**2.1 Принятие решений и обучение с подкреплением**

Поведенческая нейробиология решает различные вопросы, в числе которых вопрос принятия решений. Выбор оптимального действия основан на прогнозах долгосрочных последствий. [29] Например, 1) часто вкусу предшествует запах: почувствовав приятный аромат с кухни мы предполагаем, что вскоре сможем вкусно поесть; 2) мы не будем переходить дорогу даже на зеленый сигнал светофора, если на нас с высокой скоростью летит автомобиль.

Вопрос принятия решений был подробно исследован в поведенческой психологии на основе парадигмы Павлова. Так же появились и другие доказательства, что существуют области, контролирующие различные аспекты усвоенного поведения. В частности, множество доказательств свидетельствует о том, что нейромедиатор дофамин обеспечивает целевые структуры базальных ганглиев фазовыми сигналами, которые передают ошибку прогнозирования, что в свою очередь влияет на обучение и выбор действия. [30 - 32] Так как вопрос принятия решений связан с прогнозированием, очевидно, что он связан и с обучением.

Современная интерпретация обучения с подкреплением исторически возникла из двух отдельных исследований. Первое исследование провели Ричард Саттон и Эндрю Барто, заинтересованные изучением создания искусственного интеллекта. Они разработали основные алгоритмы и концепции обучения с подкреплением. [33] Вторым исследованием занимались Димитрий Бертсекас и Джон Цициклис, специалисты в области исследования операций и оптимального управления. Их разработки заключаются в стохастических приближениях к методам динамического программирования, которые были названы нейродинамическим программированием. [34] Слияние этих двух направлений исследований сформулировало поведенчески вдохновленные алгоритмы эвристического обучения с подкреплением и предоставило инструменты для анализа свойств этих алгоритмов. [29]

«Обучение с подкреплением - это обучение некого агента тому, что надо делать, как следует отображать ситуации в действия, чтобы максимизировать некоторый сигнал поощрения (вознаграждения), принимающий числовые значения. Обучение с подкреплением состоит из следующих элементов: стратегия, функция поощрения, функция ценности и модель среды (в системах, имитирующих обучение с подкреплением)». [35] Стратегия определяет поведение агента в каждый момент времени, функция поощрения является характеристикой вознаграждения, функция ценности характеризует состояние с учетом долгосрочных перспектив, а модель среды имитирует действия рассматриваемой среды. [35]

В процессе обучения агент взаимодействует со средой, получает от нее ответ, в зависимости от которого принимает дальнейшие действия. Каждое новое состояние характеризуется новым вознаграждением. Максимизация награды является целью обучения с подкреплением, для достижения которой могут использоваться различные алгоритмы: динамическое программирование, Прогнозирование и управление Монте-Карло, SARSA, TD-обучение и Q-обучение. [36 - 41] Все они были рассмотрены Саттоном и Барто. [35]

Список литературы для 2.1

29. Yael Niv **Reinforcement learning in the brain**

Psychology Department & Princeton Neuroscience Institute, Princeton University

2009

30. Barto, A. G. (1995). Adaptive critic and the basal ganglia. In J. C. Houk, J. L. Davis, &

D. G. Beiser (Eds.), *Models of information processing in the basal ganglia* (p. 215-232).

Cambridge: MIT Press.

31. Schultz, W., Dayan, P., & Montague, P. R. (1997). A neural substrate of prediction and reward.

*Science*, *275*, 1593-1599.

32. Wickens, J. R., & Kotter, R. (1995). Cellular models of reinforcement. In J. C. Houk, J. L.¨

Davis, & D. G. Beiser (Eds.), *Models of information processing in the basal ganglia* (p.

187-214). MIT Press.

33. Barto, A. G., Sutton, R. S., & Watkins, C. J. C. H. (1990). Learning and sequential decision

making. In M. Gabriel & J. Moore (Eds.), Learning and computational neuroscience:

Foundations of adaptive networks (p. 539-602). Cambridge, MA: MIT Press.

34. Bertsekas, D. P., & Tsitsiklis, J. N. (1996). Neuro-dynamic programming. Athena Sc.

35. Sutton, R. S. and Barto, A. G. (1998). Reinforcement Learning: An Introduction. Bradford Books, MIT Press, Cambridge, MA, 2002 edition.

36. Bellman, R. E. (1957). Dynamic Programming. Princeton University Press, Princeton, NJ.

37. Sutton, R. S. (1988). Learning to predict by the methods of temporal differences. Mach. Learn.

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

39. Rummery, G. A. (1995). Problem solving with reinforcement learning. PhD thesis, Cambridge University, Cambridge.

40. Watkins, C. J. C. H. (1989). Learning from delayed rewards. PhD thesis, University of Cambridge, Cambridge, England.

41. Watkins, C. J. C. H. and Dayan, P. (1992). Technical note:Q-Learning. Mach. Learn., 8:279–292.