awareness. Previous examples have already elaborated on how battery power could inform more appropriate decision making in robotic agents. However, resource awareness can also comprise allogenic components. An autonomous factory or robotics warehouse could exploit knowledge of resource availability to increase production during peak renewable energy generation and reduce it to essential operations when only non-renewable energy is available. This extends beyond energy considerations and allows a manufacturing system to adjust its production schedule based on material availability or even carbon credit budgets.

Similarly, a more simple gardening robot might reduce water usage during droughts or defer watering if the weather forecast reliably predicts rain. In agricultural settings, such systems could further consider soil moisture levels, seasonal water allocation permits, and competing demands from other agricultural zones, demonstrating how resource contexts can be both physical and regulatory in nature. Meanwhile, energy, compute, and network bandwidth are naturally modeled as continuous autogenic variables. By conditioning policies on such contexts, robots can automatically trade off task urgency against resource consumption wich is particularly essential for long-duration missions.

Regulatory / Ethical Context. Finally, our world is governed not only by the laws of physics, but also by those of our societies [30]. A self-driving car might leverage the fact that the German Autobahn has no general speed limit, but must immediately adhere to speed limits upon crossing into neighboring countries. Such agents must further adapt to right-of-way rules, permitted lane-changing behaviors, and even culturally-specific expectations about pedestrian interactions. What is considered polite yielding behavior in one country might be seen as dangerous hesitation in another [46].

Legal regimes (such as speed zones) and ethical constraints (such as privacy budgets) and other forms of human preferences are allogenic yet abstract. Adequately encoding this form of context will be far from trivial but necessary to enable safe coexistence in our shared physical world. The challenge lies not just in representing these constraints, but in enabling agents to reason about their interactions. Thus, while RL from human feedback [14] has helped in shaping preference based-reward signals, we do not believe that it is enough to indirectly expose agents to human preferences but explicitly condition their behavior on these preferences.

These examples illustrate that contextual intelligence extends beyond low-level physics; it encompasses any factor that influences optimal decision-making. This observation highlights our final fundamental open problem: **how can we incorporate high-level and abstract contexts into the learning process?** To tackle this challenge an important aspect is to expose non-physical contexts to learning agents in settings where they might already be available or otherwise derive simulators that enable us to make progress towards this open problem.

6 CONCLUSION

The future of autonomous agents hinges on contextual intelligence. While the field has made great progress, often in isolated domains, we argue that crucial characteristics of what constitute context have not yet been taken into account. A unified view will result in agents that understand the difference between what they can change and what they must adapt to, that operate across temporal scales, and that reason about abstract contexts like roles and regulations will enable entirely new classes of applications. We call on the community to abandon the notion that context is merely another input feature. Context is the foundation on which we can build dedicated architectures, learning mechanisms, and theoretical frameworks. Otherwise, RL will remain confined to narrow domains with perfect simulators.

ACKNOWLEDGMENTS

André Biedenkapp acknowledges funding through the research network "Responsive and Scalable Learning for Robots Assisting Humans" (ReScaLe) of the University of Freiburg. The ReScaLe project is funded by the Carl Zeiss Foundation.

REFERENCES

[1] M. Abdolshah, H. Le, T. K. George, S. Gupta, S. Rana, and S. Venkatesh. 2021. A New Representation of Successor Features for Transfer across Dissimilar Environments. In *Proceedings of the 38th International Conference on Machine Learning (ICML'21) (Proceedings of Machine Learning Research, Vol. 139)*, M. Meila and T. Zhang (Eds.). PMLR, 1–9.

[2] S. V. Albrecht, F. Christianos, and L. Schäfer. 2024. *Multi-agent reinforcement learning: Foundations and modern approaches.* MIT Press.

[3] M. Andrychowicz, B. Baker, M. Chociej, R. Józefowicz, B. McGrew, J. Pachocki, A. Petron, M. Plappert, G. Powell, A. Ray, J. Schneider, S. Sidor, J. Tobin, P. Welinder, L. Weng, and W. Zaremba. 2020. Learning dexterous in-hand manipulation. *International Journal of Robotics Research* 39, 1 (2020).

[4] J. Beck, R. Vuorio, E. Z. Liu, Z. Xiong, L. M. Zintgraf, C. Finn, and S. Whiteson. 2025. A Tutorial on Meta-Reinforcement Learning. *Found. Trends Mach. Learn.* 18, 2-3 (2025), 224–384. https://doi.org/10.1561/2200000080

[5] M. G. Bellemare, S. Candido, P. Samuel Castro, J. Gong, M. C. Machado, S. Moitra, S. S. Ponda, and Z. Wang. 2020. Autonomous navigation of stratospheric balloons using reinforcement learning. *Nature* 588, 7836 (2020), 77–82.

[6] R. Bellman. 1957. A Markovian decision process. *Journal of Mathematics and Mechanics* (1957), 679–684.

[7] J. Benad, F. Röder, M. V. Butz, and M. Eppe. 2025. Shared dynamic model aligned hypernetworks for contextual reinforcement learning. In *Eighteenth European Workshop on Reinforcement Learning.* https://openreview.net/forum?id=6gdvQqkFKT

[8] C. Benjamins, T. Eimer, F. Schubert, A. Mohan, S. Döhler, A. Biedenkapp, B. Rosenhan, F. Hutter, and M. Lindauer. 2023. Contextualize Me – The Case for Context in Reinforcement Learning. *Transactions on Machine Learning Research* (2023).

[9] M. Beukman, D. Jarvis, R. Klein, S. James, and B. Rosman. 2023. Dynamics Generalisation in Reinforcement Learning via Adaptive Context-Aware Policies. In *Proceedings of the 36th International Conference on Advances in Neural Information Processing Systems (NeurIPS'23)*, A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (Eds.). Curran Associates.

[10] A Biedenkapp. 2026. https://github.com/AndreBiedenkapp/AndreBiedenkapp.github.io/blob/8b6dfd1df95e83a2eb5551ee54f48fdcce2da118/assets/pdf/paper/26-AAMAS-Blue-Sky.pdf

[11] A. Biedenkapp, R. Rajan, F. Hutter, and M. Lindauer. 2021. TempoRL: Learning When to Act. In *Proceedings of the 38th International Conference on Machine Learning (ICML'21) (Proceedings of Machine Learning Research, Vol. 139)*, M. Meila and T. Zhang (Eds.). PMLR, 914–924.

[12] A. Biedenkapp, D. Speck, S. Sievers, F. Hutter, M. Lindauer, and J. Seipp. 2022. Learning Domain-Independent Policies for Open List Selection. In *Workshop on Bridging the Gap Between AI Planning and Reinforcement Learning (PRL@ICAPS'22)*, M. Katz, H. Palacios, and V. Gómez (Eds.).

[13] P. Bordne, M. A. Hasan, E. Bergman, N. Awad, and A. Biedenkapp. 2024. CANDID DAC: Leveraging Coupled Action Dimensions with Importance Differences in DAC. In *Proc. of AutoML'24, Workshop Track.* https://openreview.net/forum?id=ZCCZYfstkG

[14] P. F. Christiano, J. Leike, T. B. Brown, M. Martic, S. Legg, and D. Amodei. 2017. Deep Reinforcement Learning from Human Preferences. In *Proceedings of the 30th International Conference on Advances in Neural Information Processing Systems (NeurIPS'17)*, I. Guyon, U. von Luxburg, S. Bengio, H. M. Wallach, R. Fergus, S. V. N. Vishwanathan, and R. Garnett (Eds.). 4299–4307.

[15] K. Cobbe, C. Hesse, J. Hilton, and J. Schulman. 2020. Leveraging Procedural Generation to Benchmark Reinforcement Learning. In *Proceedings of the 37th International Conference on Machine Learning (ICML'20)*, H. Daume III and A. Singh (Eds.), Vol. 98. Proceedings of Machine Learning Research.

[16] J. Degrave, F. Felici, J. Buchli, M. Neunert, B. Tracey, F. Carpanese, T. Ewalds, R. Hafner, A. Abdolmaleki, D. de las Casas, C. Donner, L. Fritz, C. Galperti, A. Huber, J. Keeling, M. Tsimpoukelli, J. Kay, A. Merle, Jean-M. Moret, S. Noury, F. Pesamosca, D. Pfau, O. Sauter, C. Sommariva, S. Coda, B. Duval, A. Fasoli, P. Kohli, K. Kavukcuoglu, D. Hassabis, and M. Riedmiller. 2022. Magnetic control of tokamak plasmas through deep reinforcement learning. *Nature* 602, 7897 (2022), 414–419.

[17] C. Ding, L. Zhou, Y. Li, and X. Rong. 2020. Locomotion Control of Quadruped Robots With Online Center of Mass Adaptation and Payload Identification. *IEEE Access* 8 (2020), 224578–224587. https://doi.org/10.1109/ACCESS.2020.3044933

[18] T. Eimer, C. Benjamins, and M. Lindauer. 2021. Hyperparameters in Contextual RL are Highly Situational. In *Workshop on Ecological Theory of Reinforcement Learning (EcoRL@NeurIPS'21).*

[19] T. Eimer, A. Biedenkapp, F. Hutter, and M. Lindauer. 2021. Self-Paced Context Evaluation for Contextual Reinforcement Learning. In *Proceedings of the 38th International Conference on Machine Learning (ICML'21) (Proceedings of Machine Learning Research, Vol. 139)*, M. Meila and T. Zhang (Eds.). PMLR, 2948–2958.

[20] L. Engwegen, D. Brinks, and W. Boehmer. 2025. Modular Recurrence in Contextual MDPs for Universal Morphology Control. In *Eighteenth European Workshop on Reinforcement Learning.* https://openreview.net/forum?id=0fn0ii1njp

[21] B. Evans, A. Thankaraj, and L. Pinto. 2022. Context is Everything: Implicit Identification for Dynamics Adaptation. In *2022 International Conference on Robotics and Automation, ICRA.* IEEE, 2642–2648.

[22] B. Eysenbach, R. R. Salakhutdinov, and S. Levine. 2019. Search on the Replay Buffer: Bridging Planning and Reinforcement Learning. In *Proceedings of the 32nd International Conference on Advances in Neural Information Processing Systems (NeurIPS'19)*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alche Buc, E. Fox, and R. Garnett (Eds.), Vol. 32. Curran Associates, Inc.

[23] D. Ghosh, J. Rahme, A. Kumar, A. Zhang, R. P. Adams, and S. Levine. 2021. Why Generalization in RL is Difficult: Epistemic POMDPs and Implicit Partial Observability. In *Proceedings of the 34th International Conference on Advances in Neural Information Processing Systems (NeurIPS'21)*, M. Ranzato, A. Beygelzimer, K. Nguyen, P. Liang, J. Vaughan, and Y. Dauphin (Eds.). Curran Associates, 25502–25515.

[24] Sebastian Griesbach and Carlo D'Eramo. 2025. Learning to Explore in Diverse Reward Settings via Temporal-Difference-Error Maximization. *Reinforcement Learning Journal* 6 (2025), 1140–1157.

[25] B. Grooten, P. MacAlpine, K. Subramanian, P. R. Wurman, and P. Stone. 2026. Out-of-Distribution Generalization with a SPARC: Racing 100 Unseen Vehicles with a Single Policy. In *Proceedings of the Fourtieth AAAI Conference on Artificial Intelligence.* AAAI Press.

[26] C. Gumbsch, N. Sajid, G. Martius, and M. V. Butz. 2024. Learning Hierarchical World Models with Adaptive Temporal Abstractions from Discrete Latent Dynamics. In *The Twelfth International Conference on Learning Representations (ICLR'24).* ICLR. https://openreview.net/forum?id=TjCDNssXKU

[27] D. Ha, A. M. Dai, and Q. V. Le. 2017. HyperNetworks. In *The Fifth International Conference on Learning Representations (ICLR'17).* ICLR, OpenReview.net.

[28] D. Hafner, J. Pasukonis, J. Ba, and T. P. Lillicrap. 2025. Mastering diverse control tasks through world models. *Nat.* 640, 8059 (2025), 647–653.

[29] A. Hallak, D. Di Castro, and S. Mannor. 2015. Contextual Markov Decision Processes. *arXiv:1502.02259 [stat.ML]* (2015).

[30] High-Level Expert Group on AI. 2019. *Ethics guidelines for trustworthy AI.* Report. European Commission. https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai

[31] M. Iannotta, Y. Yang, J. A. Stork, E. Schaffernicht, and T. Stoyanov. 2025. Can Context Bridge the Reality Gap? Sim-to-Real Transfer of Context-Aware Policies. *arXiv:2511.04249 [cs.RO]* (2025). https://doi.org/10.48550/arXiv.2511.04249

[32] E. Kaufmann, L. Bauersfeld, A. Loquercio, M. Müller, V. Koltun, and D. Scaramuzza. 2023. Champion-level drone racing using deep reinforcement learning. *Nat.* 620, 7976 (2023), 982–987. https://doi.org/10.1038/S41586-023-06419-4

[33] R. Kirk, A. Zhang, E. Grefenstette, and T. Rocktäschel. 2023. A Survey of Zero-shot Generalisation in Deep Reinforcement Learning. *Journal of Artificial Intelligence Research (JAIR)* 76 (2023), 201–264.

[34] P. Klink, C. D'Eramo, J. Peters, and J. Pajarinen. 2020. Self-Paced Deep Reinforcement Learning. In *Proceedings of the 33rd International Conference on Advances in Neural Information Processing Systems (NeurIPS'20)*, H. Larochelle, M. Ranzato, R. Hadsell, M.-F. Balcan, and H. Lin (Eds.). Curran Associates, 9216–9227.

[35] S. Levine, A. Kumar, G. Tucker, and J. Fu. 2020. Offline Reinforcement Learning: Tutorial, Review, and Perspectives on Open Problems. *arXiv:2005.01643 [cs.LG]* abs/2005.01643 (2020).

[36] V. Mnih, K. Kavukcuoglu, D. Silver, A. Rusu, J. Veness, M. Bellemare, A. Graves, M. Riedmiller, A. Fidjeland, G. Ostrovski, S. Petersen, C. Beattie, A. Sadik, I. Antonoglou, H. King, D. Kumaran, D. Wierstra, S. Legg, and D. Hassabis. 2015. Human-level control through deep reinforcement learning. *Nature* 518, 7540 (26 02 2015), 529–533.

[37] A. Modi, N. Jiang, S. Singh, and A. Tewari. 2018. Markov Decision Processes With Continuous Side Information. In *Algorithmic Learning Theory (ALT'18)*, Vol. 83. 597–618.

[38] A. Mohan, A. Zhang, and M. Lindauer. 2024. Structure in Deep Reinforcement Learning: A Survey and Open Problems. *Journal of Artificial Intelligence Research* 79 (2024).

[39] T. Camaret Ndir, A. Biedenkapp, and N. Awad. 2024. Inferring Behavior-Specific Context Improves Zero-Shot Generalization in Reinforcement Learning. In *Seventeenth European Workshop on Reinforcement Learning.* https://openreview.net/forum?id=51XSWH0mgN

[40] K. Panaganti, Z. Xu, D. Kalathil, and M. Ghavamzadeh. 2022. Robust Reinforcement Learning using Offline Data. In *Proceedings of the 35th International Conference on Advances in Neural Information Processing Systems (NeurIPS'22)*, S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (Eds.). Curran Associates.

[41] X. Bin Peng, M. Andrychowicz, W. Zaremba, and P. Abbeel. 2018. Sim-to-Real Transfer of Robotic Control with Dynamics Randomization. In *Proc. of ICRA'18*. IEEE, 1–8.

[42] C. Perez, F. P. Such, and T. Karaletsos. 2020. Generalized hidden parameter mdps: Transferable model-based rl in a handful of trials. In *Proceedings of the AAAI Conference on Artificial Intelligence*, F. Rossi, V. Conitzer, and F. Sha (Eds.). Association for the Advancement of Artificial Intelligence, AAAI Press, 5403–5411.

[43] S. Prasanna, K. Farid, R. Rajan, and A. Biedenkapp. 2024. Dreaming of Many Worlds: Learning Contextual World Models Aids Zero-Shot Generalization. *Reinforcement Learning Journal* 1 (2024).

[44] R. Rajan, J. Diaz, S. Guttikonda, F. Ferreira, A. Biedenkapp, J. Ole von H., and F. Hutter. 2023. MDP Playground: An Analysis and Debug Testbed for Reinforcement Learning. *Journal of Artificial Intelligence Research (JAIR)* 77 (2023), 821–890.

[45] S. Reed, K. Zolna, E. Parisotto, S. Gómez C., A. Novikov, G. Barth-maron, M. Giménez, Y. Sulsky, J. Kay, J. T. Springenberg, T. Eccles, J. Bruce, A. Razavi, A. Edwards, N. Heess, Y. Chen, R. Hadsell, O. Vinyals, M. Bordbar, and N. de Freitas. 2022. A Generalist Agent. *Transactions on Machine Learning Research* (2022). https://openreview.net/forum?id=1ikK0kHjvj Featured Certification, Outstanding Certification.

[46] S. Russell. 2022. Human-Compatible Artificial Intelligence. In *Human-Like Machine Intelligence*, S. H. Muggleton and N. Chater (Eds.). Oxford University Press, 3–23.

[47] V. Saxena, J. Ba, and D. Hafner. 2021. Clockwork Variational Autoencoders. In *Proceedings of the 34th International Conference on Advances in Neural Information Processing Systems (NeurIPS'21)*, M. A. Ranzato, A. Beygelzimer, Y. N. Dauphin, P. Liang, and J. W. Vaughan (Eds.). Curran Associates, 29246–29257.

[48] S. Sodhani, A. Zhang, and J. Pineau. 2021. Multi-Task Reinforcement Learning with Context-based Representations. In *Proceedings of the 38th International Conference on Machine Learning (ICML'21) (Proceedings of Machine Learning Research, Vol. 139)*, M. Meila and T. Zhang (Eds.). PMLR, 9767–9779.

[49] R. S. Sutton and A. G. Barto. 2018. *Reinforcement learning: An introduction* (2 ed.). MIT Press.

[50] J. Tobin, R. Fong, A. Ray, J. Schneider, W. Zaremba, and P. Abbeel. 2017. Domain randomization for transferring deep neural networks from simulation to the real world. In *Proc. of (IROS'17)*. IEEE, 23–30.

[51] O. Vinyals, I. Babuschkin, W. Czarnecki, M. Mathieu, A. Dudzik, J. Chung, D. Choi, R. Powell, T. Ewalds, P. Georgiev, J. Oh, D. Horgan, M. Kroiss, I. Danihelka, A. Huang, L. Sifre, T. Cai, J. Agapiou, M. Jaderberg, A. Vezhnevets, R. Leblond, T. Pohlen, V. Dalibard, D. Budden, Y. Sulsky, J. Molloy, T. Paine, C. Gulcehre, Z. Wand, T. Pfaff, Y. Wu, R. Ring, D. Yogatama, D. Wünsch, K McKinney, O. Smith, T. Schaul, T. Lillicrap, K. Kavukcuoglu, D. Hassabis, C. Apps, and D. Silver. 2019. Grandmaster level in StarCraft II using multi-agent reinforcement learning. *Nature* 7782 (2019), 350–354.

[52] J. Wang, M. King, N. Porcel, Z. Kurth-Nelson, T. Zhu, C. Deck, P. Choy, M. Cassin, M. Reynolds, H. F. Song, G. Buttimore, D. P. Reichert, N. C. Rabinowitz, L. Matthey, D. Hassabis, A. Lerchner, and M. M. Botvinick. 2021. Alchemy: A benchmark and analysis toolkit for meta-reinforcement learning agents. In *Proceedings of the 34th International Conference on Advances in Neural Information Processing Systems (NeurIPS'21)*, M. Ranzato, A. Beygelzimer, K. Nguyen, P. Liang, J. Vaughan, and Y. Dauphin (Eds.). Curran Associates.

[53] P. R. Wurman, S. Barrett, K. Kawamoto, J. MacGlashan, K. Subramanian, T. J. Walsh, R. Capobianco, A. Devlic, F. Eckert, F. Fuchs, L. Gilpin, P. Khandelwal, V. Raj Kompella, H. Lin, P. MacAlpine, D. Oller, T. Seno, C. Sherstan, M. D. Thomure, H. Aghabozorgi, L. Barrett, R. Douglas, D. Whitehead, P. Dürr, P. Stone, M. Spranger, and H. Kitano. 2022. Outracing champion Gran Turismo drivers with deep reinforcement learning. *Nat.* 602, 7896 (2022), 223–228. https://doi.org/10.1038/s41586-021-04357-7

[54] D. Yarats, R. Fergus, A. Lazaric, and L. Pinto. 2021. Reinforcement Learning with Prototypical Representations. In *Proceedings of the 38th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 139)*, M. Meila and T. Zhang (Eds.). PMLR, 11920–11931.

[55] W. Yu, J. Tan, C. K. Liu, and G. Turk. 2017. Preparing for the unknown: Learning a universal policy with online system identification. *arXiv preprint arXiv:1702.02453* (2017).

[56] W. Yu, J. Tan, C. K. Liu, and G. Turk. 2017. Preparing for the Unknown: Learning a Universal Policy with Online System Identification. In *Robotics: Science and Systems XIII*.

[57] A. Zhang, S. Sodhani, K. Khetarpal, and J. Pineau. 2021. Learning Robust State Abstractions for Hidden-parameter Block MDPs. In *The Ninth International Conference on Learning Representations (ICLR'21)*. ICLR.

[58] H. Zhang, H. Chen, C. Xiao, B. Li, M. Liu, D. S. Boning, and C.-J. Hsieh. 2020. Robust Deep Reinforcement Learning against Adversarial Perturbations on State Observations. In *Proceedings of the 33rd International Conference on Advances in Neural Information Processing Systems (NeurIPS'20)*, H. Larochelle, M. Ranzato, R. Hadsell, M.-F. Balcan, and H.-T. Lin (Eds.). Curran Associates.

[59] W. Zhou, L. Pinto, and A. Gupta. 2019. Environment Probing Interaction Policies. In *The Seventh International Conference on Learning Representations (ICLR'19)*. ICLR, OpenReview.net.