References

Adams, C. D. and Dickinson, A. (1981). Instrumental responding following reinforcer devaluation. *The Quarterly Journal of Experimental Psychology*, 33(2):109–121.

Agrawal, R. (1995). Sample mean based index policies with $O(logn)$ regret for the multi-armed bandit problem. *Advances in Applied Probability*, 27:1054–1078.

Agre, P. E. (1988). *The Dynamic Structure of Everyday Life*. Ph.D. thesis, Massachusetts Institute of Technology. AI-TR 1085, MIT Artificial Intelligence Laboratory.

Agre, P. E., Chapman, D. (1990). What are plans for? *Robotics and Autonomous Systems*, 6:17–34.

Albus, J. S. (1971). A theory of cerebellar function. *Mathematical Biosciences*, 10:25–61.

Albus, J. S. (1981). *Brain, Behavior, and Robotics.* Byte Books, Peterborough, NH.

Anderson, C. W. (1986). *Learning and Problem Solving with Multilayer Connectionist Systems.* Ph.D. thesis, University of Massachusetts, Amherst.

Anderson, C. W. (1987). Strategy learning with multilayer connectionist representations. *Proceedings of the Fourth International Workshop on Machine Learning*, pp. 103–114. Morgan Kaufmann, San Mateo, CA.

Anderson, J. A., Silverstein, J. W., Ritz, S. A., Jones, R. S. (1977). Distinctive features, categorical perception, and probability learning: Some applications of a neural model. *Psychological Review*, 84:413–451.

Andreae, J. H. (1963). STELLA: A scheme for a learning machine. In *Proceedings of the 2nd IFAC Congress, Basle*, pp. 497–502. Butterworths, London.

Andreae, J. H. (1969a). A learning machine with monologue. *International Journal of Man–Machine Studies*, 1:1–20.

Andreae, J. H. (1969b). Learning machines—a unified view. In A. R. Meetham and R. A. Hudson (eds.), *Encyclopedia of Information, Linguistics, and Control*, pp. 261–270. Pergamon, Oxford.

Andreae, J. H. (1977). *Thinking with the Teachable Machine.* Academic Press, London.

Arthur, W. B. (1991). Designing economic agents that act like human agents: A behavioral approach to bounded rationality. *The American Economic Review 81*(2):353-359.

Auer, P., Cesa-Bianchi, N., Fischer, P. (2002). Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47(2-3):235–256.

Baird, L. C. (1995). Residual algorithms: Reinforcement learning with function approximation. In *Proceedings of the Twelfth International Conference on Machine Learning*, pp. 30–37. Morgan Kaufmann, San Francisco.

Baldassarre, G. and Mirolli, M., editors (2013). *Intrinsically Motivated Learning in Natural and Artificial Systems.* Springer-Verlag, Berlin.

Balke, A., Pearl, J. (1994). Counterfactual probabilities: Computational methods, bounds and applications. In *Proceedings of the Tenth International Conference on Uncertainty in Artificial Intelligence* (pp. 46-54). Morgan Kaufmann.

Bao, G., Cassandras, C. G., Djaferis, T. E., Gandhi, A. D., Looze, D. P. (1994). Elevator dispatchers for down peak traffic. Technical report. ECE Department, University of Massachusetts, Amherst.

Barnard, E. (1993). Temporal-difference methods and Markov models. *IEEE Transactions on Systems, Man, and Cybernetics*, 23:357–365.

Barto, A. G. (1985). Learning by statistical cooperation of self-interested neuron-like computing elements. *Human Neurobiology*, 4:229–256.

Barto, A. G. (1986). Game-theoretic cooperativity in networks of self-interested units. In J. S. Denker (ed.), *Neural Networks for Computing*, pp. 41–46. American Institute of Physics, New York.

Barto, A. G. (1990). Connectionist learning for control: An overview. In T. Miller, R. S. Sutton, and P. J. Werbos (eds.), *Neural Networks for Control*, pp. 5–58. MIT Press, Cambridge, MA.

Barto, A. G. (1991). Some learning tasks from a control perspective. In L. Nadel and D. L. Stein (eds.), *1990 Lectures in Complex Systems*, pp. 195–223. Addison-Wesley, Redwood City, CA.

Barto, A. G. (1992). Reinforcement learning and adaptive critic methods. In D. A. White and D. A. Sofge (eds.), *Handbook of Intelligent Control: Neural, Fuzzy, and Adaptive Approaches*, pp. 469–491. Van Nostrand Reinhold, New York.

Barto, A. G. (1995a). Adaptive critics and the basal ganglia. In J. C. Houk, J. L. Davis, and D. G. Beiser (eds.), *Models of Information Processing in the Basal Ganglia*, pp. 215–232. MIT Press, Cambridge, MA.

Barto, A. G. (1995b). Reinforcement learning. In M. A. Arbib (ed.), *Handbook of Brain Theory and Neural Networks*, pp. 804–809. MIT Press, Cambridge, MA.

Barto, A. G. (2013). Intrinsic motivation and reinforcement learning. In Baldassarre, G. and Mirolli, M., editors, *Intrinsically Motivated Learning in Natural and Artificial Systems*, pages 17–47. Springer-Verlag, Berlin.

Barto, A. G., Anandan, P. (1985). Pattern recognizing stochastic learning automata. *IEEE Transactions on Systems, Man, and Cybernetics*, 15:360–375.

Barto, A. G., Anderson, C. W. (1985). Structural learning in connectionist systems. In *Program of the Seventh Annual Conference of the Cognitive Science Society*, pp. 43–54.

Barto, A. G., Anderson, C. W., Sutton, R. S. (1982). Synthesis of nonlinear control surfaces by a layered associative search network. *Biological Cybernetics*, 43:175–185.

Barto, A. G., Bradtke, S. J., Singh, S. P. (1991). Real-time learning and control using asynchronous dynamic programming. Technical Report 91-57. Department of Computer and Information Science, University of Massachusetts, Amherst.

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

Barto, A. G., Duff, M. (1994). Monte Carlo matrix inversion and reinforcement learning. In J. D. Cohen, G. Tesauro, and J. Alspector (eds.), *Advances in Neural Information Processing Systems: Proceedings of the 1993 Conference*, pp. 687–694. Morgan Kaufmann, San Francisco.

Barto, A. G., Jordan, M. I. (1987). Gradient following without back-propagation in layered networks. In M. Caudill and C. Butler (eds.), *Proceedings of the IEEE First Annual Conference on Neural Networks*, pp. II629–II636. SOS Printing, San Diego, CA.

Barto, A. G., Singh, S., and Chentanez, N. (2004). Intrinsically motivated learning of hierarchical collections of skills. In *International Conference on Developmental Learning (ICDL)*, LaJolla, CA.

Barto, A. G., Sutton, R. S. (1981a). Goal seeking components for adaptive intelligence: An initial assessment. Technical Report AFWAL-TR-81-1070. Air Force Wright Aeronautical Laboratories/Avionics Laboratory, Wright-Patterson AFB, OH.

Barto, A. G., Sutton, R. S. (1981b). Landmark learning: An illustration of associative search. *Biological Cybernetics*, 42:1–8.

Barto, A. G., Sutton, R. S. (1982). Simulation of anticipatory responses in classical conditioning by a neuron-like adaptive element. *Behavioural Brain Research*, 4:221–235.

Barto, A. G., Sutton, R. S., Anderson, C. W. (1983). Neuronlike elements that can solve difficult learning control problems. *IEEE Transactions on Systems, Man, and Cybernetics*, 13:835–846. Reprinted in J. A. Anderson and E. Rosenfeld (eds.), *Neurocomputing: Foundations of Research*, pp. 535–549. MIT Press, Cambridge, MA, 1988.

Barto, A. G., Sutton, R. S., Brouwer, P. S. (1981). Associative search network: A reinforcement learning associative memory. *Biological Cybernetics*, 40:201–211.

Bellman, R. E. (1956). A problem in the sequential design of experiments. *Sankhya*, 16:221–229.

Bellman, R. E. (1957a). *Dynamic Programming*. Princeton University Press, Princeton.

Bellman, R. E. (1957b). A Markov decision process. *Journal of Mathematical Mechanics*, 6:679–684.

Bellman, R. E., Dreyfus, S. E. (1959). Functional approximations and dynamic programming. *Mathematical Tables and Other Aids to Computation*, 13:247–251.

Bellman, R. E., Kalaba, R., Kotkin, B. (1973). Polynomial approximation—A new computational technique in dynamic programming: Allocation processes. *Mathematical Computation*, 17:155–161.

Bernoulli, D. (1954). Exposition of a new theory on the measurement of risk. *Econometrica*, 22(1):23–36. English translation of the 1738 paper.

Berridge, K. C. and Robinson, T. E. (1998). What is the role of dopamine in reward: hedonic impact, reward learning, or incentive salience? *Brain Research Reviews*, 28(3):309–369.

Berry, D. A., Fristedt, B. (1985). *Bandit Problems*. Chapman and Hall, London.

Bertsekas, D. P. (1982). Distributed dynamic programming. *IEEE Transactions on Automatic Control*, 27:610–616.

Bertsekas, D. P. (1983). Distributed asynchronous computation of fixed points. *Mathematical Programming*, 27:107–120.

Bertsekas, D. P. (1987). *Dynamic Programming: Deterministic and Stochastic Models*. Prentice-Hall, Englewood Cliffs, NJ.

Bertsekas, D. P. (2005). *Dynamic Programming and Optimal Control, Volume1,* third edition. Athena Scientific, Belmont, MA.

Bertsekas, D. P. (2012). *Dynamic Programming and Optimal Control, Volume 2: Approximate Dynamic Programming*, fourth edition. Athena Scientific, Belmont, MA.

Bertsekas, D. P., Tsitsiklis, J. N. (1989). *Parallel and Distributed Computation: Numerical Methods*. Prentice-Hall, Englewood Cliffs, NJ.

Bertsekas, D. P., Tsitsiklis, J. N. (1996). *Neuro-Dynamic Programming.* Athena Scientific, Belmont, MA.

Bertsekas, D. P., Yu, H. (2009). Projected equation methods for approximate solution of large linear systems. *Journal of Computational and Applied Mathematics*, 227(1):27–50.

Biermann, A. W., Fairfield, J. R. C., Beres, T. R. (1982). Signature table systems and learning. *IEEE Transactions on Systems, Man, and Cybernetics*, 12:635–648.

Bishop, C. M. (1995). *Neural Networks for Pattern Recognition.* Clarendon, Oxford.

Blodgett, H. C. (1929). The effect of the introduction of reward upon the maze performance of rats. *University of California Publications in Psychology*, 4:113–134.

Boakes, R. A. and Costa, D. S. J. (2014). Temporal contiguity in associative learning: Iinterference and decay from an historical perspective. *Journal of Experimental Psychology: Animal Learning and Cognition*, 40(4):381–400.

Booker, L. B. (1982). *Intelligent Behavior as an Adaptation to the Task Environment.* Ph.D. thesis, University of Michigan, Ann Arbor.

Boone, G. (1997). Minimum-time control of the acrobot. In *1997 International Conference on Robotics and Automation*, pp. 3281–3287. IEEE Robotics and Automation Society.

Boutilier, C., Dearden, R., Goldszmidt, M. (1995). Exploiting structure in policy construction. In *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence*, pp. 1104–1111. Morgan Kaufmann.

Boyan, J. A., Moore, A. W. (1995). Generalization in reinforcement learning: Safely approximating the value function. In G. Tesauro, D. S. Touretzky, and T. Leen (eds.), *Advances in Neural Information Processing Systems: Proceedings of the 1994 Conference*, pp. 369–376. MIT Press, Cambridge, MA.

Bradtke, S. J. (1993). Reinforcement learning applied to linear quadratic regulation. In S. J. Hanson, J. D. Cowan, and C. L. Giles (eds.), *Advances in Neural Information Processing Systems: Proceedings of the 1992 Conference*, pp. 295–302. Morgan Kaufmann, San Mateo, CA.

Bradtke, S. J. (1994). *Incremental Dynamic Programming for On-Line Adaptive Optimal Control.* Ph.D. thesis, University of Massachusetts, Amherst. Appeared as CMPSCI Technical Report 94-62.

Bradtke, S. J., Barto, A. G. (1996). Linear least–squares algorithms for temporal difference learning. *Machine Learning*, 22:33–57.

Bradtke, S. J., Ydstie, B. E., Barto, A. G. (1994). Adaptive linear quadratic control using policy iteration. In *Proceedings of the American Control Conference*, pp. 3475–3479. American Automatic Control Council, Evanston, IL.

Bradtke, S. J., Duff, M. O. (1995). Reinforcement learning methods for continuous-time Markov decision problems. In G. Tesauro, D. Touretzky, and T. Leen (eds.), *Advances in Neural Information Processing Systems: Proceedings of the 1994 Conference*, pp. 393–400. MIT Press, Cambridge, MA.

Brafman, R. I., Tennenholtz, M. (2003). R-max – a general polynomial time algorithm for near-optimal reinforcement learning. *Journal of Machine Learning Research*, 3, 213–231.

Breland, K. and Breland, M. (1961). The misbehavior of organisms. *American Psychologist*, 16(11):681–684.

Bridle, J. S. (1990). Training stochastic model recognition algorithms as networks can lead to maximum mutual information estimates of parameters. In D. S. Touretzky (ed.), *Advances in Neural Information Processing Systems: Proceedings of the 1989 Conference*, pp. 211–217. Morgan Kaufmann, San Mateo, CA.

Broomhead, D. S., Lowe, D. (1988). Multivariable functional interpolation and adaptive networks. *Complex Systems*, 2:321–355.

Bryson, A. E., Jr. (1996). Optimal control—1950 to 1985. *IEEE Control Systems*, 13(3):26–33.

Buchanan, B. G., Mitchell, T., Smith, R. G., and Jr., C. R. J. (1978). Models of learning systems. *Encyclopeadia of Computer Science and technology*, 11.

Burke, C. J., Dreher, J.-C., Seymour, B., and Tobler, P. N. (2014). State-dependent value representation: evidence from the stiatum. *Frontiers in Neuroscience*, 8.

Bush, R. R., Mosteller, F. (1955). *Stochastic Models for Learning.* Wiley, New York.

Byrne, J. H., Gingrich, K. J., Baxter, D. A. (1990). Computational capabilities of single neurons: Relationship to simple forms of associative and nonassociative learning in *aplysia.* In R. D. Hawkins and G. H. Bower (eds.), *Computational Models of Learning*, pp. 31–63. Academic Press, New York.

Camerer, C. (2003). *Behavioral game theory: Experiments in strategic interaction.* Princeton University Press.

Campbell, D. T. (1960). Blind variation and selective survival as a general strategy in knowledge-processes. In M. C. Yovits and S. Cameron (eds.), *Self-Organizing Systems*, pp. 205–231. Pergamon, New York.

Carlström, J., Nordström, E. (1997). Control of self-similar ATM call traffic by reinforcement learning. In *Proceedings of the International Workshop on Applications of Neural Networks to Telecommunications 3*, pp. 54–62. Erlbaum, Hillsdale, NJ.

Chapman, D., Kaelbling, L. P. (1991). Input generalization in delayed reinforcement learning: An algorithm and performance comparisons. In *Proceedings of the Twelfth International Conference on Artificial Intelligence*, pp. 726–731. Morgan Kaufmann, San Mateo, CA.

Chow, C.-S., Tsitsiklis, J. N. (1991). An optimal one-way multigrid algorithm for discrete-time stochastic control. *IEEE Transactions on Automatic Control*, 36:898–914.

Chrisman, L. (1992). Reinforcement learning with perceptual aliasing: The perceptual distinctions approach. In *Proceedings of the Tenth National Conference on Artificial Intelligence*, pp. 183–188. AAAI/MIT Press, Menlo Park, CA.

Christensen, J., Korf, R. E. (1986). A unified theory of heuristic evaluation functions and its application to learning. In *Proceedings of the Fifth National Conference on Artificial Intelligence*, pp. 148–152. Morgan Kaufmann, San Mateo, CA.

Cichosz, P. (1995). Truncating temporal differences: On the efficient implementation of TD(λ) for reinforcement learning. *Journal of Artificial Intelligence Research*, 2:287–318.

Clark, W. A., Farley, B. G. (1955). Generalization of pattern recognition in a self-organizing system. In *Proceedings of the 1955 Western Joint Computer Conference*, pp. 86–91.

Clouse, J. (1996). *On Integrating Apprentice Learning and Reinforcement Learning TITLE2.* Ph.D. thesis, University of Massachusetts, Amherst. Appeared as CMPSCI Technical Report 96-026.

Clouse, J., Utgoff, P. (1992). A teaching method for reinforcement learning systems. In *Proceedings of the Ninth International Machine Learning Conference*, pp. 92–101. Morgan Kaufmann, San Mateo, CA.

Colombetti, M., Dorigo, M. (1994). Training agent to perform sequential behavior. *Adaptive Behavior*, 2(3):247–275.

Connell, J. (1989). A colony architecture for an artificial creature. Technical Report AI-TR-1151. MIT Artificial Intelligence Laboratory, Cambridge, MA.

Connell, J., Mahadevan, S. (1993). *Robot Learning*. Kluwer Academic, Boston.

Courville, A. C., Daw, N. D., and Touretzky, D. S. (2006). Bayesian theories of conditioning in a changing world. *Trends in Cognitive Science*, 10(7):294–300.

Craik, K. J. W. (1943). *The Nature of Explanation*. Cambridge University Press, Cambridge.

Crites, R. H. (1996). *Large-Scale Dynamic Optimization Using Teams of Reinforcement Learning Agents*. Ph.D. thesis, University of Massachusetts, Amherst.

Crites, R. H., Barto, A. G. (1996). Improving elevator performance using reinforcement learning. In D. S. Touretzky, M. C. Mozer, and M. E. Hasselmo (eds.), *Advances in Neural Information Processing Systems: Proceedings of the 1995 Conference*, pp. 1017–1023. MIT Press, Cambridge, MA.

Cross, J. G. (1973). A stochastic learning model of economic behavior. *The Quarterly Journal of Economics 87*(2):239-266.

Curtiss, J. H. (1954). A theoretical comparison of the efficiencies of two classical methods and a Monte Carlo method for computing one component of the solution of a set of linear algebraic equations. In H. A. Meyer (ed.), *Symposium on Monte Carlo Methods*, pp. 191–233. Wiley, New York.

Cziko, G. (1995). *Without Miracles: Universal Selection Theory and the Second Darvinian Revolution*. MIT Press, Cambridge, MA.

Daniel, J. W. (1976). Splines and efficiency in dynamic programming. *Journal of Mathematical Analysis and Applications*, 54:402–407.

Daw, N., Niv, Y., and Dayan, P. (2005). Uncertainty based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nature Neuroscience*, 8(12):1704–1711.

Daw, N. D. and Shohamy, D. (2008). The cognitive neuroscience of motivation and learning. *Social Cognition*, 26(5):593–620.

Dayan, P. (1991). Reinforcement comparison. In D. S. Touretzky, J. L. Elman, T. J. Sejnowski, and G. E. Hinton (eds.), *Connectionist Models: Proceedings of the 1990 Summer School*, pp. 45–51. Morgan Kaufmann, San Mateo, CA.

Dayan, P. (1992). The convergence of TD(λ) for general λ. *Machine Learning*, 8:341–362.

Dayan, P. (2008). The role of value systems in decision making. In Engel, C. and Singer, W., editors, *Better Than Conscious?: Decision Making, the Human Mind, and Implications For Institutions (Strüngmann Forum Reports)*, pages 51–70. MIT Press, Cambridge, MA.

Dayan, P. and Berridge, K. C. (2014). Model-based and model-free Pavlovian reward learning: Revaluation, revision, and revaluation. *Cognitive, Affective, & Behavioral Neuroscience*, 14(2):473–492.

Dayan, P., Hinton, G. E. (1993). Feudal reinforcement learning. In S. J. Hanson, J. D. Cohen, and C. L. Giles (eds.), *Advances in Neural Information Processing Systems: Proceedings of the 1992 Conference*, pp. 271–278. Morgan Kaufmann, San Mateo, CA.

Dayan, P., Sejnowski, T. (1994). TD(λ) converges with probability 1. *Machine Learning*, 14:295–301.

Dean, T., Lin, S.-H. (1995). Decomposition techniques for planning in stochastic domains. In *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence*, pp. 1121–1127. Morgan Kaufmann. See also Technical Report CS-95-10, Brown

University, Department of Computer Science, 1995.

DeJong, G., Spong, M. W. (1994). Swinging up the acrobot: An example of intelligent control. In *Proceedings of the American Control Conference*, pp. 2158–2162. American Automatic Control Council, Evanston, IL.

Denardo, E. V. (1967). Contraction mappings in the theory underlying dynamic programming. *SIAM Review*, 9:165–177.

Dennett, D. C. (1978). *Brainstorms*, pp. 71–89. Bradford/MIT Press, Cambridge, MA.

Dick, T. (2015). *A Regret-full Perspective on Policy Gradient Methods for Reinforcement Learning.* MSc Thesis, University of Alberta.

Dickinson, A. (1980). *Contemporary Animal Learning Theory.* Cambridge University Press, Cambridge.

Dickinson, A. (1985). Actions and habits: the development of behavioral autonomy. *Phil. Trans. R. Soc. Lond. B*, 308(1135):67–78.

Dickinson, A. and Balleine, B. W. (2002). The role of learning in motivation. In Gallistel, C. R., editor, *Stevens handbook of experimental psychology*, volume 3, pages 497–533. Wiley, NY.

Dietterich, T. and Buchanan, B. G. (1984). The role of the critic in learning systems. In Selfridge, O. G., Rissland, E. L., and Arbib, M. A., editors, *Adaptive Control of Ill-Defined Systems*, pages 127–147. Plenum Press, NY. Proceedings of the NATO Advanced Research Institute on Adaptive Control of Ill-defined Systems, NATO Conference Series II, Systems Science, Vol. 16.

Dietterich, T. G., Flann, N. S. (1995). Explanation-based learning and reinforcement learning: A unified view. In A. Prieditis and S. Russell (eds.), *Proceedings of the Twelfth International Conference on Machine Learning*, pp. 176–184. Morgan Kaufmann, San Francisco.

Dolan, R. J. and Dayan, P. (2013). Goals and habits in the brain. *Neuron*, 80(2):312–325.

Donahoe, J. W. and Burgos, J. E. (2000). Behavior analysis and revaluation. *Journal of the Experimental Analysis of Behavior*, 74(3):331–346.

Dorigo, M. and Colombetti, M. (1994). Robot shaping: Developing autonomous agents through learning. *Artificial Intelligence*, 71(2):321–370.

Doya, K. (1996). Temporal difference learning in continuous time and space. In D. S. Touretzky, M. C. Mozer, and M. E. Hasselmo (eds.), *Advances in Neural Information Processing Systems: Proceedings of the 1995 Conference*, pp. 1073–1079. MIT Press, Cambridge, MA.

Doyle, P. G., Snell, J. L. (1984). *Random Walks and Electric Networks.* The Mathematical Association of America. Carus Mathematical Monograph 22.

Dreyfus, S. E., Law, A. M. (1977). *The Art and Theory of Dynamic Programming.* Academic Press, New York.

Duda, R. O., Hart, P. E. (1973). *Pattern Classification and Scene Analysis.* Wiley, New York.

Duff, M. O. (1995). Q-learning for bandit problems. In A. Prieditis and S. Russell (eds.), *Proceedings of the Twelfth International Conference on Machine Learning*, pp. 209–217. Morgan Kaufmann, San Francisco.

Estes, W. K. (1950). Toward a statistical theory of learning. *Psychololgical Review*, 57:94–107.

Farley, B. G., Clark, W. A. (1954). Simulation of self-organizing systems by digital computer.

IRE Transactions on Information Theory, 4:76–84.

Feldbaum, A. A. (1965). *Optimal Control Systems.* Academic Press, New York.

Finnsson, H., Björnsson, Y. (2008). Simulation-based approach to general game playing. In *Proceedings of the Association for the Advancement of Artificial Intelligence*, 259–264.

Fogel, L. J., Owens, A. J., Walsh, M. J. (1966). *Artificial intelligence through simulated evolution.* John Wiley and Sons.

Friston, K. J., Tononi, G., Reeke, G. N., Sporns, O., Edelman, G. M. (1994). Value-dependent selection in the brain: Simulation in a synthetic neural model. *Neuroscience*, 59:229–243.

Fu, K. S. (1970). Learning control systems—Review and outlook. *IEEE Transactions on Automatic Control*, 15:210–221.

Galanter, E., Gerstenhaber, M. (1956). On thought: The extrinsic theory. *Psychological Review*, 63:218–227.

Gallant, S. I. (1993). *Neural Network Learning and Expert Systems.* MIT Press, Cambridge, MA.

Gallistel, C. R. (2005). Deconstructing the law of effect. *Games and Economic Behavior 52*(2), 410-423.

Gällmo, O., Asplund, L. (1995). Reinforcement learning by construction of hypothetical targets. In J. Alspector, R. Goodman, and T. X. Brown (eds.), *Proceedings of the International Workshop on Applications of Neural Networks to Telecommunications 2*, pp. 300–307. Erlbaum, Hillsdale, NJ.

Gardner, M. (1973). Mathematical games. *Scientific American*, 228(1):108–115.

Gelperin, A., Hopfield, J. J., Tank, D. W. (1985). The logic of *limax* learning. In A. Selverston (ed.), *Model Neural Networks and Behavior*, pp. 247–261. Plenum Press, New York.

Genesereth, M., Thielscher, M. (2014). General game playing. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, 8(2), 1–229.

Gershman, S. J. and Niv, Y. (2010). Learning latent structure: Carving nature at its joints. *Current Opinions in Neurobiology*, 20:251–256.

Gittins, J. C., Jones, D. M. (1974). A dynamic allocation index for the sequential design of experiments. In J. Gani, K. Sarkadi, and I. Vincze (eds.), *Progress in Statistics*, pp. 241–266. North-Holland, Amsterdam–London.

Goldberg, D. E. (1989). *Genetic Algorithms in Search, Optimization, and Machine Learning.* Addison-Wesley, Reading, MA.

Goldstein, H. (1957). *Classical Mechanics.* Addison-Wesley, Reading, MA.

Goodwin, G. C., Sin, K. S. (1984). *Adaptive Filtering Prediction and Control.* Prentice-Hall, Englewood Cliffs, NJ.

Gopnik, A., Glymour, C., Sobel, D., Schulz, L. E., Kushnir, T., and Danks, D. (2004). A theory of causal learning in children: Causal maps and Bayes nets. *Psychological Review*, 111(1):3–32.

Gordon, G. J. (1995). Stable function approximation in dynamic programming. In A. Prieditis and S. Russell (eds.), *Proceedings of the Twelfth International Conference on Machine Learning*, pp. 261–268. Morgan Kaufmann, San Francisco. An expanded version was published as Technical Report CMU-CS-95-103. Carnegie Mellon University, Pittsburgh, PA, 1995.

Gordon, G. J. (1996). Chattering in SARSA(λ). CMU learning lab internal report.

Gordon, G. J. (1996). Stable fitted reinforcement learning. In D. S. Touretzky, M. C. Mozer, M. E. Hasselmo (eds.), *Advances in Neural Information Processing Systems: Proceedings of the 1995 Conference*, pp. 1052–1058. MIT Press, Cambridge, MA.

Gordon, G. J. (2001). Reinforcement learning with function approximation converges to a region. *Advances in neural information processing systems.*

Greensmith, E., Bartlett, P. L., Baxter, J. (2001). Variance reduction techniques for gradient estimates in reinforcement learning. In *Advances in Neural Information Processing Systems: Proceedings of the 2000 Conference*, pp. 1507–1514.

Greensmith, E., Bartlett, P. L., Baxter, J. (2004). Variance reduction techniques for gradient estimates in reinforcement learning. *Journal of Machine Learning Research 5*, 1471-1530.

Griffith, A. K. (1966). A new machine learning technique applied to the game of checkers. Technical Report Project MAC, Artificial Intelligence Memo 94. Massachusetts Institute of Technology, Cambridge, MA.

Griffith, A. K. (1974). A comparison and evaluation of three machine learning procedures as applied to the game of checkers. *Artificial Intelligence*, 5:137–148.

Grossberg, S. (1975). A neural model of attention, reinforcement, and discrimination learning. *International Review of Neurobiology*, 18:263–327.

Gullapalli, V. (1990). A stochastic reinforcement algorithm for learning real-valued functions. *Neural Networks*, 3:671–692.

Gurvits, L., Lin, L.-J., Hanson, S. J. (1994). Incremental learning of evaluation functions for absorbing Markov chains: New methods and theorems. Preprint.

Hampson, S. E. (1983). *A Neural Model of Adaptive Behavior.* Ph.D. thesis, University of California, Irvine.

Hampson, S. E. (1989). *Connectionist Problem Solving: Computational Aspects of Biological Learning.* Birkhauser, Boston.

Hawkins, R. D., Kandel, E. R. (1984). Is there a cell-biological alphabet for simple forms of learning? *Psychological Review*, 91:375–391.

Herrnstein, R. J. (1970). On the Law of Effect. *Journal of the Experimental Analysis of Behavior 13*(2), 243-266.

Hersh, R., Griego, R. J. (1969). Brownian motion and potential theory. *Scientific American*, 220:66–74.

Hesterberg, T. C. (1988), *Advances in importance sampling*, Ph.D. Dissertation, Statistics Department, Stanford University.

Hilgard, E. R. (1956). *Theories of Learning, Second Edition.* Appleton-Century-Cofts, Inc., New York.

Hilgard, E. R., Bower, G. H. (1975). *Theories of Learning.* Prentice-Hall, Englewood Cliffs, NJ.

Hinton, G. E. (1984). Distributed representations. Technical Report CMU-CS-84-157. Department of Computer Science, Carnegie-Mellon University, Pittsburgh, PA.

Hochreiter, S., Schmidhuber, J. (1997). LTSM can solve hard time lag problems. In *Advances in Neural Information Processing Systems: Proceedings of the 1996 Conference*, pp. 473–479. MIT Press, Cambridge, MA.

Holland, J. H. (1975). *Adaptation in Natural and Artificial Systems.* University of Michigan Press, Ann Arbor.

Holland, J. H. (1976). Adaptation. In R. Rosen and F. M. Snell (eds.), *Progress in Theoretical Biology*, vol. 4, pp. 263–293. Academic Press, New York.

Holland, J. H. (1986). Escaping brittleness: The possibility of general-purpose learning algorithms applied to rule-based systems. In R. S. Michalski, J. G. Carbonell, and T. M. Mitchell (eds.), *Machine Learning: An Artificial Intelligence Approach*, vol. 2, pp. 593–623. Morgan Kaufmann, San Mateo, CA.

Houk, J. C., Adams, J. L., Barto, A. G. (1995). A model of how the basal ganglia generates and uses neural signals that predict reinforcement. In J. C. Houk, J. L. Davis, and D. G. Beiser (eds.), *Models of Information Processing in the Basal Ganglia*, pp. 249–270. MIT Press, Cambridge, MA.

Howard, R. (1960). *Dynamic Programming and Markov Processes.* MIT Press, Cambridge, MA.

Hull, C. L. (1932). The goal-gradient hypothesis and maze learning. *Psychological Review*, 39(1):25–43.

Hull, C. L. (1943). *Principles of Behavior.* Appleton-Century, New York.

Hull, C. L. (1952). *A Behavior System.* Wiley, New York.

Jaakkola, T., Jordan, M. I., Singh, S. P. (1994). On the convergence of stochastic iterative dynamic programming algorithms. *Neural Computation*, 6:1185–1201.

Jaakkola, T., Singh, S. P., Jordan, M. I. (1995). Reinforcement learning algorithm for partially observable Markov decision problems. In G. Tesauro, D. S. Touretzky, T. Leen (eds.), *Advances in Neural Information Processing Systems: Proceedings of the 1994 Conference*, pp. 345–352. MIT Press, Cambridge, MA.

Johanson, E. B., Killeen, P. R., Russell, V. A., Tripp, G., Wickens, J. R., Tannock, R., Williams, J., and Sagvolden, T. (2009). Origins of altered reinforcement effects in ADHD. *Behavioral and Brain Functions*, 5(7).

Kaelbling, L. P. (1993a). Hierarchical learning in stochastic domains: Preliminary results. In *Proceedings of the Tenth International Conference on Machine Learning*, pp. 167–173. Morgan Kaufmann, San Mateo, CA.

Kaelbling, L. P. (1993b). *Learning in Embedded Systems.* MIT Press, Cambridge, MA.

Kaelbling, L. P. (ed.). (1996). Special issue of *Machine Learning* on reinforcement learning, 22.

Kaelbling, L. P., Littman, M. L., Moore, A. W. (1996). Reinforcement learning: A survey. *Journal of Artificial Intelligence Research*, 4:237–285.

Kahneman, D. and Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica: Journal of the Econometric Society*, 47:263–291.

Kakutani, S. (1945). Markov processes and the Dirichlet problem. *Proceedings of the Japan Academy*, 21:227–233.

Kalos, M. H., Whitlock, P. A. (1986). *Monte Carlo Methods.* Wiley, New York.

Kamin, L. J. (1968). "Attention-like" processes in classical conditioning. In Jones, M. R., editor, *Miami Symposium on the Prediction of Behavior, 1967: Aversive Stimulation*, pages 9–31. University of Miami Press, Coral Gables, Florida.

Kamin, L. J. (1969). Predictability, surprise, attention, and conditioning. In Campbell, B. A. and Church, R. M., editors, *Punishment and Aversive Behavior*, pages 279–296. Appleton-Century-Crofts, New York, NY.

Kanerva, P. (1988). *Sparse Distributed Memory.* MIT Press, Cambridge, MA.

Kanerva, P. (1993). Sparse distributed memory and related models. In M. H. Hassoun (ed.), *Associative Neural Memories: Theory and Implementation*, pp. 50–76. Oxford University Press, New York.

Kashyap, R. L., Blaydon, C. C., Fu, K. S. (1970). Stochastic approximation. In J. M. Mendel and K. S. Fu (eds.), *Adaptive, Learning, and Pattern Recognition Systems: Theory and Applications*, pp. 329–355. Academic Press, New York.

Kearns, M., Singh, S. (2002). Near-optimal reinforcement learning in polynomial time. *Machine Learning*, 49(2-3), 209–232.

Keerthi, S. S., Ravindran, B. (1997). Reinforcement learning. In E. Fiesler and R. Beale (eds.), *Handbook of Neural Computation*, C3. Oxford University Press, New York.

Kehoe, E. J., Schreurs, B. G., and Graham, P. (1987). Temporal primacy overrides prior training in serial compound conditioning of the rabbits nictitating membrane response. *Animal Learning & Behavior*, 15(4):455–464.

Kimble, G. A. (1961). *Hilgard and Marquis' Conditioning and Learning*. Appleton-Century-Crofts, New York.

Kimble, G. A. (1967). *Foundations of Conditioning and Learning*. Appleton-Century-Crofts, New York.

Klopf, A. H. (1972). Brain function and adaptive systems—A heterostatic theory. Technical Report AFCRL-72-0164, Air Force Cambridge Research Laboratories, Bedford, MA. A summary appears in *Proceedings of the International Conference on Systems, Man, and Cybernetics*. IEEE Systems, Man, and Cybernetics Society, Dallas, TX, 1974.

Klopf, A. H. (1975). A comparison of natural and artificial intelligence. *SIGART Newsletter*, 53:11–13.

Klopf, A. H. (1982). *The Hedonistic Neuron: A Theory of Memory, Learning, and Intelligence*. Hemisphere, Washington, DC.

Klopf, A. H. (1988). A neuronal model of classical conditioning. *Psychobiology*, 16:85–125.

Kocsis, L., Szepesvári, Cs. (2006). Bandit based Monte-Carlo planning. In *Proceedings of the European Conference on Machine Learning*, 282–293. Springer Berlin Heidelberg.

Kohonen, T. (1977). *Associative Memory: A System Theoretic Approach*. Springer-Verlag, Berlin.

Koller, D., Friedman, N. (2009). *Probabilistic Graphical Models: Principles and Techniques*. MIT Press, 2009.

Korf, R. E. (1988). Optimal path finding algorithms. In L. N. Kanal and V. Kumar (eds.), *Search in Artificial Intelligence*, pp. 223–267. Springer Verlag, Berlin.

Koza, J. R. (1992). *Genetic programming: On the programming of computers by means of natural selection* (Vol. 1). MIT press.

Kraft, L. G., Campagna, D. P. (1990). A summary comparison of CMAC neural network and traditional adaptive control systems. In T. Miller, R. S. Sutton, and P. J. Werbos (eds.), *Neural Networks for Control*, pp. 143–169. MIT Press, Cambridge, MA.

Kraft, L. G., Miller, W. T., Dietz, D. (1992). Development and application of CMAC neural network-based control. In D. A. White and D. A. Sofge (eds.), *Handbook of Intelligent Control: Neural, Fuzzy, and Adaptive Approaches*, pp. 215–232. Van Nostrand Reinhold, New York.

Kumar, P. R., Varaiya, P. (1986). *Stochastic Systems: Estimation, Identification, and Adaptive Control*. Prentice-Hall, Englewood Cliffs, NJ.

Kumar, P. R. (1985). A survey of some results in stochastic adaptive control. *SIAM Journal of Control and Optimization*, 23:329–380.

Kumar, V., Kanal, L. N. (1988). The CDP: A unifying formulation for heuristic search, dynamic programming, and branch-and-bound. In L. N. Kanal and V. Kumar (eds.), *Search in Artificial Intelligence*, pp. 1–37. Springer-Verlag, Berlin.

Kushner, H. J., Dupuis, P. (1992). *Numerical Methods for Stochastic Control Problems in Continuous Time*. Springer-Verlag, New York.

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

Lang, K. J., Waibel, A. H., Hinton, G. E. (1990). A time-delay neural network architecture for isolated word recognition. *Neural Networks*, 3:33–43.

Lewis, F. L., Liu, D. (Eds.). (2013). *Reinforcement Learning and Approximate Dynamic Programming for Feedback Control*. John Wiley and Sons.

Lewis, R. L., Howes, A., and Singh, S. (2014). Computational rationality: Linking mechanism and behavior through utility maximization. *Topics in Cognitive Science*, 6(2):279–311.

Lin, C.-S., Kim, H. (1991). CMAC-based adaptive critic self-learning control. *IEEE Transactions on Neural Networks*, 2:530–533.

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

Lin, L.-J., Mitchell, T. (1992). Reinforcement learning with hidden states. In *Proceedings of the Second International Conference on Simulation of Adaptive Behavior: From Animals to Animats*, pp. 271–280. MIT Press, Cambridge, MA.

Littman, M. L. (1994). Markov games as a framework for multi-agent reinforcement learning. In *Proceedings of the Eleventh International Conference on Machine Learning*, pp. 157–163. Morgan Kaufmann, San Francisco.

Littman, M. L., Cassandra, A. R., Kaelbling, L. P. (1995). Learning policies for partially observable environments: Scaling up. In A. Prieditis and S. Russell (eds.), *Proceedings of the Twelfth International Conference on Machine Learning*, pp. 362–370. Morgan Kaufmann, San Francisco.

Littman, M. L., Dean, T. L., Kaelbling, L. P. (1995). On the complexity of solving Markov decision problems. In *Proceedings of the Eleventh Annual Conference on Uncertainty in Artificial Intelligence*, pp. 394–402.

Liu, J. S. (2001). *Monte Carlo strategies in scientific computing*. Berlin, Springer-Verlag.

Ljung, L. (1998). System identification. In Procházka, A., Uhl´*i*r̂, J., Rayner, P. W. J., and Kingsbury, N. G., editors, *Signal Analysis and Prediction*, pages 163–173. Springer Science + Business Media New York, LLC.

Ljung, L., Söderstrom, T. (1983). *Theory and Practice of Recursive Identification*. MIT Press, Cambridge, MA.

Lovejoy, W. S. (1991). A survey of algorithmic methods for partially observed Markov decision processes. *Annals of Operations Research*, 28:47–66.

Luce, D. (1959). *Individual Choice Behavior*. Wiley, New York.

Ludvig, E. A., Sutton, R. S., and Kehoe, E. J. (2008). Stimulus representation and the timing of reward-prediction errors in models of the dopamine system. *Neural Computation*, 20(12):3034–3054.

Ludvig, E. A., Sutton, R. S., and Kehoe, E. J. (2012). Evaluating the TD model of classical conditioning. *Learning & behavior*, 40(3):305–319.

Mackintosh, N. J. (1975). A theory of attention: Variations in the associability of stimuli with reinforcement. *Psychological Review*, 82(4):276–298.

Maclin, R., Shavlik, J. W. (1994). Incorporating advice into agents that learn from reinforcements. In *Proceedings of the Twelfth National Conference on Artificial Intelligence*, pp. 694–699. AAAI Press, Menlo Park, CA.

Mahadevan, S. (1996). Average reward reinforcement learning: Foundations, algorithms, and empirical results. *Machine Learning*, 22:159–196.

Mahadevan, S. and Connell, J. (1992). Automatic programming of behavior-based robots using reinforcement learning. *Artificial Intelligence*, 55:311–365.

Markey, K. L. (1994). Efficient learning of multiple degree-of-freedom control problems with quasi-independent Q-agents. In M. C. Mozer, P. Smolensky, D. S. Touretzky, J. L. Elman, and A. S. Weigend (eds.), *Proceedings of the 1990 Connectionist Models Summer School*. Erlbaum, Hillsdale, NJ.

Mataric, M. J. (1994). Reward functions for accelerated learning. In *Machine Learning: Proceedings of the Eleventh international conference*, pages 181–189.

Mazur, J. E. (1994). *Learning and Behavior*, 3rd ed. Prentice-Hall, Englewood Cliffs, NJ.

McCallum, A. K. (1993). Overcoming incomplete perception with utile distinction memory. In *Proceedings of the Tenth International Conference on Machine Learning*, pp. 190–196. Morgan Kaufmann, San Mateo, CA.

McCallum, A. K. (1995). *Reinforcement Learning with Selective Perception and Hidden State*. Ph.D. thesis, University of Rochester, Rochester, NY.

Mendel, J. M. (1966). A survey of learning control systems. *ISA Transactions*, 5:297–303.

Mendel, J. M., McLaren, R. W. (1970). Reinforcement learning control and pattern recognition systems. In J. M. Mendel and K. S. Fu (eds.), *Adaptive, Learning and Pattern Recognition Systems: Theory and Applications*, pp. 287–318. Academic Press, New York.

Michie, D. (1961). Trial and error. In S. A. Barnett and A. McLaren (eds.), *Science Survey, Part 2*, pp. 129–145. Penguin, Harmondsworth.

Michie, D. (1963). Experiments on the mechanisation of game learning. 1. characterization of the model and its parameters. *Computer Journal*, 1:232–263.

Michie, D. (1974). *On Machine Intelligence*. Edinburgh University Press, Edinburgh.

Michie, D., Chambers, R. A. (1968). BOXES: An experiment in adaptive control. In E. Dale and D. Michie (eds.), *Machine Intelligence 2*, pp. 137–152. Oliver and Boyd, Edinburgh.

Miller, S., Williams, R. J. (1992). Learning to control a bioreactor using a neural net Dyna-Q system. In *Proceedings of the Seventh Yale Workshop on Adaptive and Learning Systems*, pp. 167–172. Center for Systems Science, Dunham Laboratory, Yale University, New Haven.

Miller, W. T., Scalera, S. M., Kim, A. (1994). Neural network control of dynamic balance for a biped walking robot. In *Proceedings of the Eighth Yale Workshop on Adaptive and Learning Systems*, pp. 156–161. Center for Systems Science, Dunham Laboratory, Yale University, New Haven.

Minsky, M. L. (1954). *Theory of Neural-Analog Reinforcement Systems and Its Application to the Brain-Model Problem*. Ph.D. thesis, Princeton University.

Minsky, M. L. (1961). Steps toward artificial intelligence. *Proceedings of the Institute of Radio Engineers*, 49:8–30. Reprinted in E. A. Feigenbaum and J. Feldman (eds.), *Computers and Thought*, pp. 406–450. McGraw-Hill, New York, 1963.

Minsky, M. L. (1967). *Computation: Finite and Infinite Machines*. Prentice-Hall, Englewood Cliffs, NJ.

Montague, P. R., Dayan, P., Sejnowski, T. J. (1996). A framework for mesencephalic dopamine systems based on predictive Hebbian learning. *Journal of Neuroscience*,

16:1936–1947.

Moore, A. W. (1990). *Efficient Memory-Based Learning for Robot Control*. Ph.D. thesis, University of Cambridge.

Moore, A. W. (1994). The parti-game algorithm for variable resolution reinforcement learning in multidimensional spaces. In J. D. Cohen, G. Tesauro and J. Alspector (eds.), *Advances in Neural Information Processing Systems: Proceedings of the 1993 Conference*, pp. 711–718. Morgan Kaufmann, San Francisco.

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

Moore, J. W. and Blazis, D. E. J. (1989). Simulation of a classically conditioned response: A cerebellar implementation of the sutton-barto-desmond model. In Byrne, J. H. and Berry, W. O., editors, *Neural Models of Plasticity*, pages 187–207. Academic Press, San Diego, CA.

Moore, J. W., Choi, J.-S., and Brunzell, D. H. (1998). Predictive timing under temporal uncertainty: The time derivative model of the conditioned response. In Rosenbaum, D. A. and Collyer, C. E., editors, *Timing of Behavior*, pages 3–34. MIT Press, Cambridge, MA.

Moore, J. W., Desmond, J. E., Berthier, N. E., Blazis, E. J., Sutton, R. S., and Barto, A. G. (1986). Simulation of the classically conditioned nictitating membrane response by a neuron-like adaptive element: I. Response topography, neuronal firing, and interstimulus intervals. *Behavioural Brain Research*, 21:143–154.

Moore, J. W., Marks, J. S., Castagna, V. E., and Polewan, R. J. (2001). Parameter stability in the TD model of complex CR topographies. Society for Neuroscience Abstract 642.2.

Moore, J. W. and Schmajuk, N. A. (2008). Kamin blocking. *Scholarpedia*, 3(5):3542.

Moore, J. W. and Stickney, K. J. (1980). Formation of attentional-associative networks in real time:Role of the hippocampus and implications for conditioning. *Physiological Psychology*, 8(2):207–217.

Narendra, K. S., Thathachar, M. A. L. (1974). Learning automata—A survey. *IEEE Transactions on Systems, Man, and Cybernetics*, 4:323–334.

Narendra, K. S., Thathachar, M. A. L. (1989). *Learning Automata: An Introduction.* Prentice-Hall, Englewood Cliffs, NJ.

Narendra, K. S., Wheeler, R. M. (1986). Decentralized learning in finite Markov chains. *IEEE Transactions on Automatic Control*, AC31(6):519–526.

Ng, A. Y. (2003). *Shaping and policy search in reinforcement learning.* PhD thesis, University of California, Berkeley, Berkeley, CA.

Ng, A. Y., Harada, D., and Russell, S. (1999). Policy invariance under reward transformations: Theory and application to reward shaping. In Bratko, I. and Dzeroski, S., editors, *Proceedings of the Sixteenth International Conference on Machine Learning (ICML 1999)*, volume 99, pages 278–287.

Nie, J., Haykin, S. (1996). A dynamic channel assignment policy through Q-learning. CRL Report 334. Communications Research Laboratory, McMaster University, Hamilton, Ontario.

Niv, Y., Daw, N. D., and Dayan, P. (2005). How fast to work: Response vigor, motivation and tonic dopamine. In Yeiss, Y., Schölkopft, B., and Platt, J., editors, *Advances in Neural Information Processing Systems 18 (NIPS 2005)*, pages 1019–1026. MIT Press, Cambridge, MA.

Niv, Y., Daw, N. D., Joel, D., and Dayan, P. (2007). Tonic dopamine: opportunity costs and the control of response vigor. *Psychopharmacology*, 191(3):507–520.

Niv, Y., Joel, D., and Dayan, P. (2006). A normative perspective on motivation. *Trends in Cognitive Sciences*, 10(8):375–381.

Nowé, A., Vrancx, P., De Hauwere, Y. M. (2012). Game theory and multi-agent reinforcement learning. In *Reinforcement Learning* (pp. 441-470). Springer Berlin Heidelberg.

Oudeyer, P.-Y. and Kaplan, F. (2007). What is intrinsic motivation? A typology of computational approaches. *Frontiers in Neurorobotics*, 1.

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

Page, C. V. (1977). Heuristics for signature table analysis as a pattern recognition technique. *IEEE Transactions on Systems, Man, and Cybernetics*, 7:77–86.

Parr, R., Russell, S. (1995). Approximating optimal policies for partially observable stochastic domains. In *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence*, pp. 1088–1094. Morgan Kaufmann.

Pavlov, P. I. (1927). *Conditioned Reflexes*. Oxford University Press, London.

Pearce, J. M. and Hall, G. (1980). A model for Pavlovian learning: Variation in the effectiveness of conditioning but not unconditioned stimuli. *Psychological Review*, 87(6):532–552.

Pearl, J. (1984). *Heuristics: Intelligent Search Strategies for Computer Problem Solving.* Addison-Wesley, Reading, MA.

Pearl, J. (1995). Causal diagrams for empirical research. *Biometrika*, 82(4), 669-688.

Peng, J. (1993). *Efficient Dynamic Programming-Based Learning for Control.* Ph.D. thesis, Northeastern University, Boston.

Peng, J., Williams, R. J. (1993). Efficient learning and planning within the Dyna framework. *Adaptive Behavior*, 1(4):437–454.

Peng, J., Williams, R. J. (1994). Incremental multi-step Q-learning. In W. W. Cohen and H. Hirsh (eds.), *Proceedings of the Eleventh International Conference on Machine Learning*, pp. 226–232. Morgan Kaufmann, San Francisco.

Peng, J., Williams, R. J. (1996). Incremental multi-step Q-learning. *Machine Learning*, 22:283–290.

Peterson, G. B. (2004). A day of great illumination: B.F. Skinner's discovery of shaping. *Journal of the Experimental Analysis of Behavior*, 82(3):317–28.

Phansalkar, V. V., Thathachar, M. A. L. (1995). Local and global optimization algorithms for generalized learning automata. *Neural Computation*, 7:950–973.

Poggio, T., Girosi, F. (1989). A theory of networks for approximation and learning. A.I. Memo 1140. Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA.

Poggio, T., Girosi, F. (1990). Regularization algorithms for learning that are equivalent to multilayer networks. *Science*, 247:978–982.

Powell, M. J. D. (1987). Radial basis functions for multivariate interpolation: A review. In J. C. Mason and M. G. Cox (eds.), *Algorithms for Approximation*, pp. 143–167. Clarendon Press, Oxford.

Powell, W. B. (2011). *Approximate Dynamic Programming: Solving the Curses of Dimensionality*, Second edition. John Wiley and Sons.

Precup, D., Sutton, R. S., Dasgupta, S. (2001). Off-policy temporal-difference learning with function approximation. In *Proceedings of the 18th International Conference on Machine Learning*.

Precup, D., Sutton, R. S., Singh, S. (2000). Eligibility traces for off-policy policy evaluation. In *Proceedings of the 17th International Conference on Machine Learning*, pp. 759–766. Morgan Kaufmann.

Puterman, M. L. (1994). *Markov Decision Problems.* Wiley, New York.

Puterman, M. L., Shin, M. C. (1978). Modified policy iteration algorithms for discounted Markov decision problems. *Management Science*, 24:1127–1137.

Randløv, J. and Alstrøm, P. (1998). Learning to drive a bicycle using reinforcement learning and shaping. In *Proceedings of the Fifteenth International Conference on Machine Learning*, pages 463–471.

Rangel, A., Camerer, C., and Montague, P. R. (2008). A framework for studying the neurobiology of value-based decision making. *Nature Reviews Neuroscience*, 9(7):545–556.

Rescorla, R. A. and Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In Black, A. H. and Prokasy, W. F., editors, *Classical Conditioning II*, pages 64–99. Appleton-Century-Crofts, New York.

Reetz, D. (1977). Approximate solutions of a discounted Markovian decision process. *Bonner Mathematische Schriften*, 98:77–92.

Revusky, S. and Garcia, J. (1970). Learned associations over long delays. In Bower, G., editor, *The psychology of learning and motivation*, volume 4, pages 1–84. Academic Press, Inc., New York.

Ring, M. B. (1994). *Continual Learning in Reinforcement Environments.* Ph.D. thesis, University of Texas, Austin.

Rivest, R. L., Schapire, R. E. (1987). Diversity-based inference of finite automata. In *Proceedings of the Twenty-Eighth Annual Symposium on Foundations of Computer Science*, pp. 78–87. Computer Society Press of the IEEE, Washington, DC.

Robbins, H. (1952). Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society*, 58:527–535.

Robertie, B. (1992). Carbon versus silicon: Matching wits with TD-Gammon. *Inside Backgammon*, 2:14–22.

Rosenblatt, F. (1962). *Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms.* Spartan Books, Washington, DC.

Ross, S. (1983). *Introduction to Stochastic Dynamic Programming.* Academic Press, New York.

Rubinstein, R. Y. (1981). *Simulation and the Monte Carlo Method.* Wiley, New York.

Rumelhart, D. E., Hinton, G. E., Williams, R. J. (1986). Learning internal representations by error propagation. In D. E. Rumelhart and J. L. McClelland (eds.), *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, vol. I, *Foundations.* Bradford/MIT Press, Cambridge, MA.

Rummery, G. A. (1995). *Problem Solving with Reinforcement Learning.* Ph.D. thesis, Cambridge University.

Rummery, G. A., Niranjan, M. (1994). On-line Q-learning using connectionist systems. Technical Report CUED/F-INFENG/TR 166. Engineering Department, Cambridge

University.

Russell, S., Norvig, P. (2009). *Artificial Intelligence: A Modern Approach.* Prentice-Hall, Englewood Cliffs, NJ.

Rust, J. (1996). Numerical dynamic programming in economics. In H. Amman, D. Kendrick, and J. Rust (eds.), *Handbook of Computational Economics*, pp. 614–722. Elsevier, Amsterdam.

Ryan, R. M. and Deci, E. L. (2000). Intrinsic and extrinsic motivations: Classic definitions and new directions. *Contemporary Educational Psychology*, 25(1):54–67.

Saksida, L. M., Raymond, S. M., and Touretzky, D. S. (1997). Shaping robot behavior using principles from instrumental conditioning. *Robotics and Autonomous Systems*, 22(3):231–249.

Samuel, A. L. (1959). Some studies in machine learning using the game of checkers. *IBM Journal on Research and Development*, 3:211–229. Reprinted in E. A. Feigenbaum and J. Feldman (eds.), *Computers and Thought*, pp. 71–105. McGraw-Hill, New York, 1963.

Samuel, A. L. (1967). Some studies in machine learning using the game of checkers. II—Recent progress. *IBM Journal on Research and Development*, 11:601–617.

Schmajuk, N. A. (2008). Computational models of classical conditioning. *Scholarpedia*, 3(3):1664.

Schmidhuber, J. (1991a). Adaptive confidence and adaptive curiosity. Technical Report FKI-149-91, Institut für Informatik, Technische Universität München, Arcisstr. 21, 800 München 2, Germany.

Schmidhuber, J. (1991b). A possibility for implementing curiosity and boredom in model-building neural controllers. In *From Animals to Animats: Proceedings of the First International Conference on Simulation of Adaptive Behavior*, pages 222–227, Cambridge, MA. MIT Press.

Schmidhuber, J. (2009). Driven by compression progress: A simple principle explains essential aspects of subjective beauty, novelty, surprise, interestingness, attention, curiosity, creativity, art, science, music, jokes. In Pezzulo, G., Butz, M. V., Sigaud, O., and Baldassarre, G., editors, *Anticipatory Behavior in Adaptive Learning Systems. From Psychological Theories to Artificial Cognitive Systems*, pages 48–76. Springer, Berlin.

Schmidhuber, J., Storck, J., and Hochreiter, S. (1994). Reinforcement driven information acquisition in nondeterministic environments. Technical report, Fakultät für Informatik, Technische Universität München, München, Germany.

Schultz, D. G., Melsa, J. L. (1967). *State Functions and Linear Control Systems.* McGraw-Hill, New York.

Schultz, W., Dayan, P., Montague, P. R. (1997). A neural substrate of prediction and reward. *Science*, 275:1593–1598.

Schwartz, A. (1993). A reinforcement learning method for maximizing undiscounted rewards. In *Proceedings of the Tenth International Conference on Machine Learning*, pp. 298–305. Morgan Kaufmann, San Mateo, CA.

Schweitzer, P. J., Seidmann, A. (1985). Generalized polynomial approximations in Markovian decision processes. *Journal of Mathematical Analysis and Applications*, 110:568–582.

Selfridge, O. J., Sutton, R. S., Barto, A. G. (1985). Training and tracking in robotics. In A. Joshi (ed.), *Proceedings of the Ninth International Joint Conference on Artificial Intelligence*, pp. 670–672. Morgan Kaufmann, San Mateo, CA.

Seo, H., Barraclough, D., and Lee, D. (2007). Dynamic signals related to choices and outcomes in the dorsolateral prefrontal cortex. *Cerebral Cortex*, 17(suppl 1):110–117.

Shah, A. (2012). Psychological and neuroscientific connections with reinforcement learning. In Wiering, M. and van Otterlo, M., editors, *Reinforcement Learning: State of the Art*, pages 507–537. Springer-Verlag, Berlin.

Shannon, C. E. (1950). Programming a computer for playing chess. *Philosophical Magazine*, 41:256–275.

Shelton, C. R. (2001). *Importance Sampling for Reinforcement Learning with Multiple Objectives*. PhD thesis, Massachusetts Institute of Technology.

Shewchuk, J., Dean, T. (1990). Towards learning time-varying functions with high input dimensionality. In *Proceedings of the Fifth IEEE International Symposium on Intelligent Control*, pp. 383–388. IEEE Computer Society Press, Los Alamitos, CA.

Si, J., Barto, A., Powell, W., Wunsch, D. (Eds.). (2004). *Handbook of learning and approximate dynamic programming*. John Wiley and Sons.

Singh, S. P. (1992a). Reinforcement learning with a hierarchy of abstract models. In *Proceedings of the Tenth National Conference on Artificial Intelligence*, pp. 202–207. AAAI/MIT Press, Menlo Park, CA.

Singh, S. P. (1992b). Scaling reinforcement learning algorithms by learning variable temporal resolution models. In *Proceedings of the Ninth International Machine Learning Conference*, pp. 406–415. Morgan Kaufmann, San Mateo, CA.

Singh, S. P. (1993). *Learning to Solve Markovian Decision Processes*. Ph.D. thesis, University of Massachusetts, Amherst. Appeared as CMPSCI Technical Report 93-77.

Singh, S., Barto, A. G., and Chentanez, N. (2005). Intrinsically motivated reinforcement learning. In *Advances in Neural Information Processing Systems 17: Proceedings of the 2004 Conference*, pages 1281–1288, Cambridge MA. MIT Press.

Singh, S. P., Bertsekas, D. (1997). Reinforcement learning for dynamic channel allocation in cellular telephone systems. In *Advances in Neural Information Processing Systems: Proceedings of the 1996 Conference*, pp. 974–980. MIT Press, Cambridge, MA.

Singh, S. P., Jaakkola, T., Jordan, M. I. (1994). Learning without state-estimation in partially observable Markovian decision problems. In W. W. Cohen and H. Hirsch (eds.), *Proceedings of the Eleventh International Conference on Machine Learning*, pp. 284–292. Morgan Kaufmann, San Francisco.

Singh, S. P., Jaakkola, T., Jordan, M. I. (1995). Reinforcement learing with soft state aggregation. In G. Tesauro, D. S. Touretzky, T. Leen (eds.), *Advances in Neural Information Processing Systems: Proceedings of the 1994 Conference*, pp. 359–368. MIT Press, Cambridge, MA.

Singh, S., Lewis, R. L., and Barto, A. G. (2009). Where do rewards come from? In Taatgen, N. and van Rijn, H., editors, *Proceedings of the 31st Annual Conference of the Cognitive Science Society*, pages 2601–2606. Cognitive Science Society.

Singh, S., Lewis, R. L., Barto, A. G., and Sorg, J. (2010). Intrinsically motivated reinforcement learning: An evolutionary perspective. *IEEE Transactions on Autonomous Mental Development*, 2(2):7082. Special issue on Active Learning and Intrinsically Motivated Exploration in Robots: Advances and Challenges.

Singh, S. P., Sutton, R. S. (1996). Reinforcement learning with replacing eligibility traces. *Machine Learning*, 22:123–158.

Sivarajan, K. N., McEliece, R. J., Ketchum, J. W. (1990). Dynamic channel assignment in cellular radio. In *Proceedings of the 40th Vehicular Technology Conference*, pp. 631–637.

Skinner, B. F. (1938). *The Behavior of Organisms: An Experimental Analysis.* Appleton-Century, New York.

Skinner, B. F. (1958). Reinforcement today. *American Psychologist*, 13(3):94–99.

Skinner, B. F. (1981). Selection by consequences. *Science 213*(4507):501–504.

Smith, K. S. and Greybiel, A. M. (2013). A dual operator view of habitual behavior reflecting cortical and striatal dynamics. *Neuron*, 79(2):361–374.

Sofge, D. A., White, D. A. (1992). Applied learning: Optimal control for manufacturing. In D. A. White and D. A. Sofge (eds.), *Handbook of Intelligent Control: Neural, Fuzzy, and Adaptive Approaches*, pp. 259–281. Van Nostrand Reinhold, New York.

Sorg, J., Singh, S., and Lewis, R. (2010). Internal rewards mitigate agent boundedness. In *Proceedings of the 27th International Conference on Machine Learning (ICML)*, pages 1007–1014.

Sorg, J. D. (2011). *The Optimal Reward Problem:Designing Effective Reward for Bounded Agents.* PhD thesis, Computer Science and Engineering, The University of Michigan.

Spence, K. W. (1947). The role of secondary reinforcement in delayed reward learning. *Psychological Review*, 54(1):1–8.

Spong, M. W. (1994). Swing up control of the acrobot. In *Proceedings of the 1994 IEEE Conference on Robotics and Automation*, pp. 2356-2361. IEEE Computer Society Press, Los Alamitos, CA.

Staddon, J. E. R. (1983). *Adaptive Behavior and Learning.* Cambridge University Press, Cambridge.

Storck, J., Hochreiter, S., and Schmidhuber, J. (1995). Reinforcement-driven information acquisition in non-deterministic environments. In *Proceedings of ICANN'95, Paris, France*, volume 2, pages 159–164.

Sugiyama, M., Hachiya, H., Morimura, T. (2013). *Statistical Reinforcement Learning: Modern Machine Learning Approaches.* Chapman & Hall/CRC.

Sutton, R. S. (1978a). Learning theory support for a single channel theory of the brain. Unpublished report.

Sutton, R. S. (1978b). Single channel theory: A neuronal theory of learning. *Brain Theory Newsletter*, 4:72–75. Center for Systems Neuroscience, University of Massachusetts, Amherst, MA.

Sutton, R. S. (1978c). A unified theory of expectation in classical and instrumental conditioning. Bachelors thesis, Stanford University.

Sutton, R. S. (1984). *Temporal Credit Assignment in Reinforcement Learning.* Ph.D. thesis, University of Massachusetts, Amherst.

Sutton, R. S. (1988). Learning to predict by the method of temporal differences. *Machine Learning*, 3:9–44.

Sutton, R. S. (1990). Integrated architectures for learning, planning, and reacting based on approximating dynamic programming. In *Proceedings of the Seventh International Conference on Machine Learning*, pp. 216–224. Morgan Kaufmann, San Mateo, CA.

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

Sutton, R. S. (1991b). Planning by incremental dynamic programming. In L. A. Birnbaum and G. C. Collins (eds.), *Proceedings of the Eighth International Workshop on Machine Learning*, pp. 353–357. Morgan Kaufmann, San Mateo, CA.

Sutton, R. S. (1995). TD models: Modeling the world at a mixture of time scales. In

A. Prieditis and S. Russell (eds.), *Proceedings of the Twelfth International Conference on Machine Learning*, pp. 531–539. Morgan Kaufmann, San Francisco.

Sutton, R. S. (1996). Generalization in reinforcement learning: Successful examples using sparse coarse coding. In D. S. Touretzky, M. C. Mozer and M. E. Hasselmo (eds.), *Advances in Neural Information Processing Systems: Proceedings of the 1995 Conference*, pp. 1038–1044. MIT Press, Cambridge, MA.

Sutton, R. S. (ed.). (1992). Special issue of *Machine Learning* on reinforcement learning, 8. Also published as *Reinforcement Learning*. Kluwer Academic, Boston, 1992.

Sutton, R. S., Barto, A. G. (1981a). Toward a modern theory of adaptive networks: Expectation and prediction. *Psychological Review*, 88:135–170.

Sutton, R. S., Barto, A. G. (1981b). An adaptive network that constructs and uses an internal model of its world. *Cognition and Brain Theory*, 3:217–246.

Sutton, R. S., Barto, A. G. (1987). A temporal-difference model of classical conditioning. In *Proceedings of the Ninth Annual Conference of the Cognitive Science Society*, pp. 355-378. Erlbaum, Hillsdale, NJ.

Sutton, R. S., Barto, A. G. (1990). Time-derivative models of Pavlovian reinforcement. In M. Gabriel and J. Moore (eds.), *Learning and Computational Neuroscience: Foundations of Adaptive Networks*, pp. 497–537. MIT Press, Cambridge, MA.

Sutton, R. S., Mahmood, A. R., Precup, D., van Hasselt, H. (2014). A new Q(λ) with interim forward view and Monte Carlo equivalence. In *Proceedings of the 31st International Conference on Machine Learning*, Beijing, China.

Sutton, R. S., Pinette, B. (1985). The learning of world models by connectionist networks. In *Proceedings of the Seventh Annual Conference of the Cognitive Science Society*, pp. 54–64.

Sutton, R. S., Singh, S. (1994). On bias and step size in temporal-difference learning. In *Proceedings of the Eighth Yale Workshop on Adaptive and Learning Systems*, pp. 91–96. Center for Systems Science, Dunham Laboratory, Yale University, New Haven.

Sutton, R. S., Whitehead, D. S. (1993). Online learning with random representations. In *Proceedings of the Tenth International Machine Learning Conference*, pp. 314–321. Morgan Kaufmann, San Mateo, CA.

Szepesvári, C. (2010). Algorithms for reinforcement learning. Synthesis Lectures on Artificial Intelligence and Machine Learning 4(1), 1–103.

Szita, I. (2012). Reinforcement learning in games. In *Reinforcement Learning* (pp. 539-577). Springer Berlin Heidelberg.

Tadepalli, P., Ok, D. (1994). H-learning: A reinforcement learning method to optimize undiscounted average reward. Technical Report 94-30-01. Oregon State University, Computer Science Department, Corvallis.

Tan, M. (1991). Learning a cost-sensitive internal representation for reinforcement learning. In L. A. Birnbaum and G. C. Collins (eds.), *Proceedings of the Eighth International Workshop on Machine Learning*, pp. 358–362. Morgan Kaufmann, San Mateo, CA.

Tan, M. (1993). Multi-agent reinforcement learning: Independent vs. cooperative agents. In *Proceedings of the Tenth International Conference on Machine Learning*, pp. 330–337. Morgan Kaufmann, San Mateo, CA.

Tesauro, G. J. (1986). Simple neural models of classical conditioning. *Biological Cybernetics*, 55:187–200.

Tesauro, G. J. (1992). Practical issues in temporal difference learning. *Machine Learning*,

8:257–277.

Tesauro, G. J. (1994). TD-Gammon, a self-teaching backgammon program, achieves master-level play. *Neural Computation*, 6(2):215–219.

Tesauro, G. J. (1995). Temporal difference learning and TD-Gammon. *Communications of the ACM*, 38:58–68.

Tesauro, G. J., Galperin, G. R. (1997). On-line policy improvement using Monte-Carlo search. In *Advances in Neural Information Processing Systems: Proceedings of the 1996 Conference*, pp. 1068–1074. MIT Press, Cambridge, MA.

Tham, C. K. (1994). *Modular On-Line Function Approximation for Scaling up Reinforcement Learning.* PhD thesis, Cambridge University.

Thathachar, M. A. L. and Sastry, P. S. (1985). A new approach to the design of reinforcement schemes for learning automata. *IEEE Transactions on Systems, Man, and Cybernetics*, 15:168–175.

Thistlethwaite, D. (1951). A critical review of latent learning and related experiments. *Psychological Bulletin*, 48(2):97–129.

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

Thompson, W. R. (1934). On the theory of apportionment. *American Journal of Mathematics*, 57:450–457.

Thorndike, E. L. (1898). Animal intelligence: An experimental study of the associative processes in animals. *The Psychological Review, Series of Monograph Supplements*, II(4).

Thorndike, E. L. (1911). *Animal Intelligence.* Hafner, Darien, CT.

Thorp, E. O. (1966). *Beat the Dealer: A Winning Strategy for the Game of Twenty-One.* Random House, New York.

Tolman, E. C. (1932). *Purposive Behavior in Animals and Men.* Century, New York.

Tolman, E. C. (1948). Cognitive maps in rats and men. *Psychological Review*, 55(4):189–208.

Tsetlin, M. L. (1973). *Automaton Theory and Modeling of Biological Systems.* Academic Press, New York.

Tsitsiklis, J. N. (1994). Asynchronous stochastic approximation and Q-learning. *Machine Learning*, 16:185–202.

Tsitsiklis, J. N. (2002). On the convergence of optimistic policy iteration. *Journal of Machine Learning Research*, 3:59–72.

Tsitsiklis, J. N. and Van Roy, B. (1996). Feature-based methods for large scale dynamic programming. *Machine Learning*, 22:59–94.

Tsitsiklis, J. N., Van Roy, B. (1997). An analysis of temporal-difference learning with function approximation. *IEEE Transactions on Automatic Control*, 42:674–690.

Tsitsiklis, J. N., Van Roy, B. (1999). Average cost temporal-difference learning. *Automatica*, 35:1799–1808.

Turing, A. M. (1950). Computing machinery and intelligence. *Mind* 433–460.

Turing, A. M. (1948). Intelligent Machinery, A Heretical Theory. *The Turing Test: Verbal Behavior as the Hallmark of Intelligence*, 105.

Ungar, L. H. (1990). A bioreactor benchmark for adaptive network-based process control. In W. T. Miller, R. S. Sutton, and P. J. Werbos (eds.), *Neural Networks for Control*, pp. 387–402. MIT Press, Cambridge, MA.

Urbanowicz, R. J., Moore, J. H. (2009). Learning classifier systems: A complete introduction, review, and roadmap. *Journal of Artificial Evolution and Applications.*

van Hasselt, H. (2010). Double Q-learning. In *Advances in Neural Information Processing Systems*, pp. 2613–2621.

van Hasselt, H. (2011). *Insights in Reinforcement Learning: Formal Analysis and Empircal Evaluation of Temporal-difference Learning.* SIKS dissertation series number 2011-04.

van Hasselt, H., Sutton, R. S. (in prep.). Learning to predict independent of span.

Van Roy, B., Bertsekas, D. P., Lee, Y., Tsitsiklis, J. N. (1997). A neuro-dynamic programming approach to retailer inventory management. In *Proceedings of the 36th IEEE Conference on Decision and Control*, Vol. 4, pp. 4052–4057.

van Seijen, H., Van Hasselt, H., Whiteson, S., Wiering, M. (2009). A theoretical and empirical analysis of Expected Sarsa. In *IEEE Symposium on Adaptive Dynamic Programming and Reinforcement Learning*, pp. 177-184.

van Seijen, H., Sutton, R. S. (2014). True online TD(λ). In *Proceedings of the 31st International Conference on Machine Learning.* JMLR W&CP 32(1):692–700.

van Seijin, H., Mahmood, A. R., Pilarski, P. M., Machado, M. C., Sutton, R. S. (in prep.). True online temporal-difference learning.

Waltz, M. D., Fu, K. S. (1965). A heuristic approach to reinforcement learning control systems. *IEEE Transactions on Automatic Control*, 10:390–398.

Watkins, C. J. C. H. (1989). *Learning from Delayed Rewards.* Ph.D. thesis, Cambridge University.

Watkins, C. J. C. H., Dayan, P. (1992). Q-learning. *Machine Learning*, 8:279–292.

Wiering, M., Van Otterlo, M. (2012). *Reinforcement Learning.* Springer Berlin Heidelberg.

Werbos, P. J. (1977). Advanced forecasting methods for global crisis warning and models of intelligence. *General Systems Yearbook*, 22:25–38.

Werbos, P. J. (1982). Applications of advances in nonlinear sensitivity analysis. In R. F. Drenick and F. Kozin (eds.), *System Modeling and Optimization*, pp. 762–770. Springer-Verlag, Berlin.

Werbos, P. J. (1987). Building and understanding adaptive systems: A statistical/numerical approach to factory automation and brain research. *IEEE Transactions on Systems, Man, and Cybernetics*, 17:7–20.

Werbos, P. J. (1988). Generalization of back propagation with applications to a recurrent gas market model. *Neural Networks*, 1:339–356.

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

Werbos, P. J. (1990). Consistency of HDP applied to a simple reinforcement learning problem. *Neural Networks*, 3:179–189.

Werbos, P. J. (1992). Approximate dynamic programming for real-time control and neural modeling. In D. A. White and D. A. Sofge (eds.), *Handbook of Intelligent Control: Neural, Fuzzy, and Adaptive Approaches*, pp. 493–525. Van Nostrand Reinhold, New York.

White, D. J. (1969). *Dynamic Programming.* Holden-Day, San Francisco.

White, D. J. (1985). Real applications of Markov decision processes. *Interfaces*, 15:73–83.

White, D. J. (1988). Further real applications of Markov decision processes. *Interfaces*,

18:55–61.

White, D. J. (1993). A survey of applications of Markov decision processes. *Journal of the Operational Research Society*, 44:1073–1096.

Whitehead, S. D., Ballard, D. H. (1991). Learning to perceive and act by trial and error. *Machine Learning*, 7:45–83.

Whitt, W. (1978). Approximations of dynamic programs I. *Mathematics of Operations Research*, 3:231–243.

Whittle, P. (1982). *Optimization over Time*, vol. 1. Wiley, New York.

Whittle, P. (1983). *Optimization over Time*, vol. 2. Wiley, New York.

Widrow, B., Gupta, N. K., Maitra, S. (1973). Punish/reward: Learning with a critic in adaptive threshold systems. *IEEE Transactions on Systems, Man, and Cybernetics*, 3:455–465.

Widrow, B., Hoff, M. E. (1960). Adaptive switching circuits. In *1960 WESCON Convention Record Part IV*, pp. 96–104. Institute of Radio Engineers, New York. Reprinted in J. A. Anderson and E. Rosenfeld, *Neurocomputing: Foundations of Research*, pp. 126–134. MIT Press, Cambridge, MA, 1988.

Widrow, B., Smith, F. W. (1964). Pattern-recognizing control systems. In J. T. Tou and R. H. Wilcox (eds.), *Computer and Information Sciences*, pp. 288–317. Spartan, Washington, DC.

Widrow, B., Stearns, S. D. (1985). *Adaptive Signal Processing.* Prentice-Hall, Englewood Cliffs, NJ.

Williams, R. J. (1986). Reinforcement learning in connectionist networks: A mathematical analysis. Technical Report ICS 8605. Institute for Cognitive Science, University of California at San Diego, La Jolla.

Williams, R. J. (1987). Reinforcement-learning connectionist systems. Technical Report NU-CCS-87-3. College of Computer Science, Northeastern University, Boston.

Williams, R. J. (1988). On the use of backpropagation in associative reinforcement learning. In *Proceedings of the IEEE International Conference on Neural Networks*, pp. I263–I270. IEEE San Diego section and IEEE TAB Neural Network Committee.

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

Williams, R. J., Baird, L. C. (1990). A mathematical analysis of actor-critic architectures for learning optimal controls through incremental dynamic programming. In *Proceedings of the Sixth Yale Workshop on Adaptive and Learning Systems*, pp. 96–101. Center for Systems Science, Dunham Laboratory, Yale University, New Haven.

Wilson, R. C., Takahashi, Y. K., Schoenbaum, G., and Niv, Y. (2014). Orbitofrontal cortex as a cognitive map of task space. *Neuron*, 81(2):267–279.

Wilson, S. W. (1994). ZCS: A zeroth order classifier system. *Evolutionary Computation*, 2:1–18.

Wise, R. A. (2004). Dopamine, learning, and motivation. *Nature Reviews Neuroscience*, 5(6):1–12.

Witten, I. H. (1976). The apparent conflict between estimation and control—A survey of the two-armed problem. *Journal of the Franklin Institute*, 301:161–189.

Witten, I. H. (1977). An adaptive optimal controller for discrete-time Markov environments. *Information and Control*, 34:286–295.

Witten, I. H., Corbin, M. J. (1973). Human operators and automatic adaptive controllers: A comparative study on a particular control task. *International Journal of Man–Machine Studies*, 5:75–104.

Woodworth, R. S., Schlosberg, H. (1938). *Experimental psychology.* New York: Henry Holt and Company.

Yee, R. C., Saxena, S., Utgoff, P. E., Barto, A. G. (1990). Explaining temporal differences to create useful concepts for evaluating states. In *Proceedings of the Eighth National Conference on Artificial Intelligence*, pp. 882–888. AAAI Press, Menlo Park, CA.

Young, P. (1984). *Recursive Estimation and Time-Series Analysis.* Springer-Verlag, Berlin.

Yu, H. (2012). Least squares temporal difference methods: An analysis under general conditions. *SIAM Journal on Control and Optimization*, 50(6), 3310–3343.

Zhang, M., Yum, T. P. (1989). Comparisons of channel-assignment strategies in cellular mobile telephone systems. *IEEE Transactions on Vehicular Technology*, 38:211-215.

Zhang, W. (1996). *Reinforcement Learning for Job-shop Scheduling.* Ph.D. thesis, Oregon State University. Technical Report CS-96-30-1.

Zhang, W., Dietterich, T. G. (1995). A reinforcement learning approach to job-shop scheduling. In *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence*, pp. 1114–1120. Morgan Kaufmann.

Zhang, W., Dietterich, T. G. (1996). High-performance job-shop scheduling with a time–delay TD(λ) network. In D. S. Touretzky, M. C. Mozer, M. E. Hasselmo (eds.), *Advances in Neural Information Processing Systems: Proceedings of the 1995 Conference*, pp. 1024–1030. MIT Press, Cambridge, MA.

Zweben, M., Daun, B., Deale, M. (1994). Scheduling and rescheduling with iterative repair. In M. Zweben and M. S. Fox (eds.), *Intelligent Scheduling*, pp. 241–255. Morgan Kaufmann, San Francisco.

Index