Model-based Reinforcement Learning: A Survey

Other titles in Foundations and Trends® in Machine Learning

Divided Differences, Falling Factorials, and Discrete Splines: Another Look at Trend Filtering and Related Problems
Ryan J. Tibshirani
ISBN: 978-1-63828-036-1

Risk-Sensitive Reinforcement Learning via Policy Gradient Search
Prashanth L. A. and Michael C. Fu
ISBN: 978-1-63828-026-2

A Unifying Tutorial on Approximate Message Passing
Oliver Y. Feng, Ramji Venkataramanan, Cynthia Rush and Richard J. Samworth
ISBN: 978-1-63828-004-0

Learning in Repeated Auctions
Thomas Nedelec, Clément Calauzènes, Noureddine El Karoui and Vianney Perchet
ISBN: 978-1-68083-938-8

Dynamical Variational Autoencoders: A Comprehensive Review
Laurent Girin, Simon Leglaive, Xiaoyu Bie, Julien Diard, Thomas Hueber and Xavier Alameda-Pineda
ISBN: 978-1-68083-912-8

Machine Learning for Automated Theorem Proving: Learning to Solve SAT and QSATe
Sean B. Holden
ISBN: 978-1-68083-898-5

Model-based Reinforcement Learning: A Survey

Thomas M. Moerland
LIACS, Leiden University
t.m.moerland@liacs.leidenuniv.nl

Joost Broekens
LIACS, Leiden University

Aske Plaat
LIACS, Leiden University

Catholijn M. Jonker
Interactive Intelligence, TU Delft
LIACS, Leiden University

the essence of knowledge
Boston — Delft

Foundations and Trends® in Machine Learning

Published, sold and distributed by:
now Publishers Inc.
PO Box 1024
Hanover, MA 02339
United States
Tel. +1-781-985-4510
www.nowpublishers.com
sales@nowpublishers.com

Outside North America:
now Publishers Inc.
PO Box 179
2600 AD Delft
The Netherlands
Tel. +31-6-51115274

The preferred citation for this publication is

T. M. Moerland *et al.*. *Model-based Reinforcement Learning: A Survey*. Foundations and Trends® in Machine Learning, vol. 16, no. 1, pp. 1–118, 2023.

ISBN: 978-1-63828-057-6

Foundations and Trends® in Machine Learning
Volume 16, Issue 1, 2023

Editorial Scope

Topics

Foundations and Trends® in Machine Learning publishes survey and tutorial articles in the following topics:

- Adaptive control and signal processing
- Applications and case studies
- Behavioral, cognitive and neural learning
- Bayesian learning
- Classification and prediction
- Clustering
- Data mining
- Dimensionality reduction
- Evaluation
- Game theoretic learning
- Graphical models
- Independent component analysis
- Inductive logic programming
- Kernel methods
- Markov chain Monte Carlo
- Model choice
- Nonparametric methods
- Online learning
- Optimization
- Reinforcement learning
- Relational learning
- Robustness
- Spectral methods
- Statistical learning theory
- Variational inference
- Visualization

Information for Librarians

Foundations and Trends® in Machine Learning, 2023, Volume 16, 6 issues. ISSN paper version 1935-8237. ISSN online version 1935-8245. Also available as a combined paper and online subscription.

Contents

Model-based Reinforcement Learning: A Survey

Thomas M. Moerland[1], Joost Broekens[1], Aske Plaat[1] and Catholijn M. Jonker[1,2]

[1]*LIACS, Leiden University, The Netherlands; t.m.moerland@liacs.leidenuniv.nl*
[2]*Interactive Intelligence, TU Delft, The Netherlands*

ABSTRACT

Sequential decision making, commonly formalized as Markov Decision Process (MDP) optimization, is an important challenge in artificial intelligence. Two key approaches to this problem are reinforcement learning (RL) and planning. This survey is an integration of both fields, better known as model-based reinforcement learning. Model-based RL has two main steps. First, we systematically cover approaches to dynamics model learning, including challenges like dealing with stochasticity, uncertainty, partial observability, and temporal abstraction. Second, we present a systematic categorization of planning-learning integration, including aspects like: where to start planning, what budgets to allocate to planning and real data collection, how to plan, and how to integrate planning in the learning and acting loop. After these two sections, we also discuss implicit model-based RL as an end-to-end alternative for model learning and planning, and we cover the potential benefits of model-based RL. Along the way, the survey also draws connections to several related RL fields, like hierarchical RL and transfer

Thomas M. Moerland, Joost Broekens, Aske Plaat and Catholijn M. Jonker (2023), "Model-based Reinforcement Learning: A Survey", Foundations and Trends® in Machine Learning: Vol. 16, No. 1, pp 1–118. DOI: 10.1561/2200000086.

learning. Altogether, the survey presents a broad conceptual overview of the combination of planning and learning for MDP optimization.

1

Introduction

Sequential decision making, commonly formalized as Markov Decision Process (MDP) (Bellman, 1954; Puterman, 2014) optimization, is a key challenge in artificial intelligence. Two successful approaches to solve this problem are *planning* (Russell and Norvig, 2016; Bertsekas *et al.*, 1995) and *reinforcement learning* (Sutton and Barto, 2018). Planning and learning may actually be combined, in a field which is known as *model-based reinforcement learning*. We define model-based RL as: 'any MDP approach that i) uses a model (known or learned) and ii) uses learning to approximate a global value or policy function'.

While model-based RL has shown great success (Silver *et al.*, 2017b; Levine and Koltun, 2013; Deisenroth and Rasmussen, 2011), literature lacks a systematic review of the field (although Hamrick *et al.* (2020) does provide an overview of mental simulation in deep learning, see Section 9 for a detailed discussion of related work). Therefore, this survey presents a combination of planning and learning. A general scheme of the possible connections between planning and learning, which we will use throughout the survey, is shown in Figure 1.1.

The survey is organized as follows. After a short introduction of the MDP optimization problem (Section 2), we first define the categories of model-based reinforcement learning and their relation to the fields of planning and model-free reinforcement learning (Section 3). Afterwards, Sections 4-7 present the main body of this survey. The crucial first step of most model-based RL algorithms is *dynamics model learning* (Figure 1.1, arrow g) which we cover in Section 4. When we have obtained a model, the second step of model-based RL is to integrate planning and learning (Figure 1.1, arrows a-f), which we discuss in Section 5. Interestingly, some model-based RL approaches do not explicitly define one or both of these steps (model learning and integration of planning and learning), but rather wrap them into a larger (end-to-end) optimization. We call these methods *implicit model-based RL*, which we cover in Section 6. Finally, we conclude the main part of this survey with a discussion of the potential benefits of these approaches, and of model-based RL in general (Section 7).

While the main focus of this survey is on the practical/empirical aspects of model-based RL, we also shortly highlight the main theoretical results on the convergence properties of model-based RL algorithms (Section 8). Additionally, note that model-based RL is a fundamental approach to sequential decision making, and many other sub-disciplines in RL have a close connection to model-based RL. For example, *hierarchical reinforcement learning* (Barto and Mahadevan, 2003) can be approached in a model-based way, where the higher-level action space defines a model with temporal abstraction. Model-based RL is also an important approach to *transfer learning* (Taylor and Stone, 2009) (through model transfer between tasks) and *targeted exploration* (Thrun, 1992). When applicable, the survey also presents short overviews of such related RL research directions. Finally, the survey finishes with Related Work (Section 9), Discussion (Section 10), and Summary (Section 11) sections.

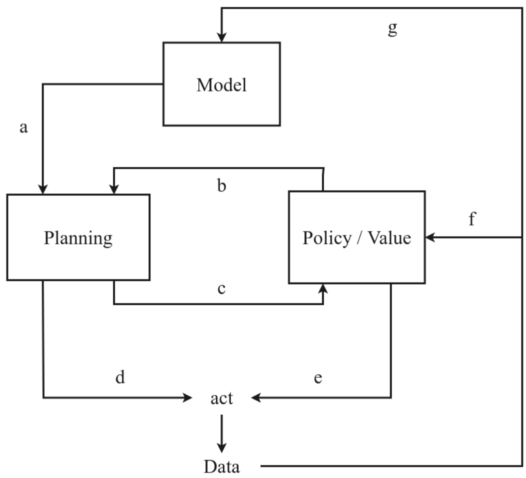

Figure 1.1: Overview of possible algorithmic connections between planning and learning. Learning can take place at two locations: in learning a dynamics model (arrow g), and/or in learning a policy/value function (arrows c and f). Most algorithms only implement a subset of the possible connections. Explanation of each arrow: a) plan over a learned model, b) use information from a policy/value network to improve the planning procedure, c) use the result from planning as training targets for a policy/value, d) act in the real world based on the planning outcome, e) act in the real world based on a policy/value function, f) generate training targets for the policy/value based on real world data, g) generate training targets for the model based on real world data.

References

Abbeel, P. and A. Y. Ng. (2005). “Learning first-order Markov models for control”. In: *Advances in neural information processing systems*. 1–8.

Achiam, J., H. Edwards, D. Amodei, and P. Abbeel. (2018). “Variational option discovery algorithms”. *arXiv preprint arXiv:1807.10299*.

Achiam, J. and S. Sastry. (2017). “Surprise-based intrinsic motivation for deep reinforcement learning”. *arXiv preprint arXiv:1703.01732*.

Agostinelli, F., S. McAleer, A. Shmakov, and P. Baldi. (2019). “Solving the Rubik’s cube with deep reinforcement learning and search”. *Nature Machine Intelligence*. 1(8): 356–363.

Agrawal, P., A. V. Nair, P. Abbeel, J. Malik, and S. Levine. (2016). “Learning to poke by poking: Experiential learning of intuitive physics”. In: *Advances in Neural Information Processing Systems*. 5074–5082.

Agrawal, S. and R. Jia. (2017). “Posterior sampling for reinforcement learning: worst-case regret bounds”. In: *Advances in Neural Information Processing Systems*. 1184–1194.

Amodei, D., C. Olah, J. Steinhardt, P. Christiano, J. Schulman, and D. Mané. (2016). “Concrete problems in AI safety”. *arXiv preprint arXiv:1606.06565*.

Anand, A., E. Racah, S. Ozair, Y. Bengio, M.-A. Côté, and R. D. Hjelm. (2019). "Unsupervised state representation learning in atari". In: *Advances in Neural Information Processing Systems*. 8766–8779.

Anthony, T., R. Nishihara, P. Moritz, T. Salimans, and J. Schulman. (2019). "Policy Gradient Search: Online Planning and Expert Iteration without Search Trees". *arXiv preprint arXiv:1904.03646*.

Anthony, T., Z. Tian, and D. Barber. (2017). "Thinking fast and slow with deep learning and tree search". In: *Advances in Neural Information Processing Systems*. 5360–5370.

Asadi, K., E. Cater, D. Misra, and M. L. Littman. (2018). "Towards a Simple Approach to Multi-step Model-based Reinforcement Learning". *arXiv preprint arXiv:1811.00128*.

Asmuth, J., L. Li, M. L. Littman, A. Nouri, and D. Wingate. (2009). "A Bayesian sampling approach to exploration in reinforcement learning". In: *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*. AUAI Press. 19–26.

Åström, K. J. and P. Eykhoff. (1971). "System identification—a survey". *Automatica*. 7(2): 123–162.

Aswani, A., H. Gonzalez, S. S. Sastry, and C. Tomlin. (2013). "Provably safe and robust learning-based model predictive control". *Automatica*. 49(5): 1216–1226.

Atkeson, C. G., A. W. Moore, and S. Schaal. (1997). "Locally weighted learning for control". In: *Lazy learning*. Springer. 75–113.

Atkeson, C. G. and J. C. Santamaria. (1997). "A comparison of direct and model-based reinforcement learning". In: *Proceedings of International Conference on Robotics and Automation*. Vol. 4. IEEE. 3557–3564.

Auer, P. (2002). "Using confidence bounds for exploitation-exploration trade-offs". *Journal of Machine Learning Research*. 3: 397–422.

Avila Belbute-Peres, F. de, K. Smith, K. Allen, J. Tenenbaum, and J. Z. Kolter. (2018). "End-to-end differentiable physics for learning and control". In: *Advances in Neural Information Processing Systems*. 7178–7189.

Azar, M. G., I. Osband, and R. Munos. (2017). "Minimax regret bounds for reinforcement learning". In: *International Conference on Machine Learning*. PMLR. 263–272.

Babaeizadeh, M., C. Finn, D. Erhan, R. H. Campbell, and S. Levine. (2017). "Stochastic variational video prediction". *arXiv preprint arXiv:1710.11252*.

Bacon, P.-L., J. Harb, and D. Precup. (2017). "The option-critic architecture". In: *Thirty-First AAAI Conference on Artificial Intelligence*.

Bagnell, J. A. and J. G. Schneider. (2001). "Autonomous helicopter control using reinforcement learning policy search methods". In: *Proceedings 2001 ICRA. IEEE International Conference on Robotics and Automation (Cat. No. 01CH37164)*. Vol. 2. IEEE. 1615–1620.

Baranes, A. and P.-Y. Oudeyer. (2009). "R-iac: Robust intrinsically motivated exploration and active learning". *IEEE Transactions on Autonomous Mental Development*. 1(3): 155–169.

Baranes, A. and P.-Y. Oudeyer. (2013). "Active learning of inverse models with intrinsically motivated goal exploration in robots". *Robotics and Autonomous Systems*. 61(1): 49–73.

Barreto, A., W. Dabney, R. Munos, J. J. Hunt, T. Schaul, H. P. van Hasselt, and D. Silver. (2017). "Successor features for transfer in reinforcement learning". In: *Advances in neural information processing systems*. 4055–4065.

Barto, A. G., S. J. Bradtke, and S. P. Singh. (1995). "Learning to act using real-time dynamic programming". *Artificial intelligence*. 72(1-2): 81–138.

Barto, A. G. and S. Mahadevan. (2003). "Recent advances in hierarchical reinforcement learning". *Discrete event dynamic systems*. 13(1-2): 41–77.

Battaglia, P., R. Pascanu, M. Lai, D. J. Rezende, *et al.* (2016). "Interaction networks for learning about objects, relations and physics". In: *Advances in neural information processing systems*. 4502–4510.

Battaglia, P. W., J. B. Hamrick, V. Bapst, A. Sanchez-Gonzalez, V. Zambaldi, M. Malinowski, A. Tacchetti, D. Raposo, A. Santoro, R. Faulkner, *et al.* (2018). "Relational inductive biases, deep learning, and graph networks". *arXiv preprint arXiv:1806.01261*.

Baxter, J., A. Tridgell, and L. Weaver. (1999). "TDLeaf (lambda): Combining temporal difference learning with game-tree search". *arXiv preprint cs/9901001*.

Beck, J., K. Ciosek, S. Devlin, S. Tschiatschek, C. Zhang, and K. Hofmann. (2020). "Amrl: Aggregated memory for reinforcement learning". In:

Bellemare, M., S. Srinivasan, G. Ostrovski, T. Schaul, D. Saxton, and R. Munos. (2016). "Unifying count-based exploration and intrinsic motivation". In: *Advances in Neural Information Processing Systems*. 1471–1479.

Bellemare, M. G., Y. Naddaf, J. Veness, and M. Bowling. (2013). "The arcade learning environment: An evaluation platform for general agents". *Journal of Artificial Intelligence Research*. 47: 253–279.

Bellman, R. (1954). "The theory of dynamic programming". *Bulletin of the American Mathematical Society*. 60(6): 503–515.

Bellman, R. (1966). "Dynamic programming". *Science*. 153(3731): 34–37.

Bengio, Y., J. Louradour, R. Collobert, and J. Weston. (2009). "Curriculum learning". In: *Proceedings of the 26th annual international conference on machine learning*. ACM. 41–48.

Berkenkamp, F., M. Turchetta, A. Schoellig, and A. Krause. (2017). "Safe model-based reinforcement learning with stability guarantees". In: *Advances in neural information processing systems*. 908–918.

Bertsekas, D. P., D. P. Bertsekas, D. P. Bertsekas, and D. P. Bertsekas. (1995). *Dynamic programming and optimal control*. Vol. 1. No. 2. Athena scientific Belmont, MA.

Bishop, C. M. (2006). *Pattern recognition and machine learning*. Springer.

Blundell, C., B. Uria, A. Pritzel, Y. Li, A. Ruderman, J. Z. Leibo, J. Rae, D. Wierstra, and D. Hassabis. (2016). "Model-free episodic control". *arXiv preprint arXiv:1606.04460*.

Boone, G. (1997). "Efficient reinforcement learning: Model-based acrobot control". In: *Proceedings of International Conference on Robotics and Automation*. Vol. 1. IEEE. 229–234.

Brafman, R. I. and M. Tennenholtz. (2002). "R-max-a general polynomial time algorithm for near-optimal reinforcement learning". *Journal of Machine Learning Research*. 3(Oct): 213–231.

Brockman, G., V. Cheung, L. Pettersson, J. Schneider, J. Schulman, J. Tang, and W. Zaremba. (2016). “OpenAI Gym”. *arXiv preprint arXiv:1606.01540*.

Browne, C. B., E. Powley, D. Whitehouse, S. M. Lucas, P. I. Cowling, P. Rohlfshagen, S. Tavener, D. Perez, S. Samothrakis, and S. Colton. (2012). “A survey of monte carlo tree search methods”. *IEEE Transactions on Computational Intelligence and AI in games*. 4(1): 1–43.

Brunskill, E. and L. Li. (2014). “Pac-inspired option discovery in lifelong reinforcement learning”. In: *International conference on machine learning*. 316–324.

Buckman, J., D. Hafner, G. Tucker, E. Brevdo, and H. Lee. (2018). “Sample-efficient reinforcement learning with stochastic ensemble value expansion”. In: *Advances in Neural Information Processing Systems*. 8224–8234.

Buesing, L., T. Weber, S. Racaniere, S. Eslami, D. Rezende, D. P. Reichert, F. Viola, F. Besse, K. Gregor, D. Hassabis, *et al.* (2018). “Learning and querying fast generative models for reinforcement learning”. *arXiv preprint arXiv:1802.03006*.

Caruana, R. (1997). “Multitask learning”. *Machine learning*. 28(1): 41–75.

Castro, P. S. and D. Precup. (2007). “Using Linear Programming for Bayesian Exploration in Markov Decision Processes.” In: *IJCAI*. Vol. 24372442.

Chang, M. B., T. Ullman, A. Torralba, and J. B. Tenenbaum. (2016). “A compositional object-based approach to learning physical dynamics”. *arXiv preprint arXiv:1612.00341*.

Chentanez, N., A. G. Barto, and S. P. Singh. (2005). “Intrinsically motivated reinforcement learning”. In: *Advances in neural information processing systems*. 1281–1288.

Chiappa, S., S. Racaniere, D. Wierstra, and S. Mohamed. (2017). “Recurrent environment simulators”. *arXiv preprint arXiv:1704.02254*.

Choi, J., Y. Guo, M. Moczulski, J. Oh, N. Wu, M. Norouzi, and H. Lee. (2018). “Contingency-aware exploration in reinforcement learning”. *arXiv preprint arXiv:1811.01483*.

Chrisman, L. (1992). "Reinforcement learning with perceptual aliasing: The perceptual distinctions approach". In: *AAAI*. Vol. 1992. Citeseer. 183–188.

Christiano, P., Z. Shah, I. Mordatch, J. Schneider, T. Blackwell, J. Tobin, P. Abbeel, and W. Zaremba. (2016). "Transfer from simulation to real world through learning deep inverse dynamics model". *arXiv preprint arXiv:1610.03518*.

Chua, K., R. Calandra, R. McAllister, and S. Levine. (2018). "Deep reinforcement learning in a handful of trials using probabilistic dynamics models". In: *Advances in Neural Information Processing Systems*. 4754–4765.

Clavera, I., J. Rothfuss, J. Schulman, Y. Fujita, T. Asfour, and P. Abbeel. (2018). "Model-Based Reinforcement Learning via Meta-Policy Optimization". In: *Conference on Robot Learning*. 617–629.

Corneil, D., W. Gerstner, and J. Brea. (2018). "Efficient model-based deep reinforcement learning with variational state tabulation". *arXiv preprint arXiv:1802.04325*.

Coulom, R. (2006). "Efficient selectivity and backup operators in Monte-Carlo tree search". In: *International conference on computers and games*. Springer. 72–83.

Craik, K. J. W. (1943). "The Nature of Explanation".

Da Silva, B. C., E. W. Basso, A. L. Bazzan, and P. M. Engel. (2006). "Dealing with non-stationary environments using context detection". In: *Proceedings of the 23rd international conference on Machine learning*. ACM. 217–224.

Daniel, C., H. Van Hoof, J. Peters, and G. Neumann. (2016). "Probabilistic inference for determining options in reinforcement learning". *Machine Learning*. 104(2-3): 337–357.

Dann, C. and E. Brunskill. (2015). "Sample complexity of episodic fixed-horizon reinforcement learning". In: *Proceedings of the 28th International Conference on Neural Information Processing Systems-Volume 2*. 2818–2826.

Dayan, P. (1993). "Improving generalization for temporal difference learning: The successor representation". *Neural Computation*. 5(4): 613–624.

Dayan, P. and G. E. Hinton. (1993). "Feudal reinforcement learning". In: *Advances in neural information processing systems*. 271–278.

Dearden, R., N. Friedman, and D. Andre. (1999). "Model based Bayesian exploration". In: *Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence*. Morgan Kaufmann Publishers Inc. 150–159.

Dearden, R., N. Friedman, and S. Russell. (1998). "Bayesian Q-learning". In: *AAAI/IAAI*. 761–768.

Degrave, J., M. Hermans, J. Dambre, *et al.* (2019). "A differentiable physics engine for deep learning in robotics". *Frontiers in neurorobotics*. 13.

Deisenroth, M. and C. E. Rasmussen. (2011). "PILCO: A model-based and data-efficient approach to policy search". In: *Proceedings of the 28th International Conference on machine learning (ICML-11)*. 465–472.

Depeweg, S., J. M. Hernández-Lobato, F. Doshi-Velez, and S. Udluft. (2016). "Learning and policy search in stochastic dynamical systems with bayesian neural networks". *arXiv preprint arXiv:1605.07127*.

Der Kiureghian, A. and O. Ditlevsen. (2009). "Aleatory or epistemic? Does it matter?" *Structural Safety*. 31(2): 105–112.

Dilokthanakul, N., C. Kaplanis, N. Pawlowski, and M. Shanahan. (2019). "Feature control as intrinsic motivation for hierarchical reinforcement learning". *IEEE transactions on neural networks and learning systems*.

Diuk, C., A. Cohen, and M. L. Littman. (2008). "An object-oriented representation for efficient reinforcement learning". In: *Proceedings of the 25th international conference on Machine learning*. ACM. 240–247.

Doll, B. B., D. A. Simon, and N. D. Daw. (2012). "The ubiquity of model-based reinforcement learning". *Current opinion in neurobiology*. 22(6): 1075–1081.

Doya, K., K. Samejima, K.-i. Katagiri, and M. Kawato. (2002). "Multiple model-based reinforcement learning". *Neural computation*. 14(6): 1347–1369.

Duff, M. O. and A. Barto. (2002). "Optimal Learning: Computational procedures for Bayes-adaptive Markov decision processes". *PhD thesis*. University of Massachusetts at Amherst.

Ecoffet, A., J. Huizinga, J. Lehman, K. O. Stanley, and J. Clune. (2019). "Go-explore: a new approach for hard-exploration problems". *arXiv preprint arXiv:1901.10995*.

Edwards, A. D., L. Downs, and J. C. Davidson. (2018). "Forward-backward reinforcement learning". *arXiv preprint arXiv:1803.10227*.

Efroni, Y., G. Dalal, B. Scherrer, and S. Mannor. (2018). "Beyond the One-Step Greedy Approach in Reinforcement Learning". In: *International Conference on Machine Learning*. 1386–1395.

Efroni, Y., M. Ghavamzadeh, and S. Mannor. (2019a). "Multi-Step Greedy and Approximate Real Time Dynamic Programming". *arXiv preprint arXiv:1909.04236*.

Efroni, Y., N. Merlis, M. Ghavamzadeh, and S. Mannor. (2019b). "Tight Regret Bounds for Model-Based Reinforcement Learning with Greedy Policies". *arXiv preprint arXiv:1905.11527*.

El Hihi, S. and Y. Bengio. (1996). "Hierarchical recurrent neural networks for long-term dependencies". In: *Advances in neural information processing systems*. 493–499.

Evans, J. S. B. (1984). "Heuristic and analytic processes in reasoning". *British Journal of Psychology*. 75(4): 451–468.

Eysenbach, B., R. R. Salakhutdinov, and S. Levine. (2019a). "Search on the Replay Buffer: Bridging Planning and Reinforcement Learning". *Advances in Neural Information Processing Systems*. 32.

Eysenbach, B., A. Gupta, J. Ibarz, and S. Levine. (2019b). "Diversity is All You Need: Learning Skills without a Reward Function". In: *International Conference on Learning Representations*.

Fairbank, M. and E. Alonso. (2012). "Value-gradient learning". In: *The 2012 International Joint Conference on Neural Networks (IJCNN)*. IEEE. 1–8.

Farquhar, G., T. Rocktäschel, M. Igl, and S. Whiteson. (2018). "Treeqn and atreec: Differentiable tree planning for deep reinforcement learning". In: International Conference on Learning Representations.

Finn, C., I. Goodfellow, and S. Levine. (2016). "Unsupervised learning for physical interaction through video prediction". In: *Advances in neural information processing systems*. 64–72.

Florensa, C., Y. Duan, and P. Abbeel. (2017). "Stochastic neural networks for hierarchical reinforcement learning". *arXiv preprint arXiv:1704.03012*.

Florensa, C., D. Held, X. Geng, and P. Abbeel. (2018). "Automatic Goal Generation for Reinforcement Learning Agents". In: *International Conference on Machine Learning*. 1514–1523.

Fortunato, M., M. Tan, R. Faulkner, S. Hansen, A. P. Badia, G. Buttimore, C. Deck, J. Z. Leibo, and C. Blundell. (2019). "Generalization of reinforcement learners with working and episodic memory". *arXiv preprint arXiv:1910.13406*.

Fox, R., S. Krishnan, I. Stoica, and K. Goldberg. (2017). "Multi-level discovery of deep options". *arXiv preprint arXiv:1703.08294*.

Fox, R., M. Moshkovitz, and N. Tishby. (2016). "Principled option learning in Markov decision processes". *arXiv preprint arXiv:1609.05524*.

Fraccaro, M., S. Kamronn, U. Paquet, and O. Winther. (2017). "A disentangled recognition and nonlinear dynamics model for unsupervised learning". In: *Advances in Neural Information Processing Systems*. 3601–3610.

Fragkiadaki, K., P. Agrawal, S. Levine, and J. Malik. (2015). "Learning visual predictive models of physics for playing billiards". *arXiv preprint arXiv:1511.07404*.

François-Lavet, V., Y. Bengio, D. Precup, and J. Pineau. (2019). "Combined reinforcement learning via abstract representations". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 33. 3582–3589.

Frans, K., J. Ho, X. Chen, P. Abbeel, and J. Schulman. (2018). "Meta learning shared hierarchies". In: *International Conference on Learning Representations*.

Fröhlich, F., F. J. Theis, and J. Hasenauer. (2014). "Uncertainty analysis for non-identifiable dynamical systems: Profile likelihoods, bootstrapping and more". In: *International Conference on Computational Methods in Systems Biology*. Springer. 61–72.

Gal, Y., R. McAllister, and C. E. Rasmussen. (2016). "Improving PILCO with Bayesian neural network dynamics models". In: *Data-Efficient Machine Learning workshop, ICML*. Vol. 4.

Gemici, M., C.-C. Hung, A. Santoro, G. Wayne, S. Mohamed, D. J. Rezende, D. Amos, and T. Lillicrap. (2017). "Generative temporal models with memory". *arXiv preprint arXiv:1702.04649*.

Gershman, S. J. and N. D. Daw. (2017). "Reinforcement learning and episodic memory in humans and animals: an integrative framework". *Annual review of psychology*. 68: 101–128.

Ghahramani, Z. and G. E. Hinton. (1996). "Parameter estimation for linear dynamical systems". *Tech. rep.* Technical Report CRG-TR-96-2, University of Totronto, Dept. of Computer Science.

Ghahramani, Z. and S. T. Roweis. (1999). "Learning nonlinear dynamical systems using an EM algorithm". In: *Advances in neural information processing systems*. 431–437.

Ghavamzadeh, M., S. Mannor, J. Pineau, A. Tamar, *et al.* (2015). "Bayesian reinforcement learning: A survey". *Foundations and Trends® in Machine Learning*. 8(5-6): 359–483. DOI: 10.1561/2200000049.

Ghosh, D., A. Gupta, and S. Levine. (2018). "Learning Actionable Representations with Goal-Conditioned Policies". *arXiv preprint arXiv:1811.07819*.

Goel, S. and M. Huber. (2003). "Subgoal discovery for hierarchical reinforcement learning using learned policies". In: *FLAIRS conference*. 346–350.

Goodfellow, I., Y. Bengio, and A. Courville. (2016). *Deep learning*. MIT press.

Gopalan, A. and S. Mannor. (2015). "Thompson sampling for learning parameterized markov decision processes". In: *Conference on Learning Theory*. PMLR. 861–898.

Graves, A., G. Wayne, and I. Danihelka. (2014). "Neural turing machines". *arXiv preprint arXiv:1410.5401*.

Gregor, K., D. J. Rezende, and D. Wierstra. (2016). "Variational intrinsic control". *arXiv preprint arXiv:1611.07507*.

Grimm, C., A. Barreto, S. Singh, and D. Silver. (2020). "The Value Equivalence Principle for Model-Based Reinforcement Learning". *Advances in Neural Information Processing Systems*.

Gu, S., T. Lillicrap, I. Sutskever, and S. Levine. (2016). "Continuous deep q-learning with model-based acceleration". In: *International Conference on Machine Learning*. 2829–2838.

Guestrin, C., D. Koller, C. Gearhart, and N. Kanodia. (2003). "Generalizing plans to new environments in relational MDPs". In: *Proceedings of the 18th international joint conference on Artificial intelligence*. Morgan Kaufmann Publishers Inc. 1003–1010.

Guez, A., M. Mirza, K. Gregor, R. Kabra, S. Racaniere, T. Weber, D. Raposo, A. Santoro, L. Orseau, T. Eccles, *et al.* (2019). "An Investigation of Model-Free Planning". In: *International Conference on Machine Learning*. 2464–2473.

Guez, A., D. Silver, and P. Dayan. (2012). "Efficient Bayes-adaptive reinforcement learning using sample-based search". In: *Advances in neural information processing systems*. 1025–1033.

Guez, A., T. Weber, I. Antonoglou, K. Simonyan, O. Vinyals, D. Wierstra, R. Munos, and D. Silver. (2018). "Learning to search with MCTSnets". *arXiv preprint arXiv:1802.04697*.

Guo, X., S. Singh, H. Lee, R. L. Lewis, and X. Wang. (2014). "Deep learning for real-time Atari game play using offline Monte-Carlo tree search planning". In: *Advances in neural information processing systems*. 3338–3346.

Ha, D. and J. Schmidhuber. (2018). "Recurrent world models facilitate policy evolution". In: *Advances in Neural Information Processing Systems*. 2450–2462.

Hafner, D., T. Lillicrap, J. Ba, and M. Norouzi. (2019a). "Dream to Control: Learning Behaviors by Latent Imagination". In: *International Conference on Learning Representations*.

Hafner, D., T. Lillicrap, I. Fischer, R. Villegas, D. Ha, H. Lee, and J. Davidson. (2019b). "Learning latent dynamics for planning from pixels". In: *International Conference on Machine Learning*. PMLR. 2555–2565.

Hamidi, M., P. Tadepalli, R. Goetschalckx, and A. Fern. (2015). "Active imitation learning of hierarchical policies". In: *Twenty-Fourth International Joint Conference on Artificial Intelligence*.

Hamrick, J. B. (2019). "Analogues of mental simulation and imagination in deep learning". *Current Opinion in Behavioral Sciences*. 29: 8–16.

Hamrick, J. B., A. J. Ballard, R. Pascanu, O. Vinyals, N. Heess, and P. W. Battaglia. (2017). "Metacontrol for adaptive imagination-based optimization". *arXiv preprint arXiv:1705.02670*.

Hamrick, J. B., V. Bapst, A. Sanchez-Gonzalez, T. Pfaff, T. Weber, L. Buesing, and P. W. Battaglia. (2020). "Combining q-learning and search with amortized value estimates". *International Conference on Learning Representations (ICLR)*.

Hasselt, H. P. van, M. Hessel, and J. Aslanides. (2019). "When to use parametric models in reinforcement learning?" In: *Advances in Neural Information Processing Systems*. 14322–14333.

Hausman, K., J. T. Springenberg, Z. Wang, N. Heess, and M. Riedmiller. (2018). "Learning an Embedding Space for Transferable Robot Skills". In: *International Conference on Learning Representations*.

Heess, N., G. Wayne, Y. Tassa, T. Lillicrap, M. Riedmiller, and D. Silver. (2016). "Learning and transfer of modulated locomotor controllers". *arXiv preprint arXiv:1610.05182*.

Heess, N., G. Wayne, D. Silver, T. Lillicrap, T. Erez, and Y. Tassa. (2015). "Learning continuous control policies by stochastic value gradients". In: *Advances in Neural Information Processing Systems*. 2944–2952.

Henderson, P., R. Islam, P. Bachman, J. Pineau, D. Precup, and D. Meger. (2018). "Deep reinforcement learning that matters". In: *Thirty-Second AAAI Conference on Artificial Intelligence*.

Hengst, B. (2017). "Hierarchical reinforcement learning". *Encyclopedia of Machine Learning and Data Mining*: 611–619.

Hester, T. and P. Stone. (2012a). "Intrinsically motivated model learning for a developing curious agent". In: *2012 IEEE international conference on development and learning and epigenetic robotics (ICDL)*. IEEE. 1–6.

Hester, T. and P. Stone. (2012b). "Learning and using models". In: *Reinforcement learning*. Springer. 111–141.

Hester, T. and P. Stone. (2013). "TEXPLORE: real-time sample-efficient reinforcement learning for robots". *Machine learning*. 90(3): 385–429.

Hochreiter, S. and J. Schmidhuber. (1997). "Long short-term memory". *Neural computation*. 9(8): 1735–1780.

Holland, G. Z., E. J. Talvitie, and M. Bowling. (2018). "The effect of planning shape on dyna-style planning in high-dimensional state spaces". *arXiv preprint arXiv:1806.01825*.

Houthooft, R., X. Chen, Y. Duan, J. Schulman, F. De Turck, and P. Abbeel. (2016). "Vime: Variational information maximizing exploration". In: *Advances in Neural Information Processing Systems*. 1109–1117.

Hu, H., J. Ye, G. Zhu, Z. Ren, and C. Zhang. (2021). "Generalizable episodic memory for deep reinforcement learning". *arXiv preprint arXiv:2103.06469*.

Jaderberg, M., V. Mnih, W. M. Czarnecki, T. Schaul, J. Z. Leibo, D. Silver, and K. Kavukcuoglu. (2016). "Reinforcement learning with unsupervised auxiliary tasks". *arXiv preprint arXiv:1611.05397*.

Jaksch, T., R. Ortner, and P. Auer. (2010). "Near-optimal Regret Bounds for Reinforcement Learning". *Journal of Machine Learning Research*. 11(4).

Janner, M., J. Fu, M. Zhang, and S. Levine. (2019). "When to trust your model: Model-based policy optimization". In: *Advances in Neural Information Processing Systems*. 12519–12530.

Jaulmes, R., J. Pineau, and D. Precup. (2005). "Learning in non-stationary partially observable Markov decision processes". In: *ECML Workshop on Reinforcement Learning in non-stationary environments*. Vol. 25. 26–32.

Jayaraman, D., F. Ebert, A. A. Efros, and S. Levine. (2018). "Time-agnostic prediction: Predicting predictable video frames". *arXiv preprint arXiv:1808.07784*.

Jiang, D., E. Ekwedike, and H. Liu. (2018). "Feedback-Based Tree Search for Reinforcement Learning". In: *International Conference on Machine Learning*. 2289–2298.

Jin, C., Z. Allen-Zhu, S. Bubeck, and M. I. Jordan. (2018). “Is Q-learning provably efficient?” In: *Proceedings of the 32nd International Conference on Neural Information Processing Systems*. 4868–4878.

Jong, N. K. and P. Stone. (2007). “Model-based function approximation in reinforcement learning”. In: *Proceedings of the 6th international joint conference on Autonomous agents and multiagent systems*. ACM. 95.

Jonschkowski, R. and O. Brock. (2015). “Learning state representations with robotic priors”. *Autonomous Robots*. 39(3): 407–428.

Jordan, M. I. and D. E. Rumelhart. (1992). “Forward models: Supervised learning with a distal teacher”. *Cognitive science*. 16(3): 307–354.

Kahneman, D. (2011). *Thinking, fast and slow*. Macmillan.

Kakade, S. and J. Langford. (2002). “Approximately optimal approximate reinforcement learning”. In: *In Proc. 19th International Conference on Machine Learning*. Citeseer.

Kakade, S., M. Wang, and L. F. Yang. (2018). “Variance reduction methods for sublinear reinforcement learning”.

Kakade, S. M. *et al.* (2003). “On the sample complexity of reinforcement learning”. *PhD thesis*. University of London London, England.

Kalchbrenner, N., A. van den Oord, K. Simonyan, I. Danihelka, O. Vinyals, A. Graves, and K. Kavukcuoglu. (2017). “Video pixel networks”. In: *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org. 1771–1779.

Kalweit, G. and J. Boedecker. (2017). “Uncertainty-driven imagination for continuous deep reinforcement learning”. In: *Conference on Robot Learning*. 195–206.

Kamthe, S. and M. P. Deisenroth. (2017). “Data-efficient reinforcement learning with probabilistic model predictive control”. *arXiv preprint arXiv:1706.06491*.

Kansky, K., T. Silver, D. A. Mély, M. Eldawy, M. Lázaro-Gredilla, X. Lou, N. Dorfman, S. Sidor, S. Phoenix, and D. George. (2017). “Schema networks: Zero-shot transfer with a generative causal model of intuitive physics”. In: *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org. 1809–1818.

Karl, M., M. Soelch, J. Bayer, and P. van der Smagt. (2016). "Deep variational bayes filters: Unsupervised learning of state space models from raw data". *arXiv preprint arXiv:1605.06432*.

Ke, N. R., A. Singh, A. Touati, A. Goyal, Y. Bengio, D. Parikh, and D. Batra. (2019). "Learning Dynamics Model in Reinforcement Learning by Incorporating the Long Term Future". *arXiv preprint arXiv:1903.01599*.

Keramati, M., A. Dezfouli, and P. Piray. (2011). "Speed/accuracy trade-off between the habitual and the goal-directed processes". *PLoS computational biology*. 7(5).

Khansari-Zadeh, S. M. and A. Billard. (2011). "Learning stable non-linear dynamical systems with gaussian mixture models". *IEEE Transactions on Robotics*. 27(5): 943–957.

Kipf, T., E. van der Pol, and M. Welling. (2020). "Contrastive Learning of Structured World Models". In: *International Conference on Learning Representations*.

Kober, J., J. A. Bagnell, and J. Peters. (2013). "Reinforcement learning in robotics: A survey". *The International Journal of Robotics Research*. 32(11): 1238–1274.

Kocsis, L. and C. Szepesvári. (2006). "Bandit based monte-carlo planning". In: *ECML*. Vol. 6. Springer. 282–293.

Kolter, J. Z. and A. Y. Ng. (2009). "Near-Bayesian exploration in polynomial time". In: *Proceedings of the 26th Annual International Conference on Machine Learning*. ACM. 513–520.

Konidaris, G. and A. G. Barto. (2007). "Building Portable Options: Skill Transfer in Reinforcement Learning." In: *IJCAI*. Vol. 7. 895–900.

Konidaris, G., S. Kuindersma, R. Grupen, and A. Barto. (2012). "Robot learning from demonstration by constructing skill trees". *The International Journal of Robotics Research*. 31(3): 360–375.

Konidaris, G. D. (2006). "A framework for transfer in reinforcement learning". In: *ICML-06 Workshop on Structural Knowledge Transfer for Machine Learning*.

Korf, R. E. (1990). "Real-time heuristic search". *Artificial intelligence*. 42(2-3): 189–211.

Krishnan, R. G., U. Shalit, and D. A. Sontag. (2015). "Deep Kalman Filters". *ArXiv*. abs/1511.05121.

Krizhevsky, A., I. Sutskever, and G. E. Hinton. (2012). “Imagenet classification with deep convolutional neural networks”. In: *Advances in neural information processing systems*. 1097–1105.

Kulkarni, T. D., K. Narasimhan, A. Saeedi, and J. Tenenbaum. (2016). “Hierarchical deep reinforcement learning: Integrating temporal abstraction and intrinsic motivation”. In: *Advances in neural information processing systems*. 3675–3683.

Kurniawati, H., D. Hsu, and W. S. Lee. (2008). “Sarsop: Efficient point-based pomdp planning by approximating optimally reachable belief spaces.” In: *Robotics: Science and systems*. Vol. 2008. Zurich, Switzerland.

Kurutach, T., A. Tamar, G. Yang, S. J. Russell, and P. Abbeel. (2018). “Learning plannable representations with causal infogan”. In: *Advances in Neural Information Processing Systems*. 8733–8744.

Lai, M. (2015). “Giraffe: Using deep reinforcement learning to play chess”. *arXiv preprint arXiv:1509.01549*.

Lai, T. L. and H. Robbins. (1985). “Asymptotically efficient adaptive allocation rules”. *Advances in applied mathematics*. 6(1): 4–22.

Lakshminarayanan, A. S., R. Krishnamurthy, P. Kumar, and B. Ravindran. (2016). “Option discovery in hierarchical reinforcement learning using spatio-temporal clustering”. *arXiv preprint arXiv:1605.05359*.

Lange, S., T. Gabel, and M. Riedmiller. (2012). “Batch reinforcement learning”. In: *Reinforcement learning*. Springer. 45–73.

LaValle, S. M. (1998). “Rapidly-exploring random trees: A new tool for path planning”.

Laversanne-Finot, A., A. Pere, and P.-Y. Oudeyer. (2018). “Curiosity Driven Exploration of Learned Disentangled Goal Spaces”. In: *Conference on Robot Learning*. 487–504.

Lazaric, A. (2012). “Transfer in reinforcement learning: a framework and a survey”. In: *Reinforcement Learning*. Springer. 143–173.

Lesort, T., N. Díaz-Rodríguez, J.-F. Goudou, and D. Filliat. (2018). “State representation learning for control: An overview”. *Neural Networks*. 108: 379–392.

Levine, S. and P. Abbeel. (2014). "Learning neural network policies with guided policy search under unknown dynamics". In: *Advances in Neural Information Processing Systems*. 1071–1079.

Levine, S. and V. Koltun. (2013). "Guided policy search". In: *International Conference on Machine Learning*. 1–9.

Levy, A., R. Platt, and K. Saenko. (2019). "Hierarchical Reinforcement Learning with Hindsight". In: *International Conference on Learning Representations*.

Lin, L.-J. (1992). "Self-improving reactive agents based on reinforcement learning, planning and teaching". *Machine learning*. 8(3-4): 293–321.

Lin, L.-J. (1993). "Reinforcement learning for robots using neural networks". *Tech. rep.* Carnegie-Mellon Univ Pittsburgh PA School of Computer Science.

Lin, L.-J. and T. M. Mitchell. (1992). *Memory approaches to reinforcement learning in non-Markovian domains*. Citeseer.

Lin, Z., T. Zhao, G. Yang, and L. Zhang. (2018). "Episodic memory deep Q-networks". *arXiv preprint arXiv:1805.07603*.

Ljung, L. (2001). "System identification". *Wiley Encyclopedia of Electrical and Electronics Engineering*.

Lopes, M., T. Lang, M. Toussaint, and P.-Y. Oudeyer. (2012). "Exploration in model-based reinforcement learning by empirically estimating learning progress". In: *Advances in neural information processing systems*. 206–214.

Lowrey, K., A. Rajeswaran, S. Kakade, E. Todorov, and I. Mordatch. (2018). "Plan online, learn offline: Efficient learning and exploration via model-based control". *arXiv preprint arXiv:1811.01848*.

Loynd, R., M. Hausknecht, L. Li, and L. Deng. (2018). "Now I Remember! Episodic Memory For Reinforcement Learning".

Lu, K., I. Mordatch, and P. Abbeel. (2019). "Adaptive Online Planning for Continual Lifelong Learning". *arXiv preprint arXiv:1912.01188*.

Machado, M. C., M. G. Bellemare, and M. Bowling. (2017). "A laplacian framework for option discovery in reinforcement learning". In: *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org. 2295–2304.

Machado, M. C., M. G. Bellemare, E. Talvitie, J. Veness, M. Hausknecht, and M. Bowling. (2018). "Revisiting the arcade learning environment: Evaluation protocols and open problems for general agents". *Journal of Artificial Intelligence Research*. 61: 523–562.

Mahadevan, S. (2009). "Learning Representation and Control in Markov Decision Processes: New Frontiers". *Foundations and Trends® in Machine Learning*. 1(4): 403–565. DOI: 10.1561/2200000003.

Mann, T. and S. Mannor. (2014). "Scaling up approximate value iteration with options: Better policies with fewer iterations". In: *International conference on machine learning*. 127–135.

Mannor, S., I. Menache, A. Hoze, and U. Klein. (2004). "Dynamic abstraction in reinforcement learning via clustering". In: *Proceedings of the twenty-first international conference on Machine learning*. ACM. 71.

Matiisen, T., A. Oliver, T. Cohen, and J. Schulman. (2017). "Teacher-student curriculum learning". *arXiv preprint arXiv:1707.00183*.

McCallum, R. (1997). "Reinforcement learning with selective perception and hidden state".

McGovern, A. and A. G. Barto. (2001). "Automatic discovery of subgoals in reinforcement learning using diverse density".

Menache, I., S. Mannor, and N. Shimkin. (2002). "Q-cut—dynamic discovery of sub-goals in reinforcement learning". In: *European Conference on Machine Learning*. Springer. 295–306.

Mishra, N., P. Abbeel, and I. Mordatch. (2017). "Prediction and control with temporal segment models". In: *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org. 2459–2468.

Mnih, V., K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, *et al.* (2015). "Human-level control through deep reinforcement learning". *Nature*. 518(7540): 529.

Moerland, T. M., J. Broekens, and C. M. Jonker. (2017a). "Efficient exploration with double uncertain value networks". *Deep Reinforcement Learning Symposium, 31st Conference on Neural Information Processing Systems (NIPS)*.

Moerland, T. M., J. Broekens, and C. M. Jonker. (2017b). "Learning Multimodal Transition Dynamics for Model-Based Reinforcement Learning". *Scaling Up Reinforcement Learning (SURL) Workshop, European Conference on Machine Learning (ECML)*.

Moerland, T. M., J. Broekens, and C. M. Jonker. (2018a). "Emotion in reinforcement learning agents and robots: a survey". *Machine Learning*. 107(2): 443–480.

Moerland, T. M., J. Broekens, A. Plaat, and C. M. Jonker. (2018b). "A0C: Alpha zero in continuous action space". *Planning and Learning Workshop, 35th International Conference on Machine Learning (ICML)*.

Moerland, T. M., A. Deichler, S. Baldi, J. Broekens, and C. M. Jonker. (2020a). "Think Too Fast Nor Too Slow: The Computational Tradeoff Between Planning And Reinforcement Learning". *arXiv preprint arXiv:2005.07404*.

Moerland, T. M., J. Broekens, and C. M. Jonker. (2020b). "A Framework for Reinforcement Learning and Planning". *arXiv preprint arXiv:2006.15009*.

Momennejad, I., E. M. Russek, J. H. Cheong, M. M. Botvinick, N. D. Daw, and S. J. Gershman. (2017). "The successor representation in human reinforcement learning". *Nature Human Behaviour*. 1(9): 680.

Moore, A. W. and C. G. Atkeson. (1993). "Prioritized sweeping: Reinforcement learning with less data and less time". *Machine learning*. 13(1): 103–130.

Müller, K.-R., A. J. Smola, G. Rätsch, B. Schölkopf, J. Kohlmorgen, and V. Vapnik. (1997). "Predicting time series with support vector machines". In: *International Conference on Artificial Neural Networks*. Springer. 999–1004.

Munos, R. (2003). "Error bounds for approximate policy iteration". In: *ICML*. Vol. 3. 560–567.

Nachum, O., S. S. Gu, H. Lee, and S. Levine. (2018). "Data-efficient hierarchical reinforcement learning". In: *Advances in Neural Information Processing Systems*. 3303–3313.

Nagabandi, A., I. Clavera, S. Liu, R. S. Fearing, P. Abbeel, S. Levine, and C. Finn. (2018a). "Learning to Adapt in Dynamic, Real-World Environments through Meta-Reinforcement Learning". In: *International Conference on Learning Representations*.

Nagabandi, A., C. Finn, and S. Levine. (2018b). "Deep online learning via meta-learning: Continual adaptation for model-based RL". *arXiv preprint arXiv:1812.07671*.

Nagabandi, A., G. Kahn, R. S. Fearing, and S. Levine. (2018c). "Neural network dynamics for model-based deep reinforcement learning with model-free fine-tuning". In: *2018 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE. 7559–7566.

Narendra, K. S. and K. Parthasarathy. (1990). "Identification and control of dynamical systems using neural networks". *IEEE Transactions on neural networks*. 1(1): 4–27.

Neitz, A., G. Parascandolo, S. Bauer, and B. Schölkopf. (2018). "Adaptive skip intervals: Temporal abstraction for recurrent dynamical models". In: *Advances in Neural Information Processing Systems*. 9816–9826.

Nguyen-Tuong, D. and J. Peters. (2011). "Model learning for robot control: a survey". *Cognitive processing*. 12(4): 319–340.

Nouri, A. and M. L. Littman. (2010). "Dimension reduction and its application to model-based exploration in continuous spaces". *Machine Learning*. 81(1): 85–98.

Oh, J., X. Guo, H. Lee, R. L. Lewis, and S. Singh. (2015). "Action-conditional video prediction using deep networks in atari games". In: *Advances in neural information processing systems*. 2863–2871.

Oh, J., S. Singh, and H. Lee. (2017). "Value prediction network". In: *Advances in Neural Information Processing Systems*. 6118–6128.

Osband, I., C. Blundell, A. Pritzel, and B. Van Roy. (2016). "Deep exploration via bootstrapped DQN". In: *Advances in Neural Information Processing Systems*. 4026–4034.

Osband, I., Y. Doron, M. Hessel, J. Aslanides, E. Sezener, A. Saraiva, K. McKinney, T. Lattimore, C. Szepezvari, S. Singh, B. V. Roy, R. Sutton, D. Silver, and H. V. Hasselt. (2019). "Behaviour Suite for Reinforcement Learning". arXiv: 1908.03568 [cs.LG].

Osband, I., D. Russo, and B. Van Roy. (2013). "(More) efficient reinforcement learning via posterior sampling". *arXiv preprint arXiv:1306.0940*.

Osband, I. and B. Van Roy. (2016). "On lower bounds for regret in reinforcement learning". *arXiv preprint arXiv:1608.02732*.

Osband, I. and B. Van Roy. (2017). "Why is posterior sampling better than optimism for reinforcement learning?" In: *International conference on machine learning*. PMLR. 2701–2710.

Ostafew, C. J., A. P. Schoellig, and T. D. Barfoot. (2016). "Robust constrained learning-based NMPC enabling reliable mobile robot path tracking". *The International Journal of Robotics Research*. 35(13): 1547–1563.

Ostrovski, G., M. G. Bellemare, A. van den Oord, and R. Munos. (2017). "Count-based exploration with neural density models". In: *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org. 2721–2730.

Oudeyer, P.-Y., F. Kaplan, and V. V. Hafner. (2007). "Intrinsic motivation systems for autonomous mental development". *IEEE transactions on evolutionary computation*. 11(2): 265–286.

Oudeyer, P.-Y., F. Kaplan, *et al.* (2008). "How can we define intrinsic motivation". In: *Proc. of the 8th Conf. on Epigenetic Robotics*. Vol. 5. 29–31.

Parlos, A. G., K. T. Chong, and A. F. Atiya. (1994). "Application of the recurrent multilayer perceptron in modeling complex process dynamics". *IEEE Transactions on Neural Networks*. 5(2): 255–266.

Parr, R., L. Li, G. Taylor, C. Painter-Wakefield, and M. L. Littman. (2008). "An analysis of linear models, linear value-function approximation, and feature selection for reinforcement learning". In: *Proceedings of the 25th international conference on Machine learning*. ACM. 752–759.

Pascanu, R., Y. Li, O. Vinyals, N. Heess, L. Buesing, S. Racanière, D. Reichert, T. Weber, D. Wierstra, and P. Battaglia. (2017). "Learning model-based planning from scratch". *arXiv preprint arXiv:1707.06170*.

Pathak, D., P. Agrawal, A. A. Efros, and T. Darrell. (2017). "Curiosity-driven exploration by self-supervised prediction". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 16–17.

Péré, A., S. Forestier, O. Sigaud, and P.-Y. Oudeyer. (2018). "Unsupervised learning of goal spaces for intrinsically motivated goal exploration". *arXiv preprint arXiv:1803.00781*.

Peshkin, L., N. Meuleau, and L. P. Kaelbling. (1999). "Learning Policies with External Memory". In: *Proceedings of the Sixteenth International Conference on Machine Learning*. Morgan Kaufmann Publishers Inc. 307–314.

Plaat, A., W. Kosters, and M. Preuss. (2020). "Model-Based Deep Reinforcement Learning for High-Dimensional Problems, a Survey". *arXiv preprint arXiv:2008.05598*.

Plappert, M., R. Houthooft, P. Dhariwal, S. Sidor, R. Y. Chen, X. Chen, T. Asfour, P. Abbeel, and M. Andrychowicz. (2017). "Parameter space noise for exploration". *arXiv preprint arXiv:1706.01905*.

Polydoros, A. S. and L. Nalpantidis. (2017). "Survey of model-based reinforcement learning: Applications on robotics". *Journal of Intelligent & Robotic Systems*. 86(2): 153–173.

Pong, V., S. Gu, M. Dalal, and S. Levine. (2018). "Temporal Difference Models: Model-Free Deep RL for Model-Based Control". In: *International Conference on Learning Representations (ICLR 2018)*. OpenReview. net.

Pritzel, A., B. Uria, S. Srinivasan, A. P. Badia, O. Vinyals, D. Hassabis, D. Wierstra, and C. Blundell. (2017). "Neural Episodic Control". In: *International Conference on Machine Learning*. 2827–2836.

Puterman, M. L. (2014). *Markov Decision Processes.: Discrete Stochastic Dynamic Programming*. John Wiley & Sons.

Racanière, S., T. Weber, D. Reichert, L. Buesing, A. Guez, D. J. Rezende, A. P. Badia, O. Vinyals, N. Heess, Y. Li, *et al.* (2017). "Imagination-augmented agents for deep reinforcement learning". In: *Advances in neural information processing systems*. 5690–5701.

Ramani, D. (2019). "A short survey on memory based reinforcement learning". *arXiv preprint arXiv:1904.06736*.

Riemer, M., M. Liu, and G. Tesauro. (2018). “Learning abstract options”. In: *Advances in Neural Information Processing Systems*. 10424–10434.

Roijers, D. M., P. Vamplew, S. Whiteson, and R. Dazeley. (2013). “A survey of multi-objective sequential decision-making”. *Journal of Artificial Intelligence Research*. 48: 67–113.

Roijers, D. M. and S. Whiteson. (2017). “Multi-objective decision making”. *Synthesis Lectures on Artificial Intelligence and Machine Learning*. 11(1): 1–129.

Rummery, G. A. and M. Niranjan. (1994). *On-line Q-learning using connectionist systems*. Vol. 37. University of Cambridge, Department of Engineering Cambridge, England.

Russell, S. J. and P. Norvig. (2016). *Artificial intelligence: a modern approach*. Malaysia; Pearson Education Limited,

Samuel, A. L. (1967). “Some studies in machine learning using the game of checkers. II - Recent progress.” *IBM Journal of research and development*. 11(6): 601–617.

Sawada, Y. (2018). “Disentangling Controllable and Uncontrollable Factors of Variation by Interacting with the World”. *arXiv preprint arXiv:1804.06955*.

Schaul, T., D. Horgan, K. Gregor, and D. Silver. (2015). “Universal value function approximators”. In: *International Conference on Machine Learning*. 1312–1320.

Schaul, T., J. Quan, I. Antonoglou, and D. Silver. (2016). “Prioritized Experience Replay”. In: *International Conference on Learning Representations (ICLR)*.

Schmidhuber, J. (1991). “Curious model-building control systems”. In: *[Proceedings] 1991 IEEE International Joint Conference on Neural Networks*. IEEE. 1458–1463.

Schrittwieser, J., I. Antonoglou, T. Hubert, K. Simonyan, L. Sifre, S. Schmitt, A. Guez, E. Lockhart, D. Hassabis, T. Graepel, *et al.* (2019). “Mastering atari, go, chess and shogi by planning with a learned model”. *arXiv preprint arXiv:1911.08265*.

Sekar, R., O. Rybkin, K. Daniilidis, P. Abbeel, D. Hafner, and D. Pathak. (2020). “Planning to Explore via Self-Supervised World Models”. *arXiv preprint arXiv:2005.05960*.

Sequeira, P., F. S. Melo, and A. Paiva. (2014). "Learning by appraising: an emotion-based approach to intrinsic reward design". *Adaptive Behavior*. 22(5): 330–349.

Sermanet, P., C. Lynch, Y. Chebotar, J. Hsu, E. Jang, S. Schaal, S. Levine, and G. Brain. (2018). "Time-contrastive networks: Self-supervised learning from video". In: *2018 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE. 1134–1141.

Sharma, A., S. Gu, S. Levine, V. Kumar, and K. Hausman. (2019). "Dynamics-Aware Unsupervised Discovery of Skills". In: *International Conference on Learning Representations*.

Shu, T., C. Xiong, and R. Socher. (2017). "Hierarchical and interpretable skill acquisition in multi-task reinforcement learning". *arXiv preprint arXiv:1712.07294*.

Shyam, P., W. Jaśkowski, and F. Gomez. (2019). "Model-Based Active Exploration". In: *International Conference on Machine Learning*. 5779–5788.

Sigaud, O., C. Salaün, and V. Padois. (2011). "On-line regression algorithms for learning mechanical models of robots: a survey". *Robotics and Autonomous Systems*. 59(12): 1115–1129.

Silver, D., H. van Hasselt, M. Hessel, T. Schaul, A. Guez, T. Harley, G. Dulac-Arnold, D. Reichert, N. Rabinowitz, A. Barreto, *et al.* (2017a). "The predictron: End-to-end learning and planning". In: *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org. 3191–3199.

Silver, D., T. Hubert, J. Schrittwieser, I. Antonoglou, M. Lai, A. Guez, M. Lanctot, L. Sifre, D. Kumaran, T. Graepel, *et al.* (2018). "A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play". *Science*. 362(6419): 1140–1144.

Silver, D., J. Schrittwieser, K. Simonyan, I. Antonoglou, A. Huang, A. Guez, T. Hubert, L. Baker, M. Lai, A. Bolton, *et al.* (2017b). "Mastering the game of go without human knowledge". *Nature*. 550(7676): 354.

Silver, D., R. S. Sutton, and M. Müller. (2008). "Sample-based learning and search with permanent and transient memories". In: *Proceedings of the 25th international conference on Machine learning*. ACM. 968–975.

Silver, D. and J. Veness. (2010). "Monte-Carlo planning in large POMDPs". In: *Advances in neural information processing systems*. 2164–2172.

Şimşek, Ö., A. P. Wolfe, and A. G. Barto. (2005). "Identifying useful subgoals in reinforcement learning by local graph partitioning". In: *Proceedings of the 22nd international conference on Machine learning*. ACM. 816–823.

Singh, S. P., T. Jaakkola, and M. I. Jordan. (1995). "Reinforcement learning with soft state aggregation". In: *Advances in neural information processing systems*. 361–368.

Spaan, M. T. and N. Spaan. (2004). "A point-based POMDP algorithm for robot planning". In: *IEEE International Conference on Robotics and Automation, 2004. Proceedings. ICRA'04. 2004*. Vol. 3. IEEE. 2399–2404.

Spelke, E. S. and K. D. Kinzler. (2007). "Core knowledge". *Developmental science*. 10(1): 89–96.

Srinivas, A., A. Jabri, P. Abbeel, S. Levine, and C. Finn. (2018). "Universal planning networks". *arXiv preprint arXiv:1804.00645*.

Stadie, B. C., S. Levine, and P. Abbeel. (2015). "Incentivizing exploration in reinforcement learning with deep predictive models". *arXiv preprint arXiv:1507.00814*.

Strehl, A. L., L. Li, and M. L. Littman. (2006). "Pac reinforcement learning bounds for rtdp and rand-rtdp". In: *Proceedings of AAAI workshop on learning for search*.

Sutton, R. S. (1990). "Integrated architectures for learning, planning, and reacting based on approximating dynamic programming". In: *Machine Learning Proceedings 1990*. Elsevier. 216–224.

Sutton, R. S. (1991). "Dyna, an integrated architecture for learning, planning, and reacting". *ACM Sigart Bulletin*. 2(4): 160–163.

Sutton, R. S. and A. G. Barto. (2018). *Reinforcement learning: An introduction*. MIT press.

Sutton, R. S., D. Precup, and S. Singh. (1999). "Between MDPs and semi-MDPs: A framework for temporal abstraction in reinforcement learning". *Artificial intelligence*. 112(1-2): 181–211.

Sutton, R. S., C. Szepesvári, A. Geramifard, and M. P. Bowling. (2012). "Dyna-style planning with linear function approximation and prioritized sweeping". *arXiv preprint arXiv:1206.3285*.

Sutton, R. S., C. Szepesvári, A. Geramifard, and M. Bowling. (2008). "Dyna-style Planning with Linear Function Approximation and Prioritized Sweeping". In: *Proceedings of the Twenty-Fourth Conference on Uncertainty in Artificial Intelligence. UAI'08*. Helsinki, Finland: AUAI Press. 528–536.

Szita, I. and C. Szepesvári. (2010). "Model-based reinforcement learning with nearly tight exploration complexity bounds". In: *ICML*.

Talvitie, E. (2014). "Model Regularization for Stable Sample Rollouts." In: *UAI*. 780–789.

Talvitie, E. (2017). "Self-correcting models for model-based reinforcement learning". In: *Thirty-First AAAI Conference on Artificial Intelligence*.

Tamar, A., Y. Wu, G. Thomas, S. Levine, and P. Abbeel. (2016). "Value iteration networks". In: *Advances in Neural Information Processing Systems*. 2154–2162.

Taylor, M. E. and P. Stone. (2009). "Transfer learning for reinforcement learning domains: A survey". *Journal of Machine Learning Research*. 10(Jul): 1633–1685.

Tesauro, G. and G. R. Galperin. (1997). "On-line Policy Improvement using Monte-Carlo Search". In: *Advances in Neural Information Processing Systems 9*. Ed. by M. C. Mozer, M. I. Jordan, and T. Petsche. MIT Press. 1068–1074.

Tessler, C., S. Givony, T. Zahavy, D. J. Mankowitz, and S. Mannor. (2017). "A deep hierarchical approach to lifelong learning in minecraft". In: *Thirty-First AAAI Conference on Artificial Intelligence*.

Thomas, V., E. Bengio, W. Fedus, J. Pondard, P. Beaudoin, H. Larochelle, J. Pineau, D. Precup, and Y. Bengio. (2018). "Disentangling the independently controllable factors of variation by interacting with the world". *arXiv preprint arXiv:1802.09484*.

Thompson, W. R. (1933). "On the likelihood that one unknown probability exceeds another in view of the evidence of two samples". *Biometrika*. 25(3/4): 285–294.

Thrun, S. and A. Schwartz. (1995). "Finding structure in reinforcement learning". In: *Advances in neural information processing systems*. 385–392.

Thrun, S. B. (1992). "Efficient exploration in reinforcement learning".

Tobin, J., 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: *2017 IEEE/RSJ international conference on intelligent robots and systems (IROS)*. IEEE. 23–30.

Todorov, E. and W. Li. (2005). "A generalized iterative LQG method for locally-optimal feedback control of constrained nonlinear stochastic systems". In: *Proceedings of the 2005, American Control Conference, 2005*. IEEE. 300–306.

Tolman, E. C. (1948). "Cognitive maps in rats and men". *Psychological review*. 55(4): 189.

Van Hoof, H., N. Chen, M. Karl, P. van der Smagt, and J. Peters. (2016). "Stable reinforcement learning with autoencoders for tactile and visual data". In: *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE. 3928–3934.

Van Seijen, H., H. Nekoei, E. Racah, and S. Chandar. (2020). "The LoCA Regret: A Consistent Metric to Evaluate Model-Based Behavior in Reinforcement Learning". *Advances in Neural Information Processing Systems*. 33.

Van Steenkiste, S., M. Chang, K. Greff, and J. Schmidhuber. (2018). "Relational neural expectation maximization: Unsupervised discovery of objects and their interactions". *arXiv preprint arXiv:1802.10353*.

Vanseijen, H. and R. Sutton. (2015). "A deeper look at planning as learning from replay". In: *International conference on machine learning*. 2314–2322.

Veness, J., D. Silver, A. Blair, and W. Uther. (2009). "Bootstrapping from game tree search". In: *Advances in neural information processing systems*. 1937–1945.

Venkatraman, A., M. Hebert, and J. A. Bagnell. (2015). "Improving multi-step prediction of learned time series models". In: *Twenty-Ninth AAAI Conference on Artificial Intelligence*.

Vezhnevets, A. S., S. Osindero, T. Schaul, N. Heess, M. Jaderberg, D. Silver, and K. Kavukcuoglu. (2017). "Feudal networks for hierarchical reinforcement learning". In: *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org. 3540–3549.

Waa, J. van der, J. van Diggelen, K. v. d. Bosch, and M. Neerincx. (2018). "Contrastive explanations for reinforcement learning in terms of expected consequences". *arXiv preprint arXiv:1807.08706*.

Wahlström, N., T. B. Schön, and M. P. Deisenroth. (2015). "From pixels to torques: Policy learning with deep dynamical models". *arXiv preprint arXiv:1502.02251*.

Wang, J., A. Hertzmann, and D. J. Fleet. (2006). "Gaussian process dynamical models". In: *Advances in neural information processing systems*. 1441–1448.

Wang, T., X. Bao, I. Clavera, J. Hoang, Y. Wen, E. Langlois, S. Zhang, G. Zhang, P. Abbeel, and J. Ba. (2019). "Benchmarking Model-Based Reinforcement Learning". *CoRR*. abs/1907.02057. arXiv: 1907.02057.

Watkins, C. J. and P. Dayan. (1992). "Q-learning". *Machine learning*. 8(3-4): 279–292.

Watson, J. S. (1966). "The development and generalization of" contingency awareness" in early infancy: Some hypotheses". *Merrill-Palmer Quarterly of Behavior and Development*. 12(2): 123–135.

Watter, M., J. Springenberg, J. Boedecker, and M. Riedmiller. (2015). "Embed to control: A locally linear latent dynamics model for control from raw images". In: *Advances in neural information processing systems*. 2746–2754.

Watters, N., L. Matthey, M. Bosnjak, C. P. Burgess, and A. Lerchner. (2019). "Cobra: Data-efficient model-based rl through unsupervised object discovery and curiosity-driven exploration". *arXiv preprint arXiv:1905.09275*.

Werbos, P. J. (1989). "Neural networks for control and system identification". In: *Proceedings of the 28th IEEE Conference on Decision and Control*. IEEE. 260–265.

Wiering, M. and J. Schmidhuber. (1998). “Efficient model-based exploration”. In: *Proceedings of the Sixth International Conference on Simulation of Adaptive Behavior: From Animals to Animats*. Vol. 6. 223–228.

Wiering, M. A., M. Withagen, and M. M. Drugan. (2014). “Model-based multi-objective reinforcement learning”. In: *2014 IEEE Symposium on Adaptive Dynamic Programming and Reinforcement Learning (ADPRL)*. IEEE. 1–6.

Willemsen, D., H. Baier, and M. Kaisers. (2020). “Value targets in off-policy AlphaZero: a new greedy backup”. In: *Adaptive and Learning Agents (ALA) Workshop*.

Williams, R. J. (1992). “Simple statistical gradient-following algorithms for connectionist reinforcement learning”. *Machine learning*. 8(3-4): 229–256.

Wilson, A., A. Fern, S. Ray, and P. Tadepalli. (2007). “Multi-task reinforcement learning: a hierarchical Bayesian approach”. In: *Proceedings of the 24th international conference on Machine learning*. ACM. 1015–1022.

Wolpert, D. M., Z. Ghahramani, and M. I. Jordan. (1995). “An internal model for sensorimotor integration”. *Science*. 269(5232): 1880–1882.

Wu, J., I. Yildirim, J. J. Lim, B. Freeman, and J. Tenenbaum. (2015). “Galileo: Perceiving physical object properties by integrating a physics engine with deep learning”. *Advances in neural information processing systems*. 28: 127–135.

Xu, Z., Z. Liu, C. Sun, K. Murphy, W. T. Freeman, J. B. Tenenbaum, and J. Wu. (2019). “Unsupervised discovery of parts, structure, and dynamics”. *arXiv preprint arXiv:1903.05136*.

Yamaguchi, T., S. Nagahama, Y. Ichikawa, and K. Takadama. (2019). “Model-Based Multi-objective Reinforcement Learning with Unknown Weights”. In: *International Conference on Human-Computer Interaction*. Springer. 311–321.

Yu, L., W. Zhang, J. Wang, and Y. Yu. (2017). “Seqgan: Sequence generative adversarial nets with policy gradient”. In: *Thirty-First AAAI Conference on Artificial Intelligence*.

Zanette, A. and E. Brunskill. (2019). "Tighter problem-dependent regret bounds in reinforcement learning without domain knowledge using value function bounds". In: *International Conference on Machine Learning*. PMLR. 7304–7312.

Zhang, L., G. Yang, and B. C. Stadie. (2021). "World model as a graph: Learning latent landmarks for planning". In: *International Conference on Machine Learning*. PMLR. 12611–12620.

Zhang, M., S. Vikram, L. Smith, P. Abbeel, M. Johnson, and S. Levine. (2019). "SOLAR: Deep Structured Representations for Model-Based Reinforcement Learning". In: *International Conference on Machine Learning*. 7444–7453.

Zhu, Z., K. Lin, and J. Zhou. (2020). "Transfer Learning in Deep Reinforcement Learning: A Survey". *arXiv preprint arXiv:2009.07888*.

Ziegler, Z. M. and A. M. Rush. (2019). "Latent normalizing flows for discrete sequences". *arXiv preprint arXiv:1901.10548*.