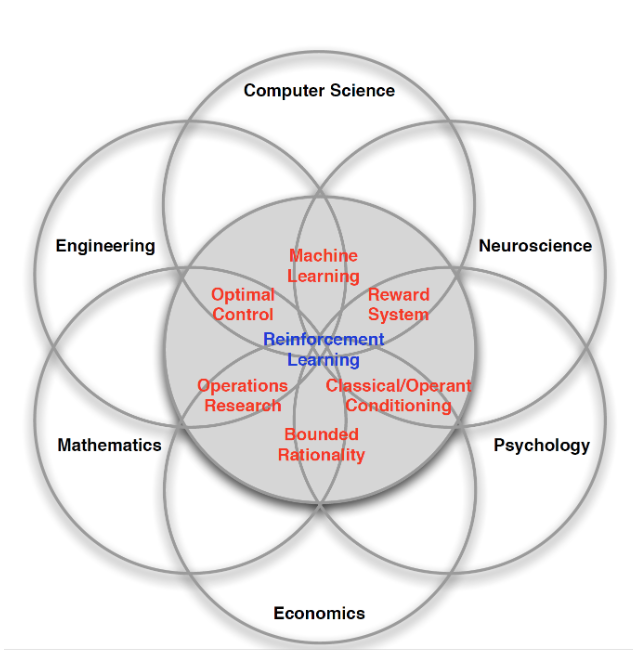
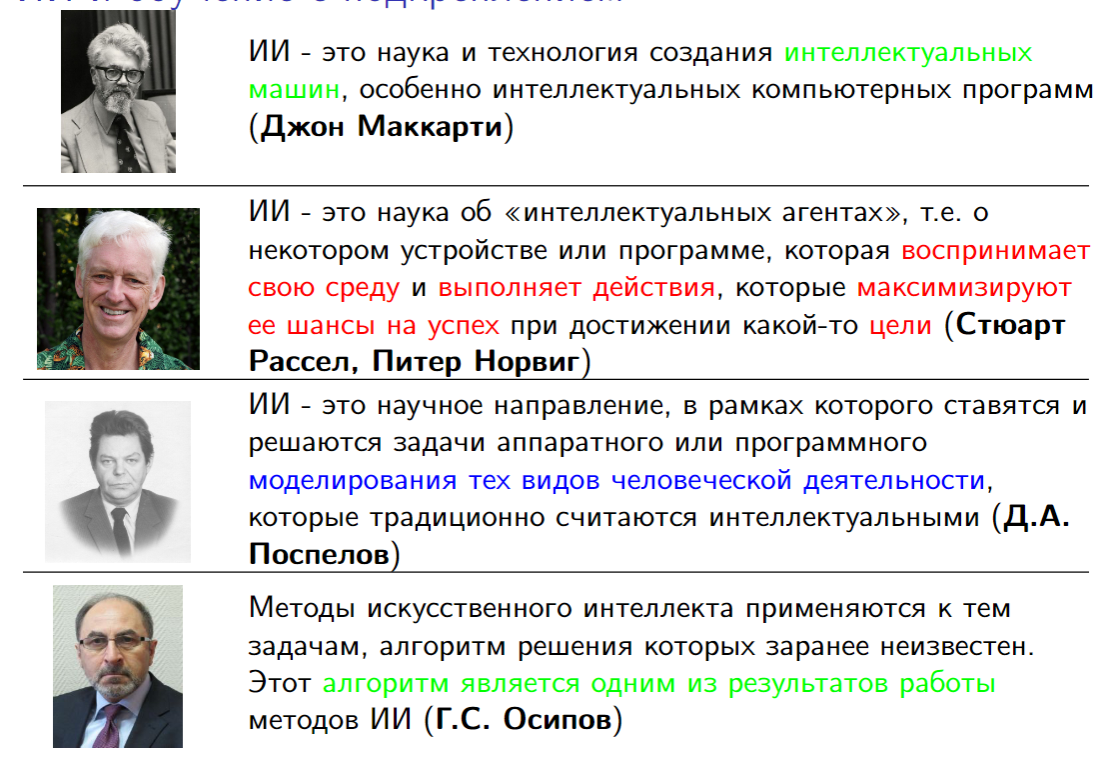
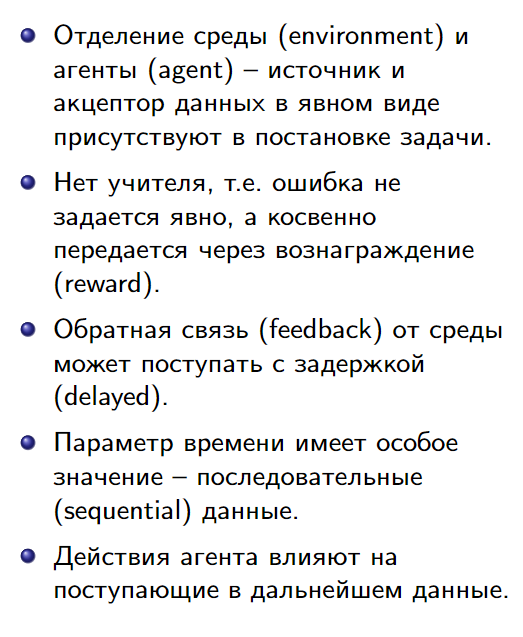
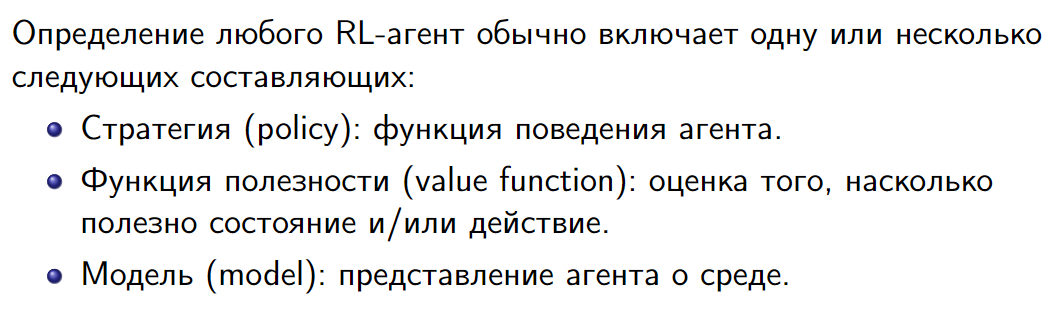
# 1. Место обучения с подкреплением в научной дисциплине искусственного интеллекта и среди других наук.

****

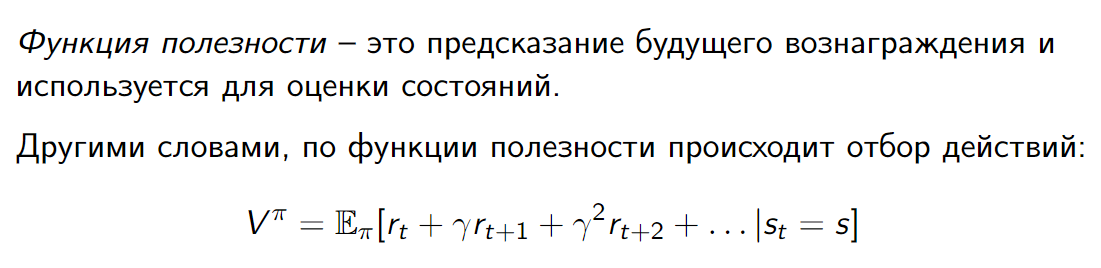
****

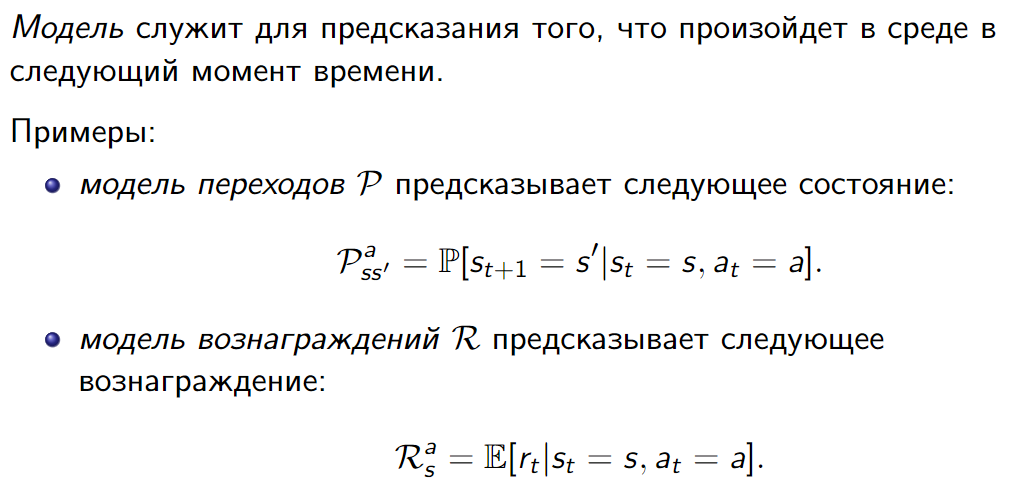
# 2. Общая постановка задачи: среда, агент, вознаграждение, стратегия. Типизация RL агентов.

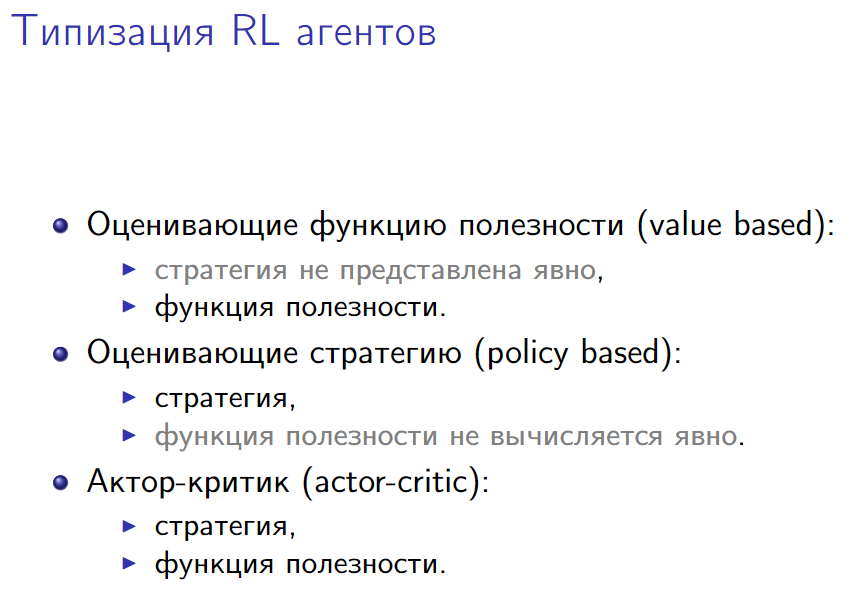
****

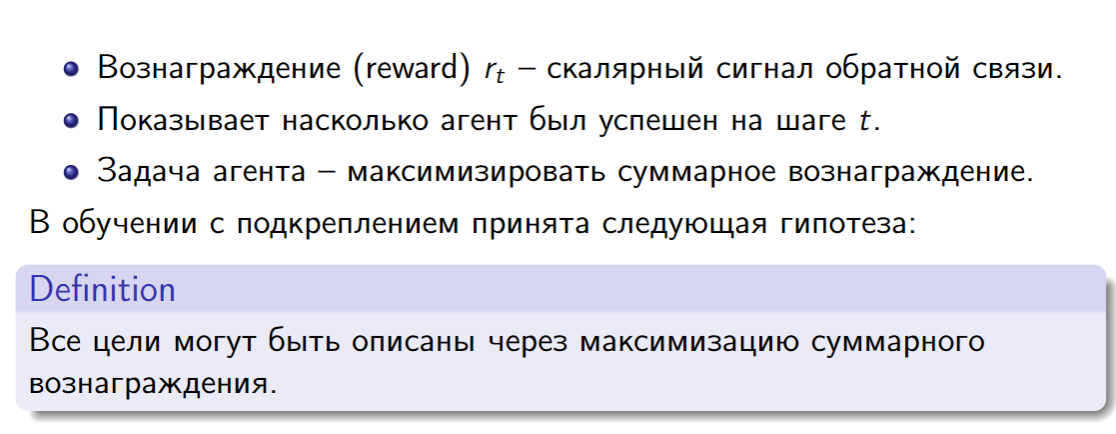
****

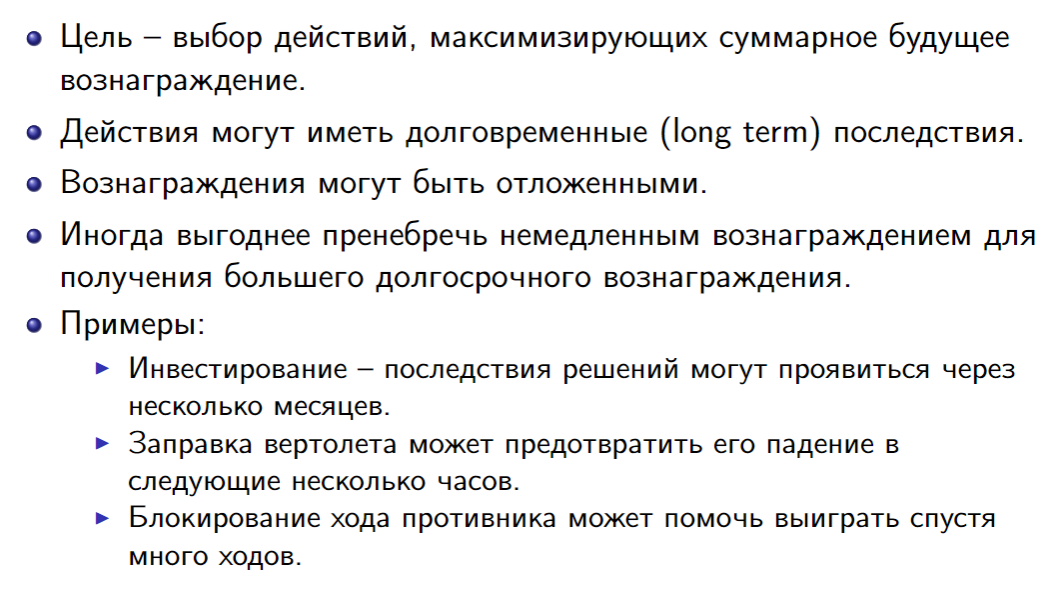
****

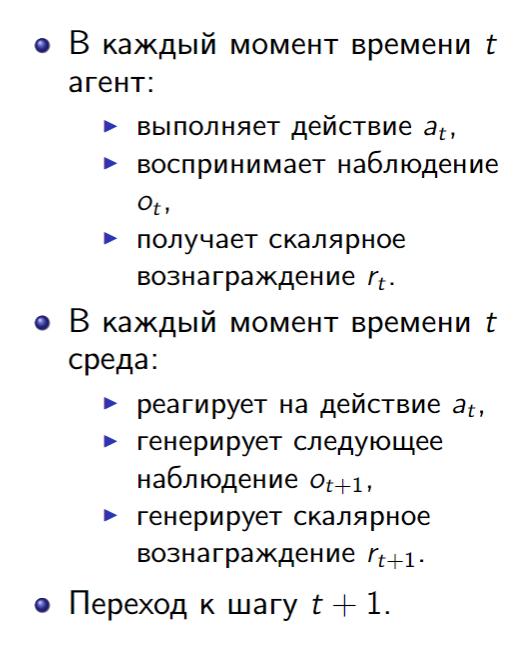
****

****

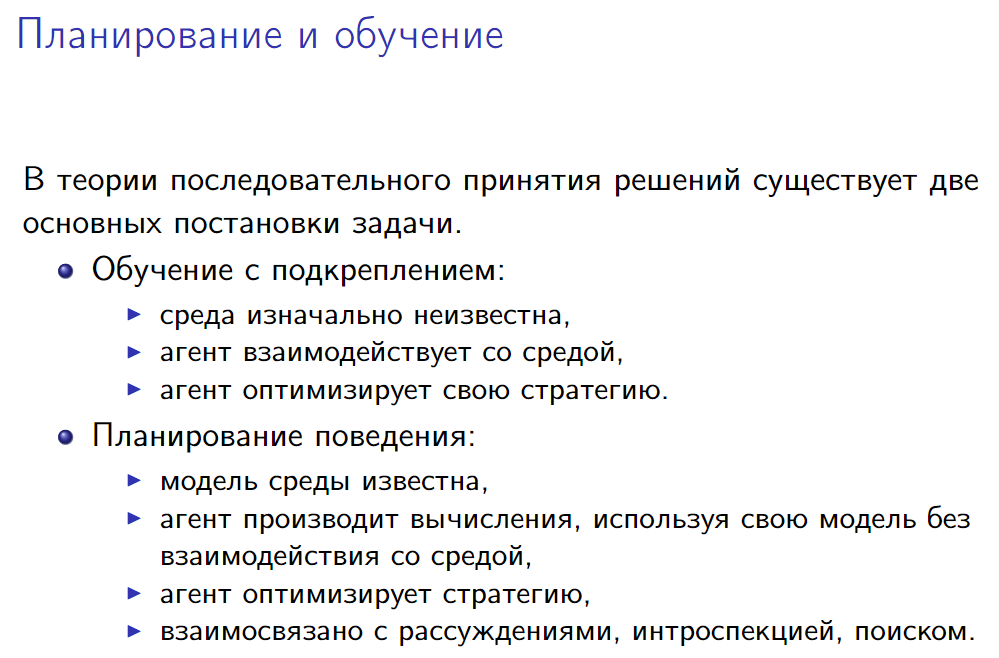
****

****

****

****

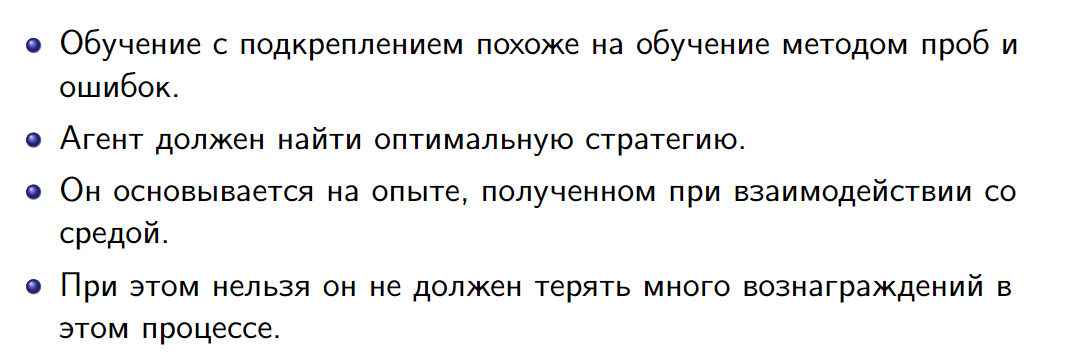
# 3. Основные подзадачи: планирование и обучение, применение и исследование, предсказание и управление.

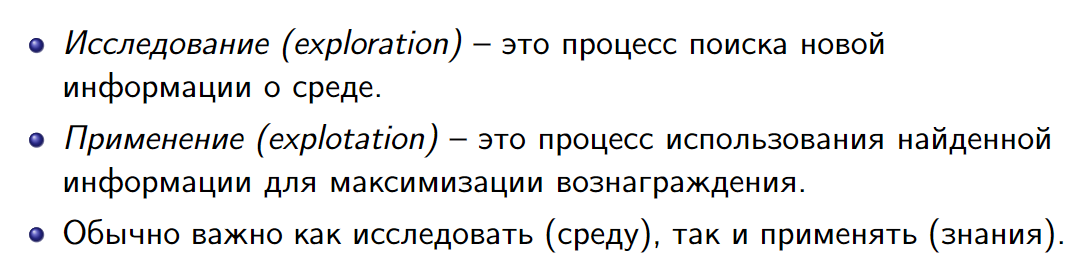
****

**Пример обучения**

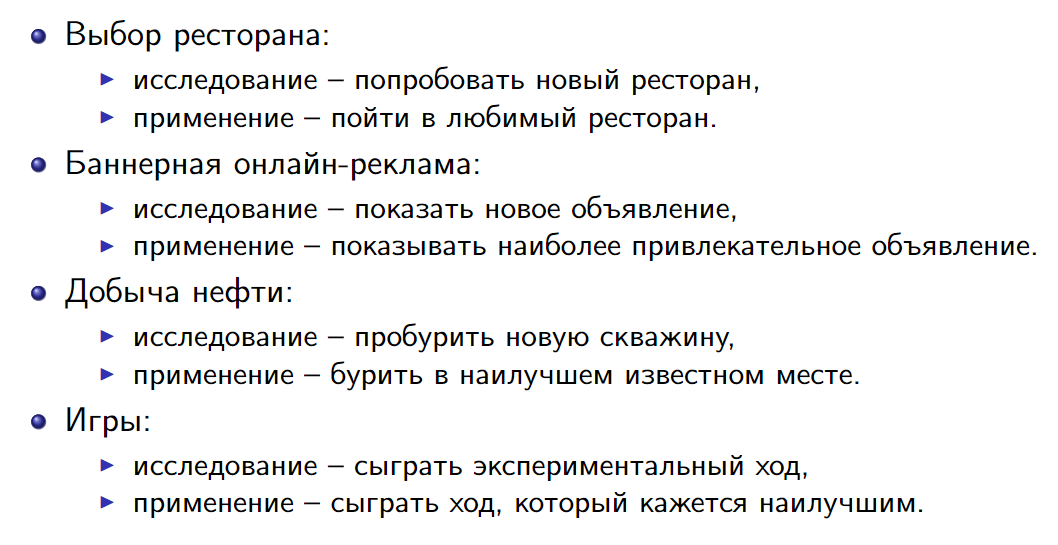
****

**Исследование и применение**

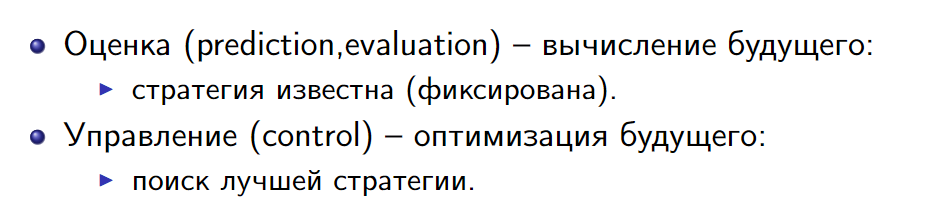
****

****

**Исследование и примеры**

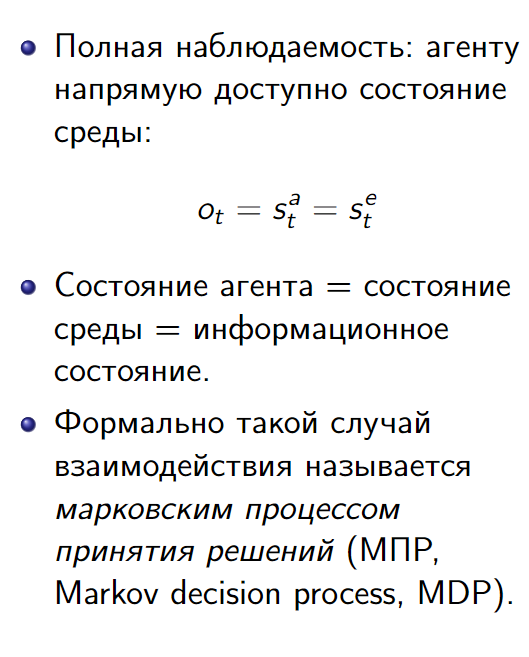
****

**Предсказание и управление**

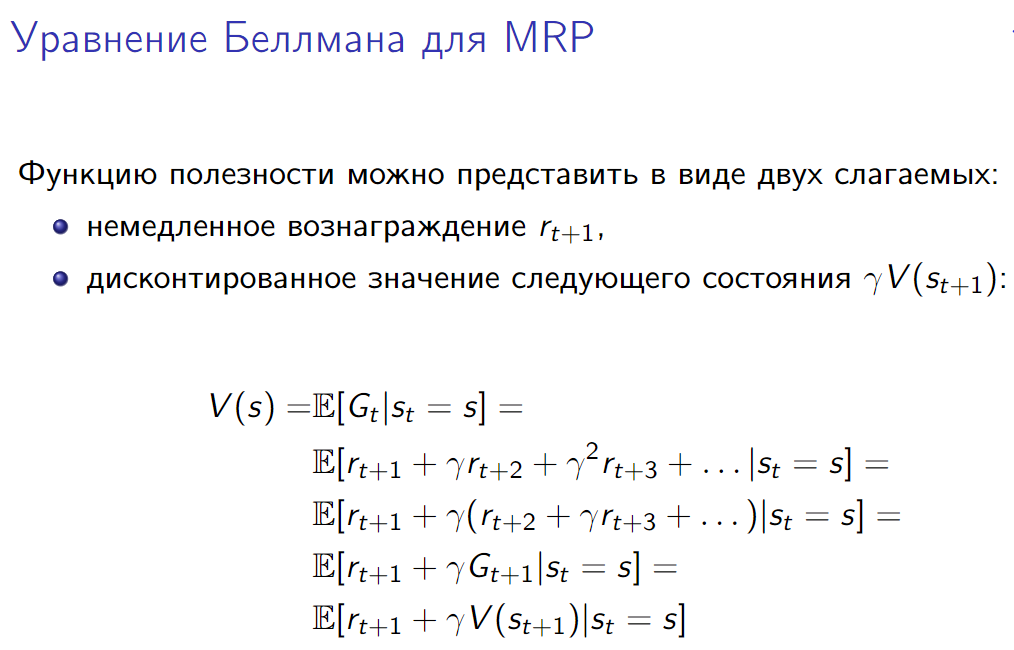
****

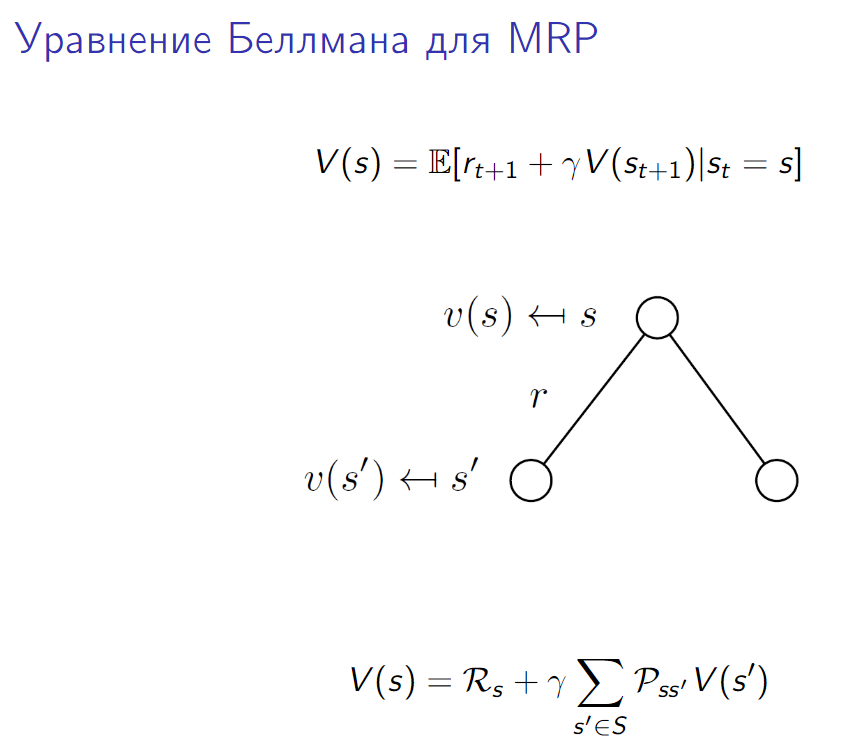
# 4. Определение и примеры марковского процесса, марковского процесса вознаграждения и марковского процесса принятия решений. Стратегия и функции полезности для МППР.

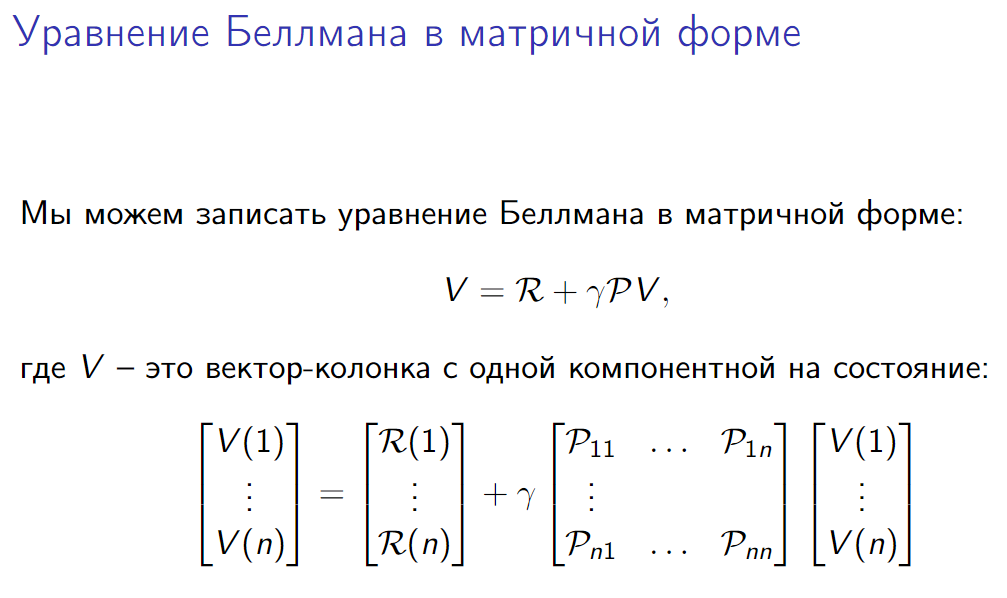
****

****

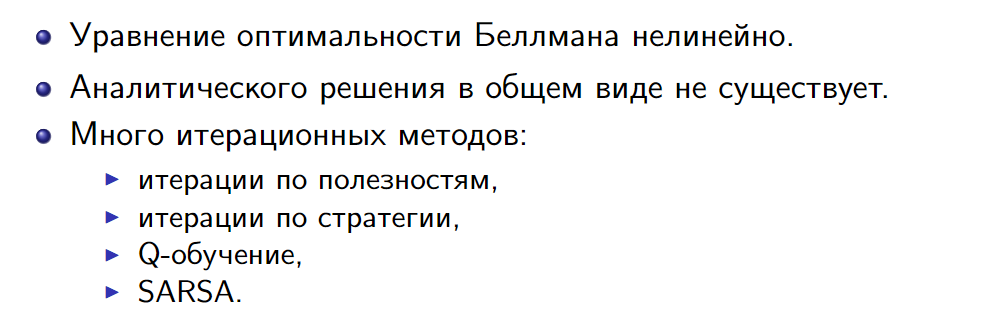
# 5. Уравнение Беллмана: формулировка, примеры, матричная форма. Оптимальность: оптимальные полезности, стратегия и уравнение Беллмана.







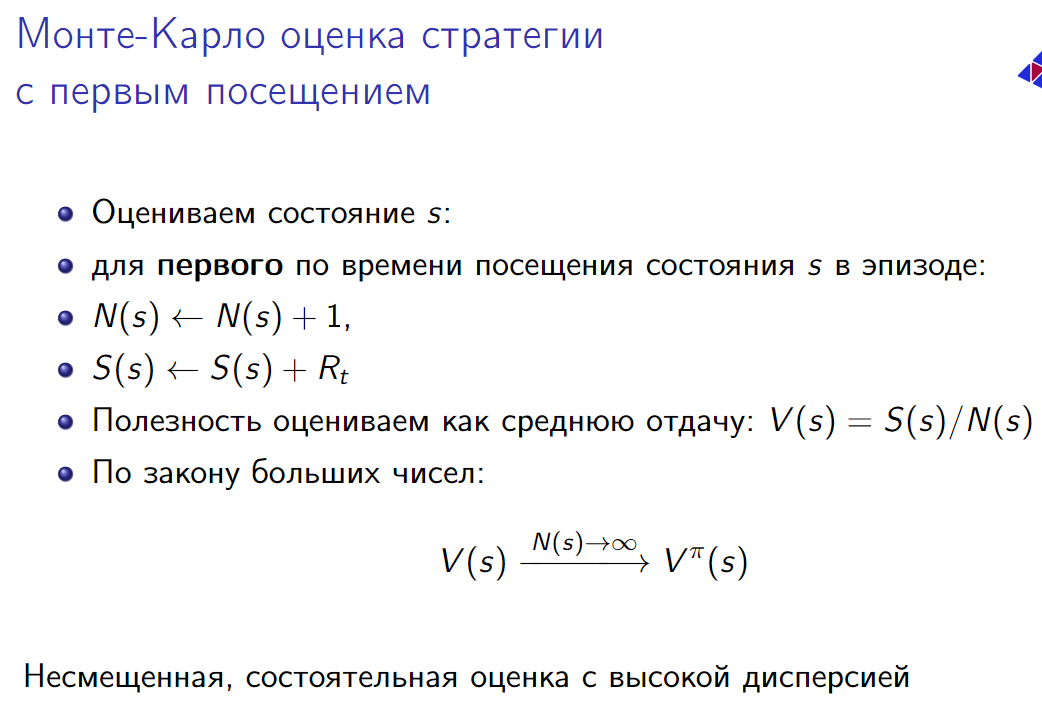
Решение уравнения оптимальности Беллмана

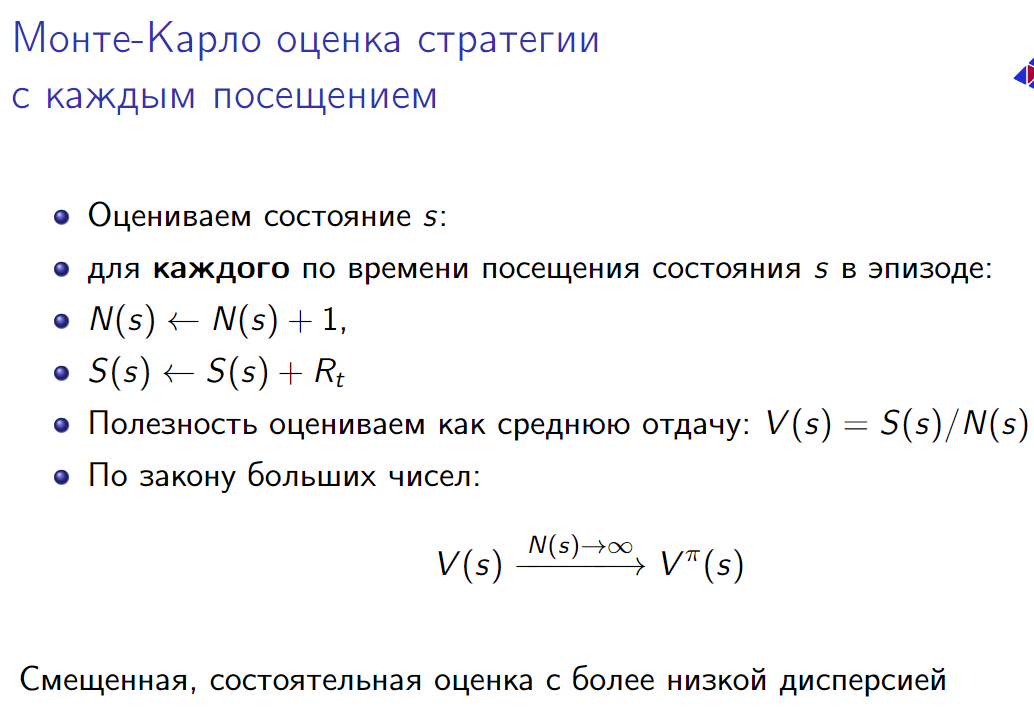


# 6. Итерационная оценка стратегии. Общий алгоритм итерации по стратегиям. Итерация по полезностям.

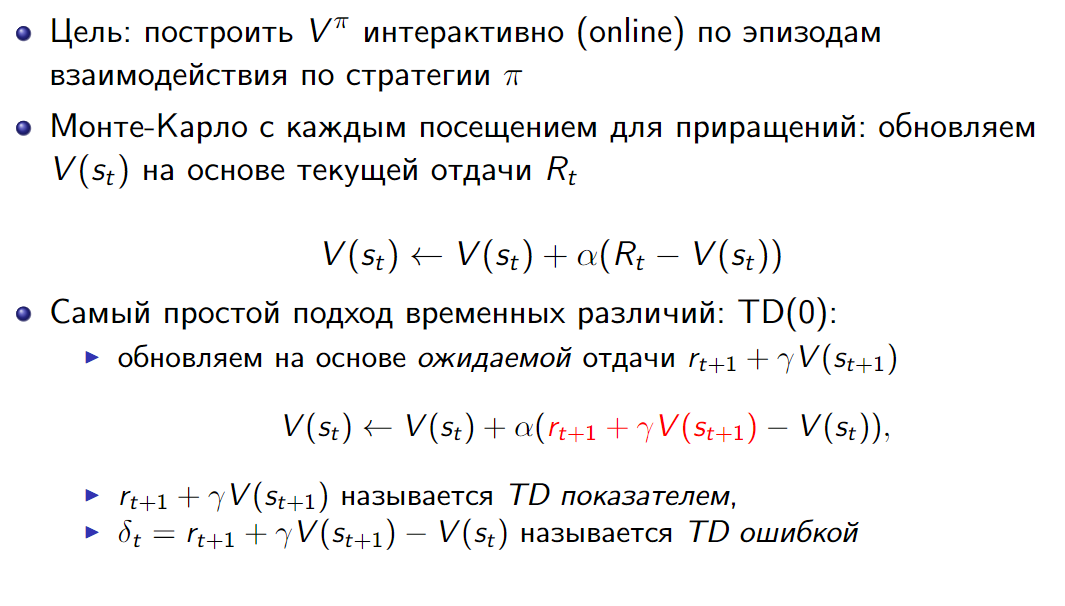
# 7. Методы Монте-Карло (MC) с первым и с каждым посещением.



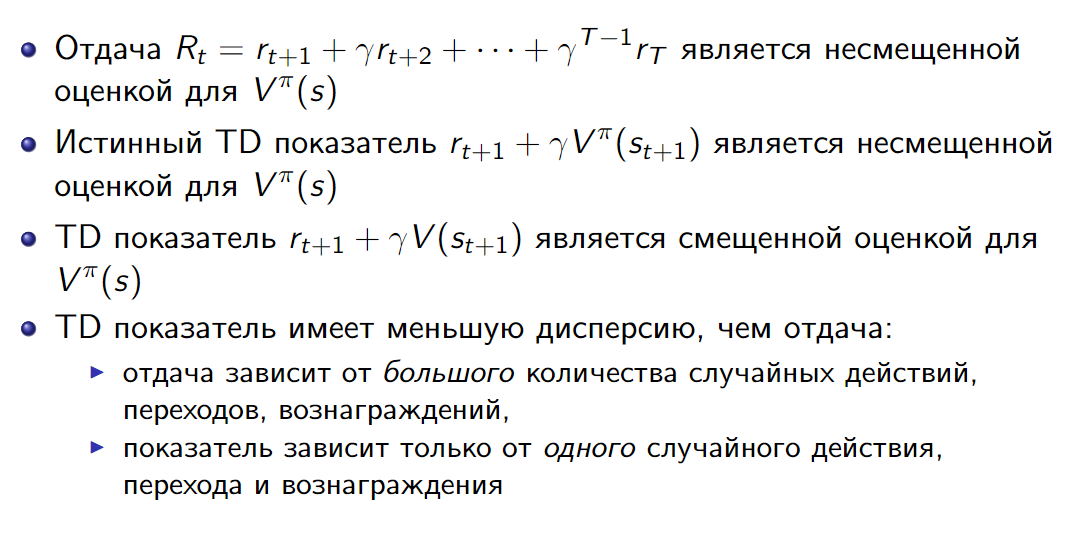




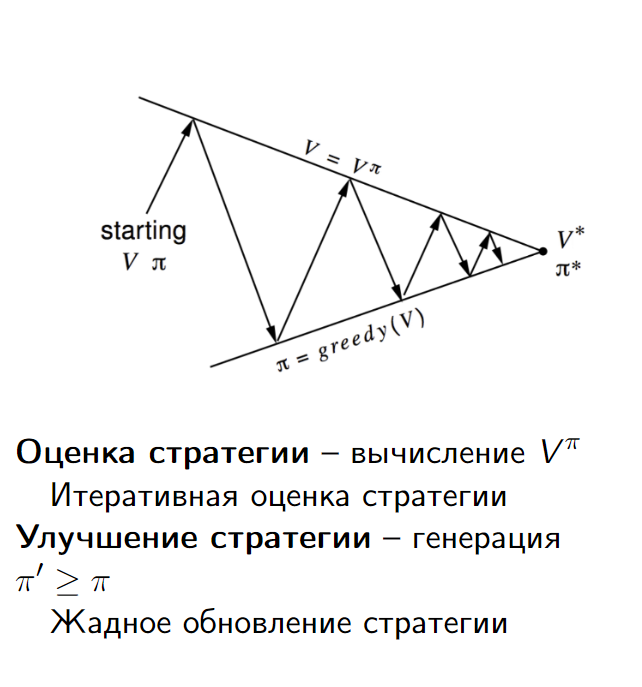
# 8. Метод временных различий (TD), TD показатель и TD ошибка.

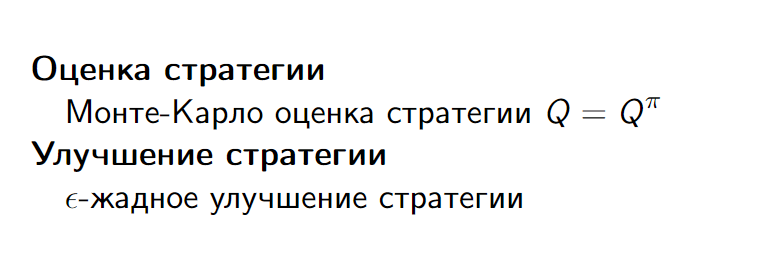


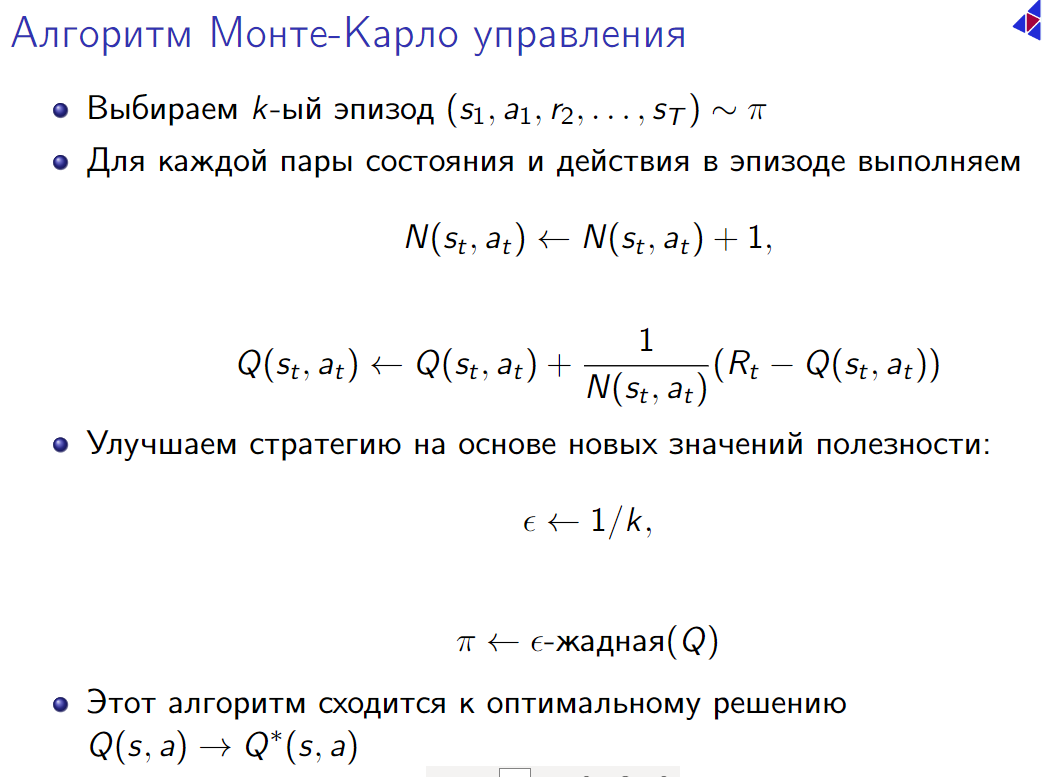
# 9. Состоятельность, несмещенность и дисперсия оценок в MC и TD.



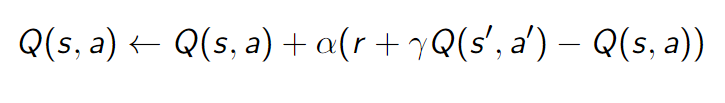
# 10. Обобщенные итерации по стратегиям. Алгоритм Монте-Карло управления. Алгоритм SARSA для управления.







Обновление функции полезности по SARSA

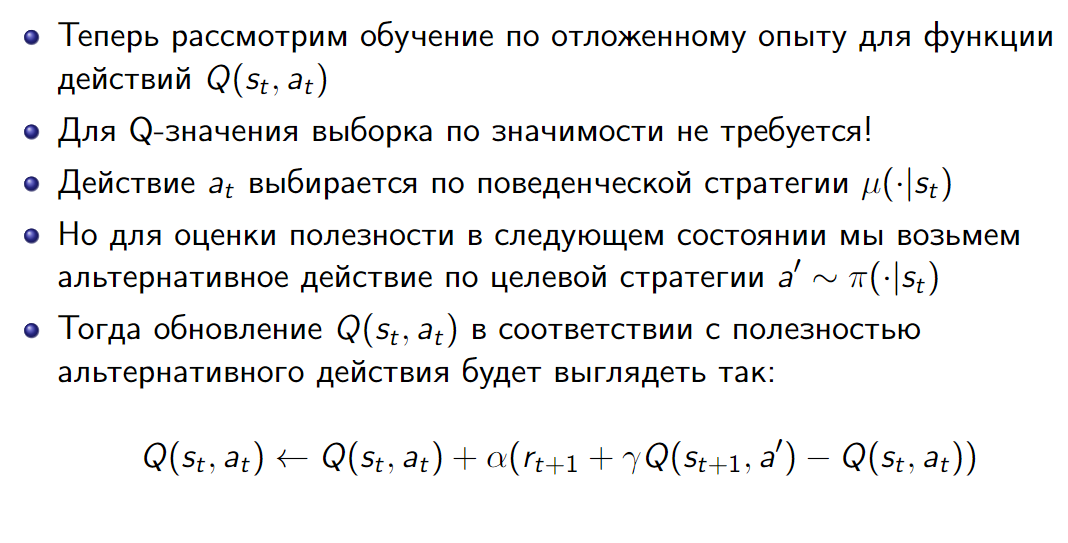


Управление по актуальному опыту

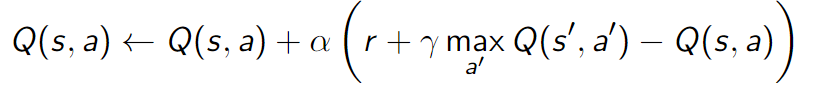
с SARSA



# 11. Обучение по собственному и чужому опыту. Алгоритм Q-обучения.



Алгоритм управления с Q-обучением (SARSAMAX)



# 12. Понятие обобщения, виды и возможные модели аппроксимации.

# 13. Линейный аппроксиматор с оценками MC и TD в задаче оценки стратегии. Линейный аппроксиматор для задачи управления.

# 14. Пакетные методы аппроксимации, пример на основе метода наименьших квадратов. Аппроксимация глубокими нейронными сетями.

# 15. DQN и его расширения: Double DQN, Prioritized Replay, Dueling DQN, Rainbow.

# 16. Назначение вознаграждений, подцели и временное абстрагирование. Полумарковский процесс принятия решений, иерархическое обучение с подкреплением.

# 17. Методы навыков (options framework). Иерархия абстрактных автоматов (HAMs).

# 18. Стохастические стратегии, преимущества и недостатки метода градиента стратегии. Функции полезности стратегии.

# 19. Результирующая функция стратегии. Теорема о градиенте стратегии. Алгоритм REINFORCE.

# 20. Метод базового уровня для уменьшения дисперсии градиента стратегии.

# 

# 

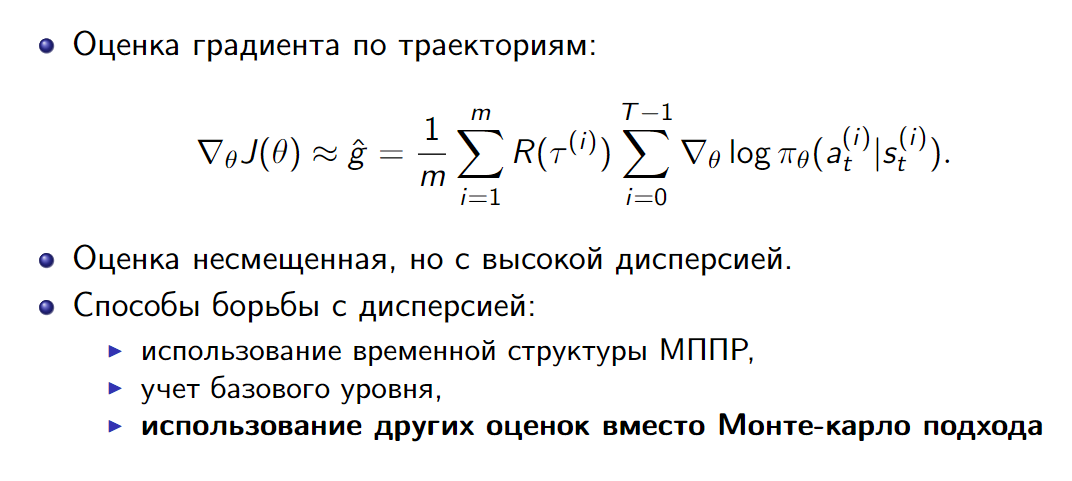
# 

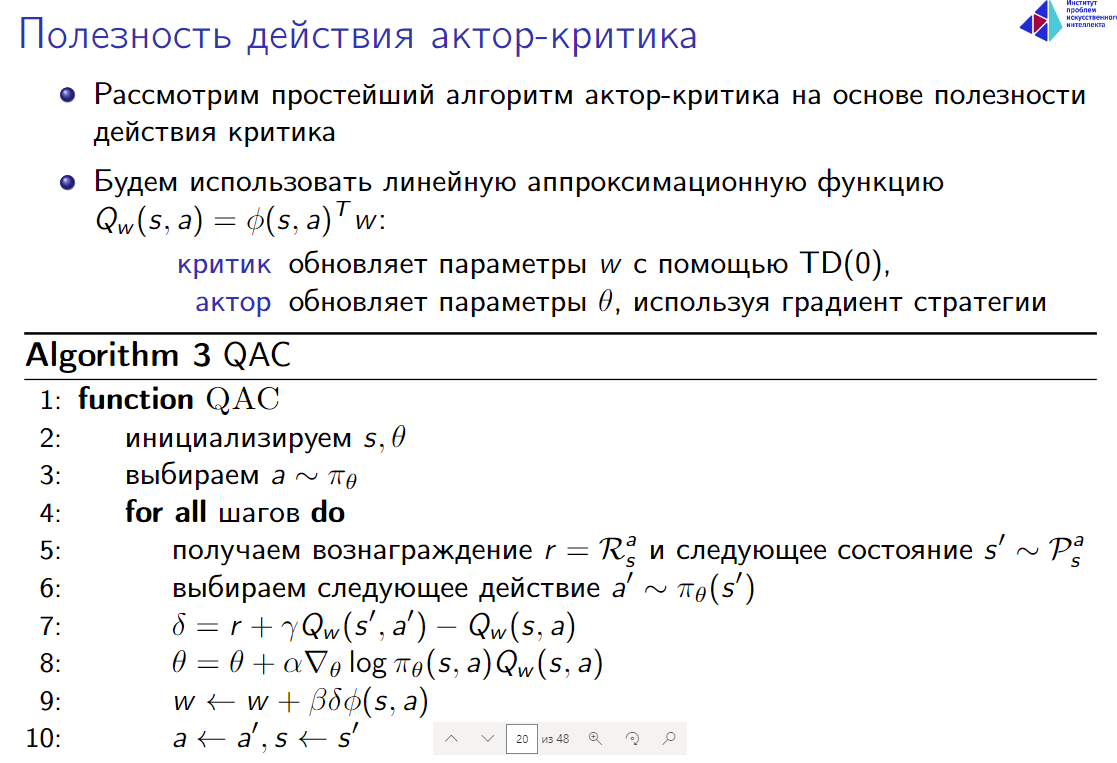
# 

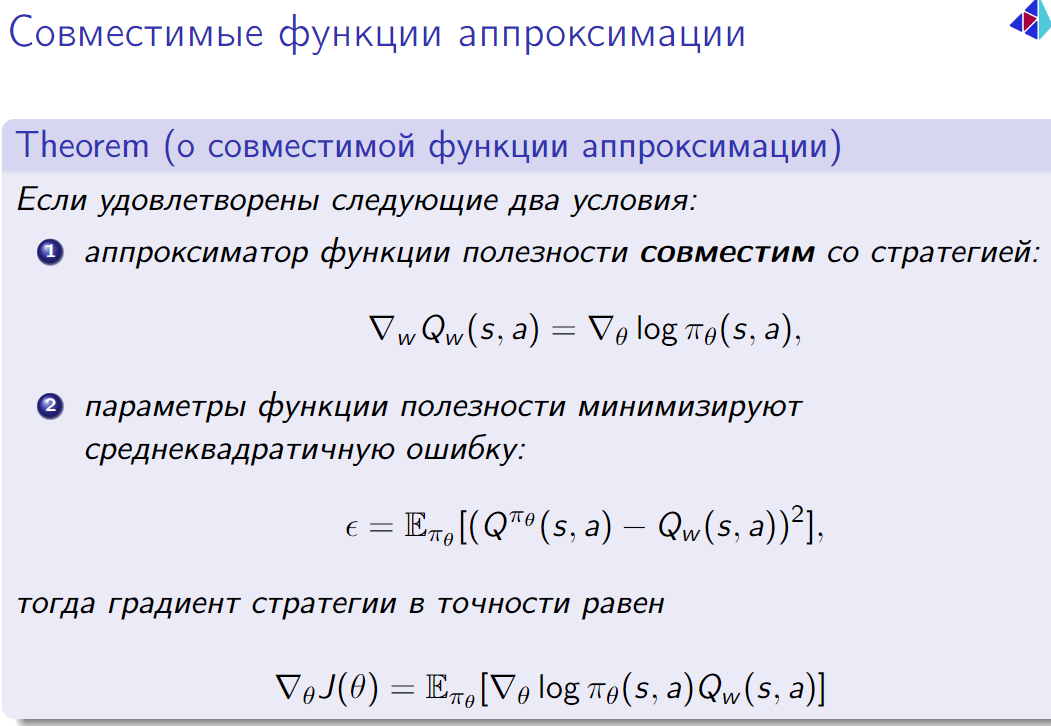
# 

# 

# 21. Оценка стратегии с помощью критика, совместимые аппроксимации функции полезности. - ЛЕКЦИЯ 6

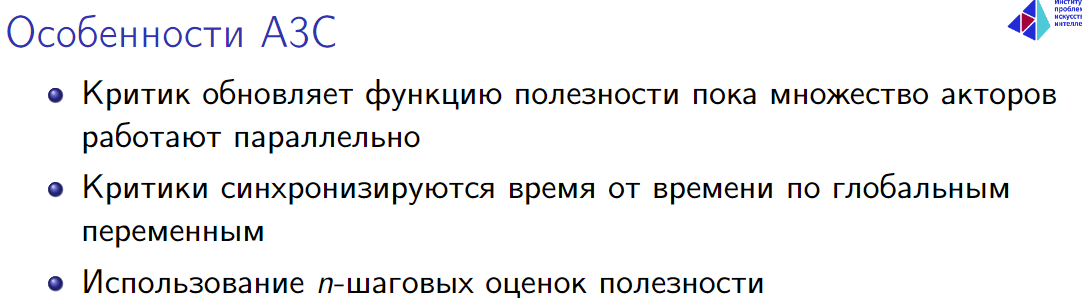


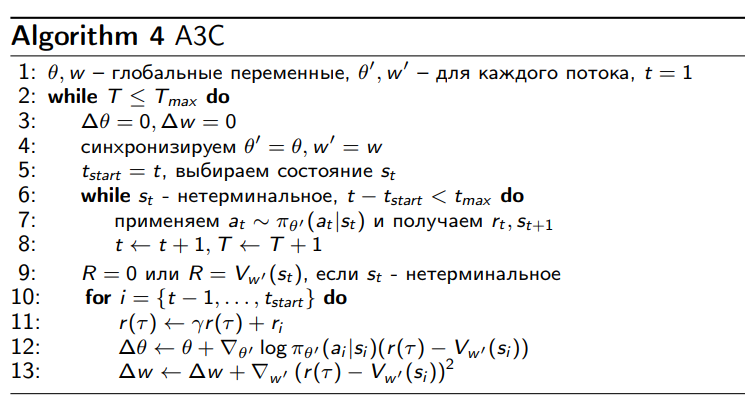


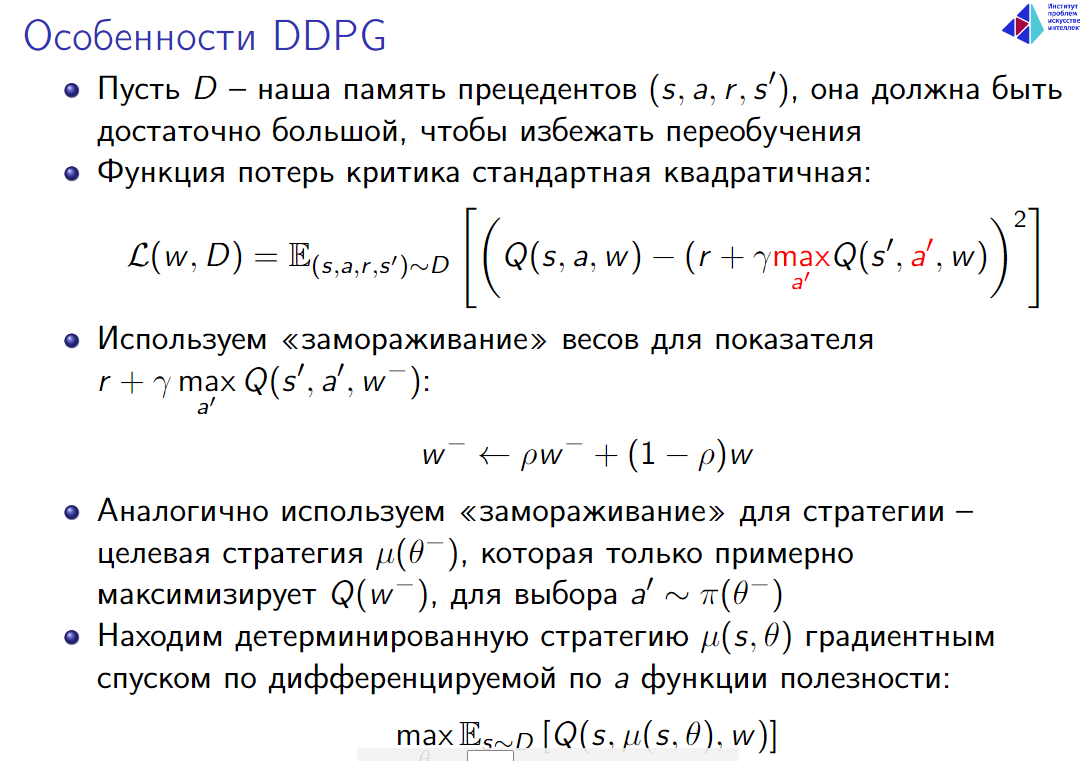


# 22. Способы оценки полезности траектории стратегии. Алгоритм Vanilla Policy Gradient (VPG).

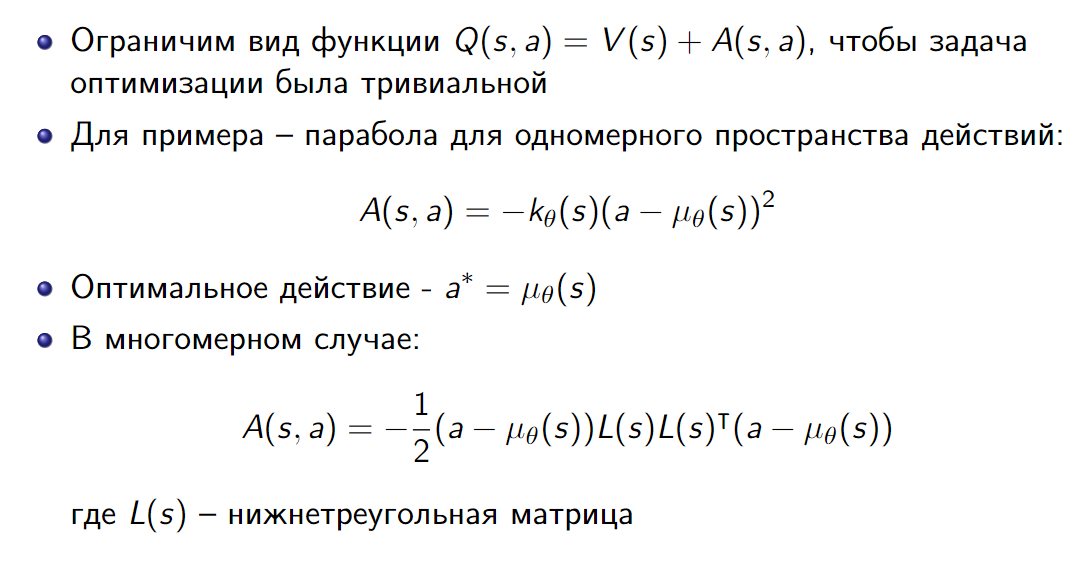
# 23. Алгоритм асинхронного актора-критика (A3C). Алгоритм глубокого градиента детерминированной стратегии (DDPG). Улучшение DDPG - TD3

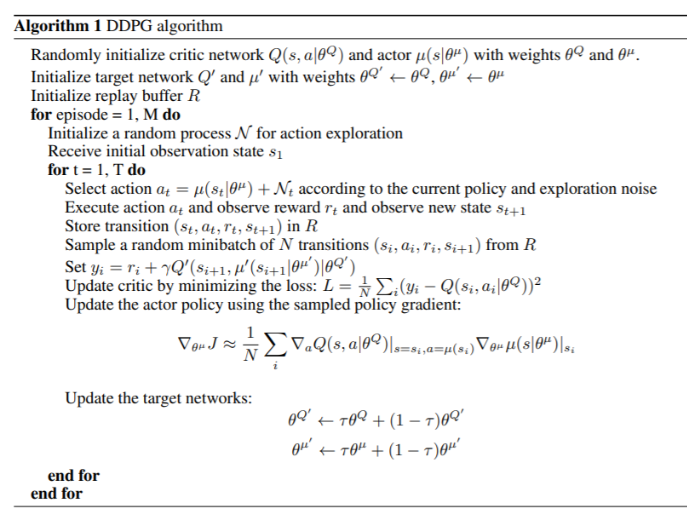




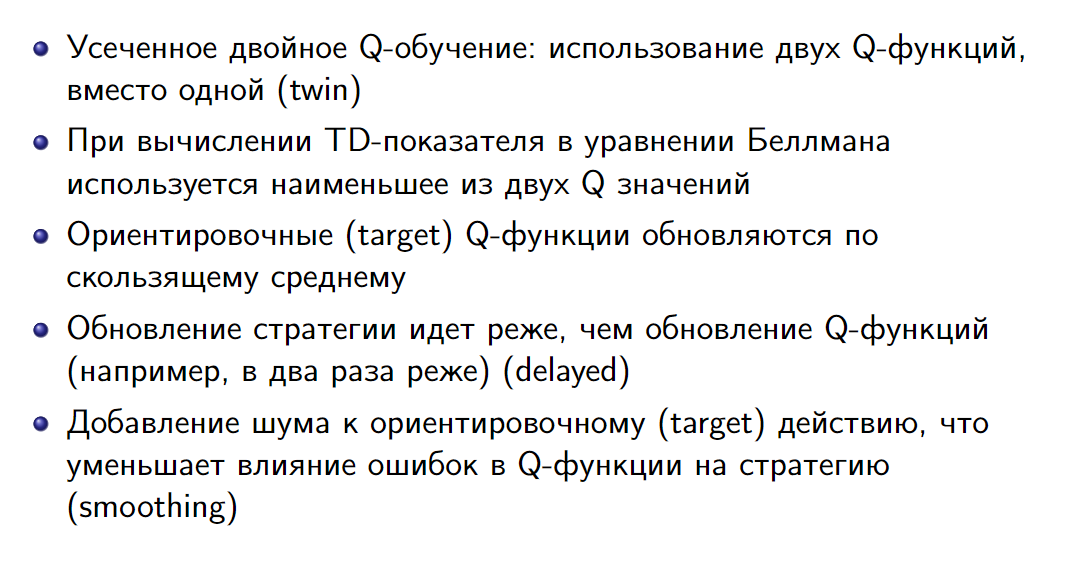


Нормализация функции преимущества

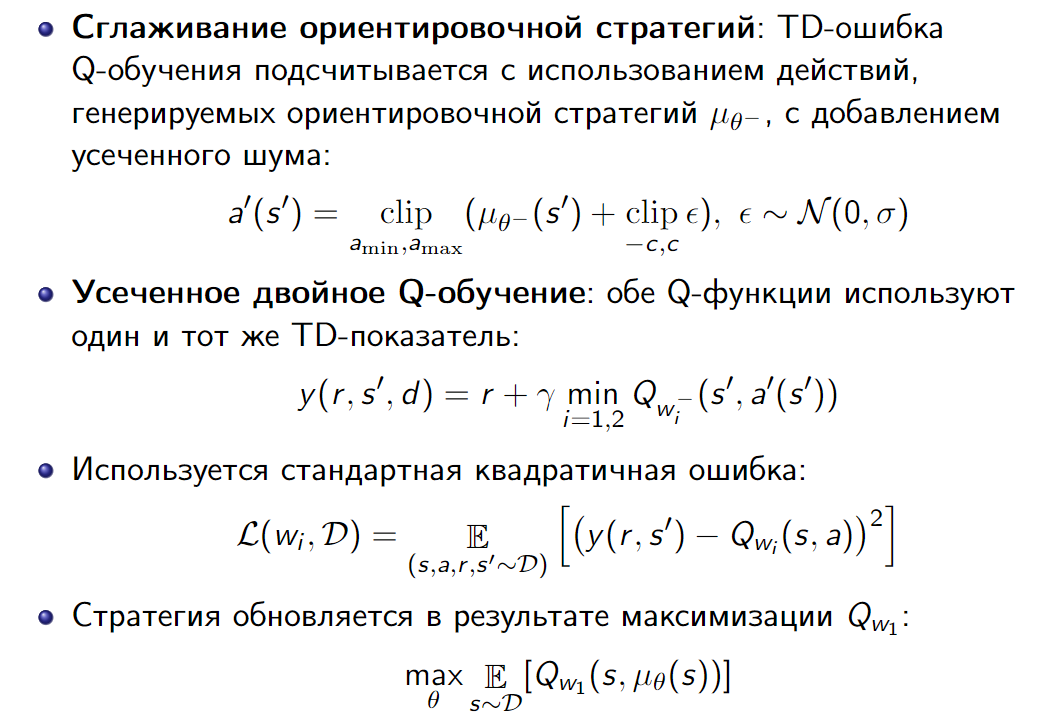




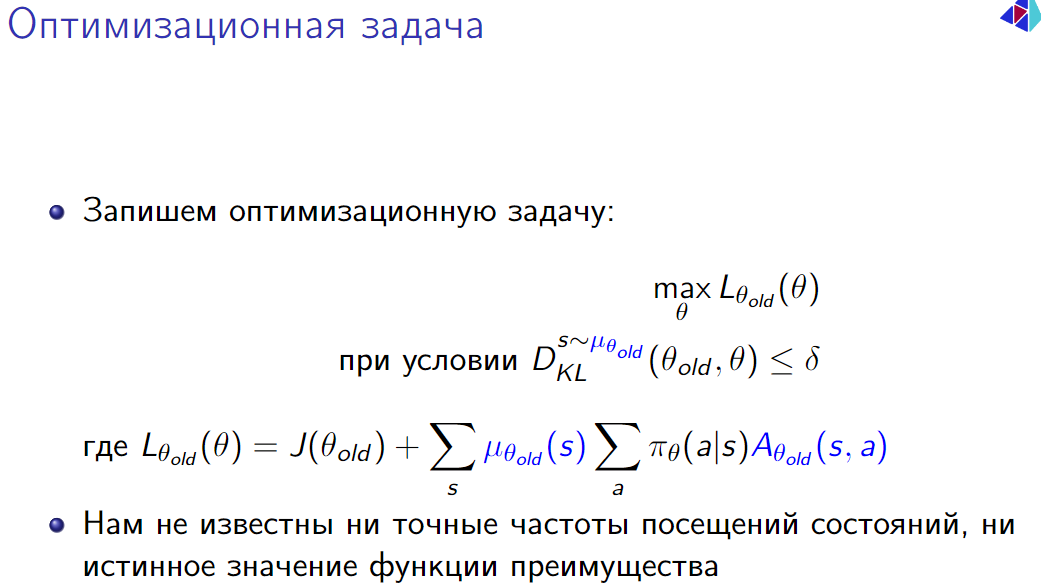
Основные особенности TD3

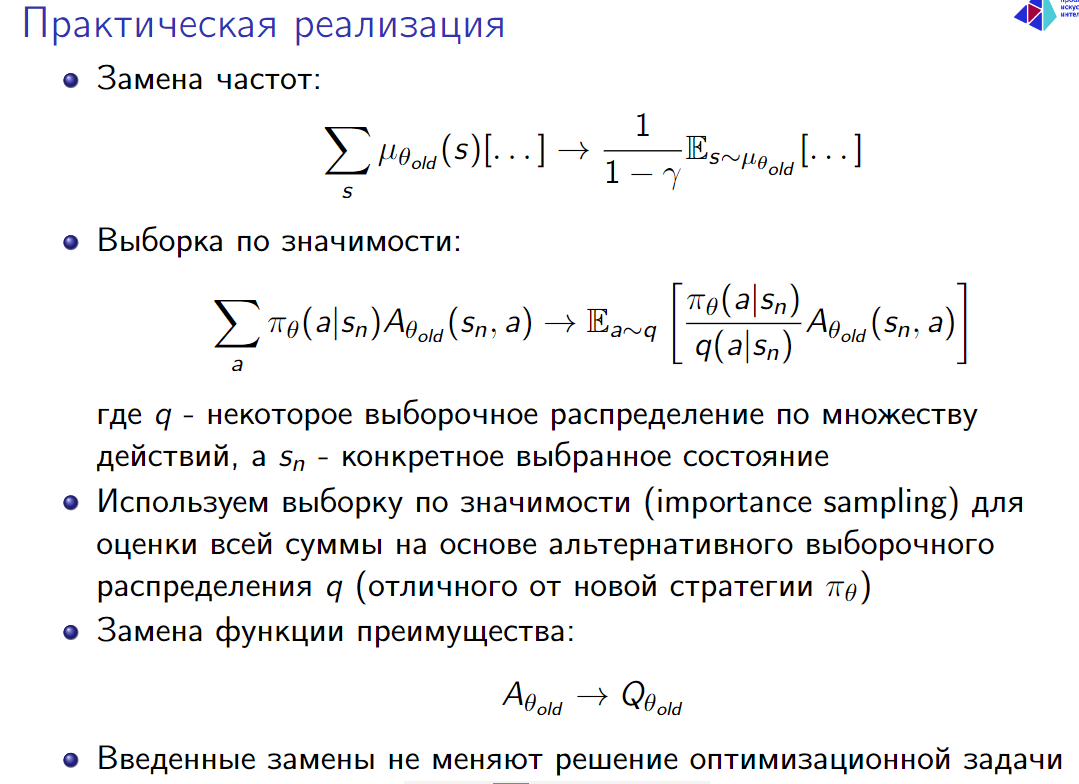


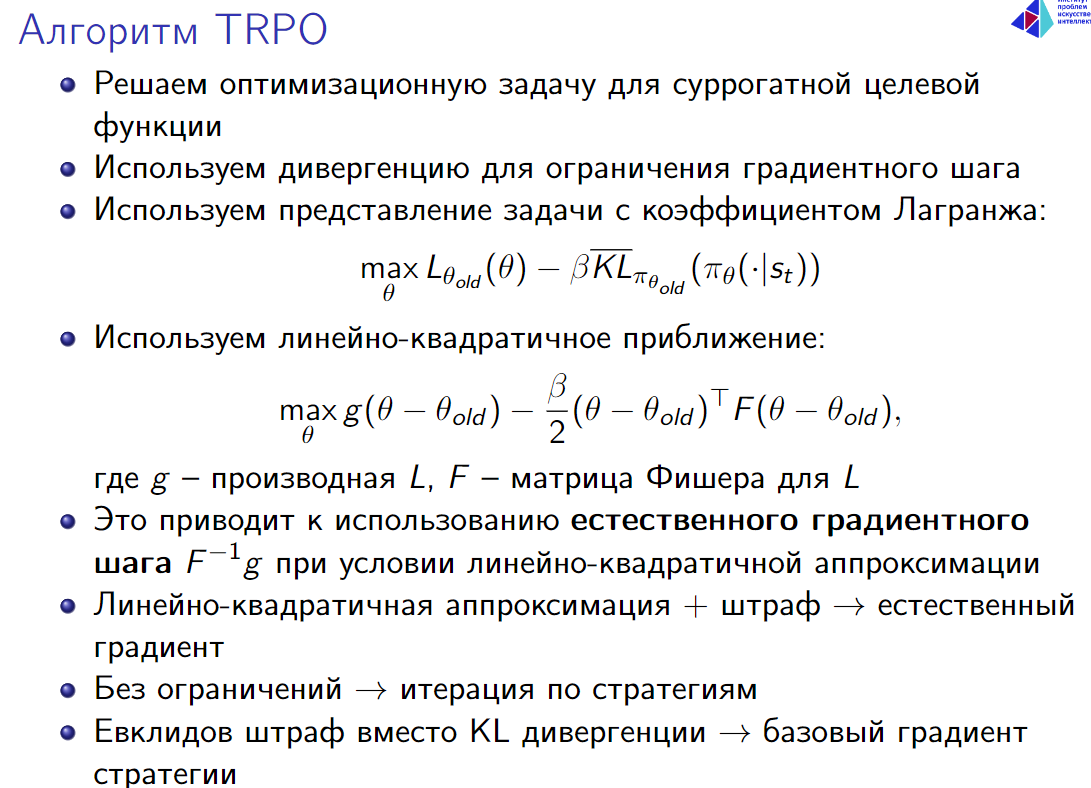
Реализация алгоритма TD3

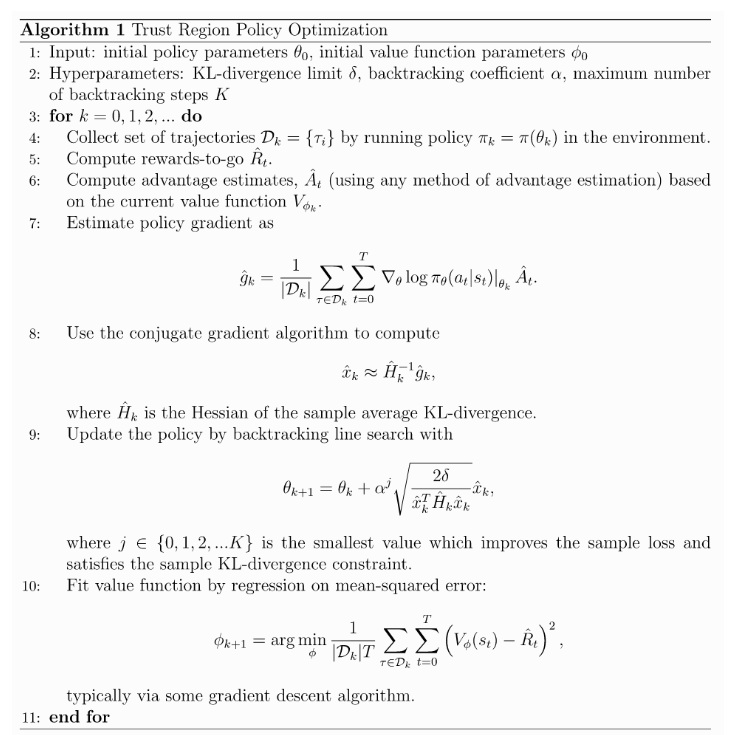


# 24. Адаптированное вычисление шага градиентного спуска для вычисления градиента стратегии. Алгоритм оптимизации стратегии по доверительным областям (TRPO).

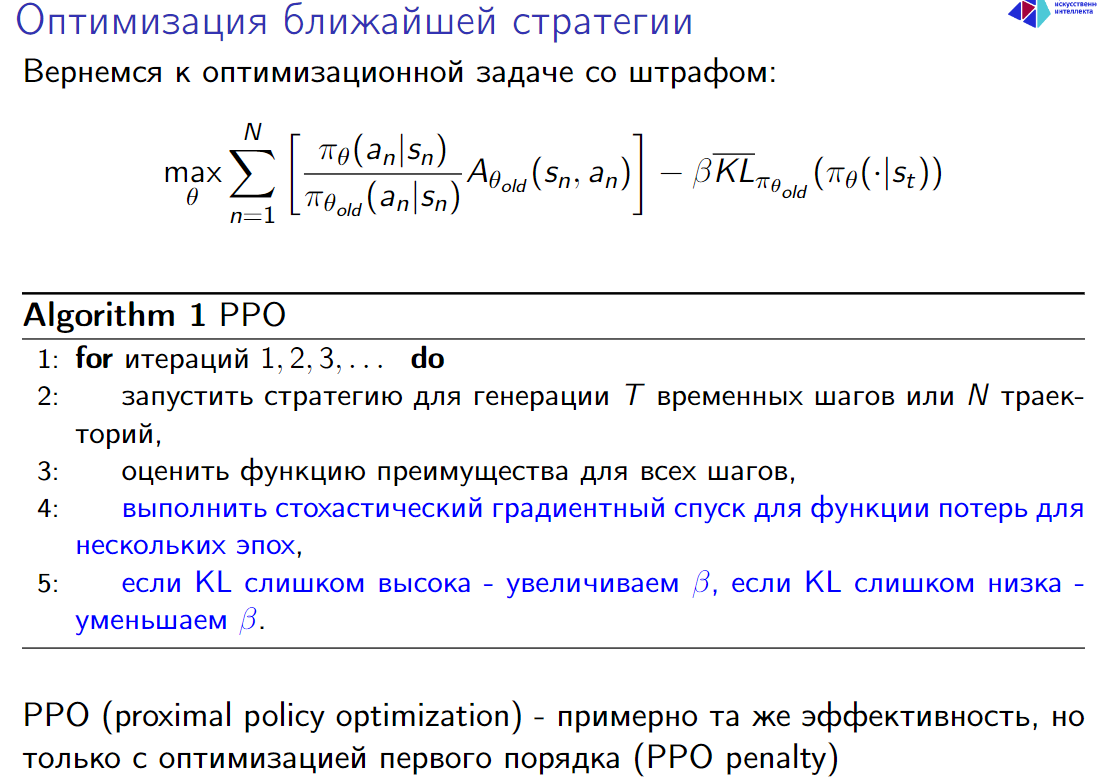


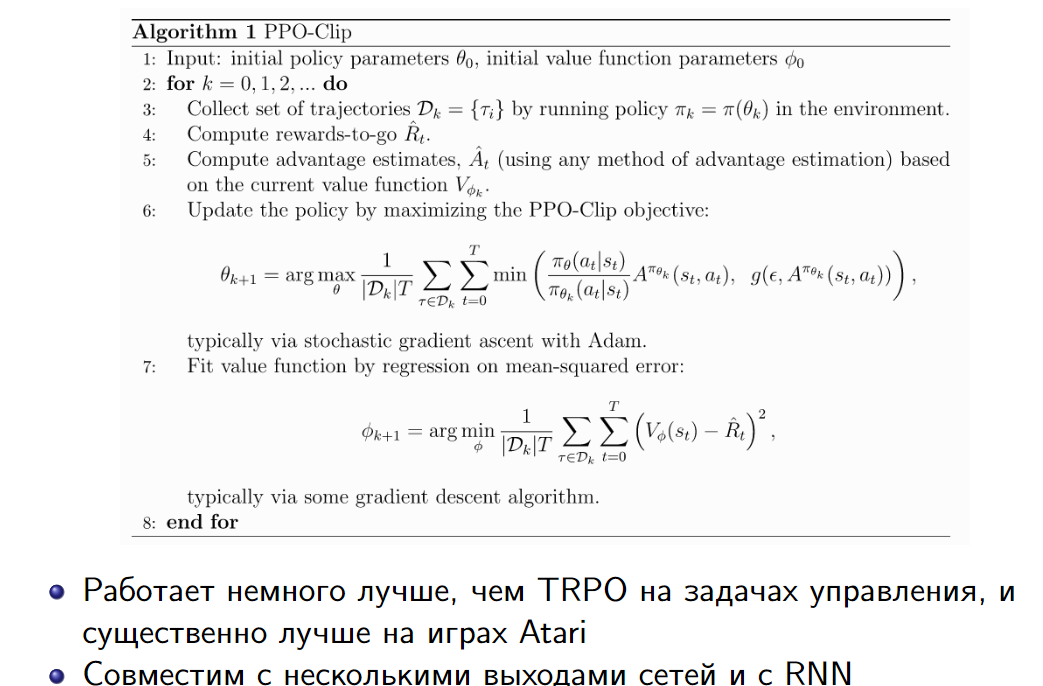


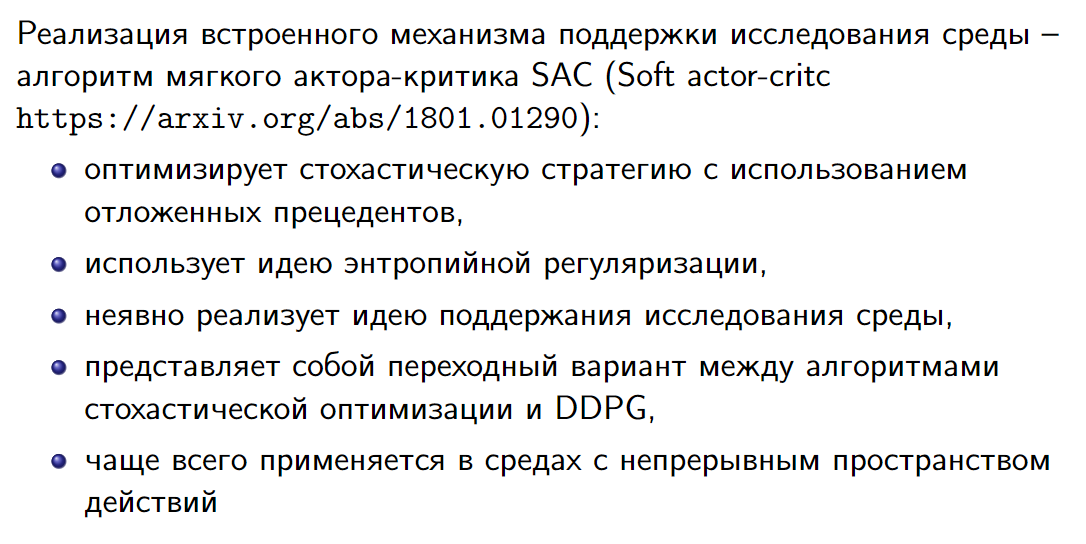


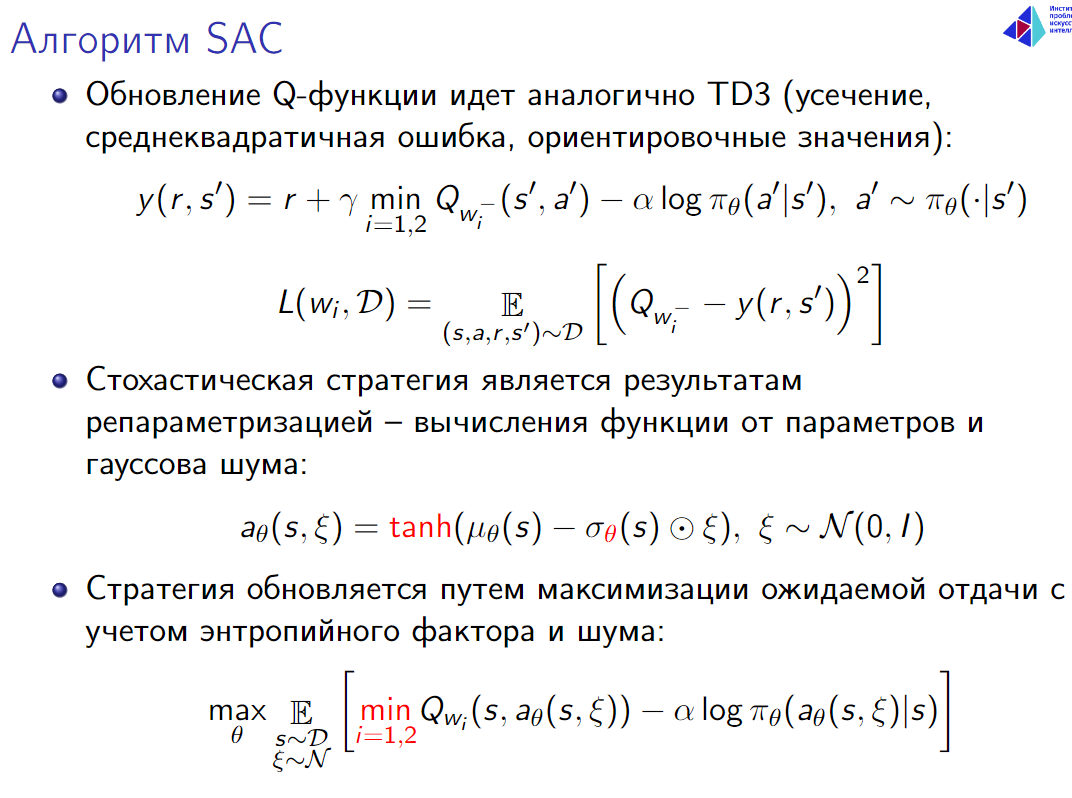


# 25. Алгоритм оптимизации ближайшей стратегии (PPO). Улучшение исследования среды в алгоритме SAC.

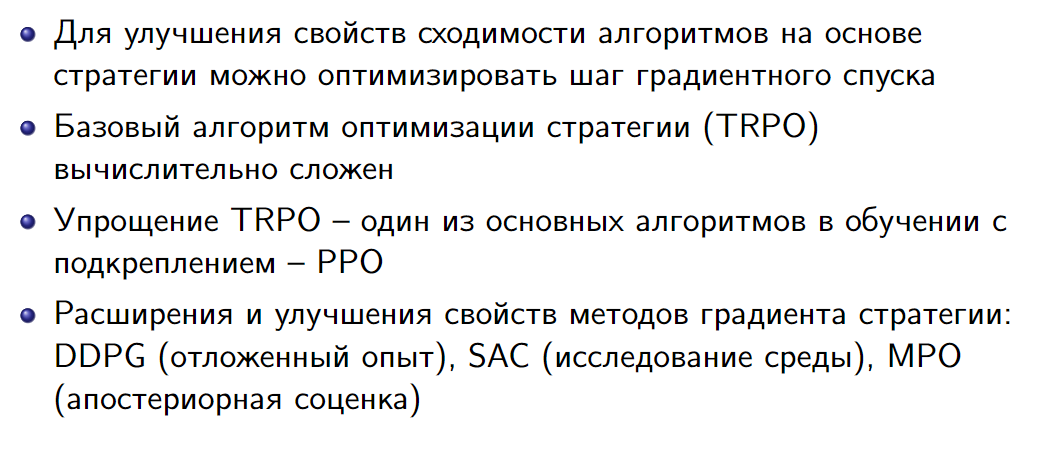






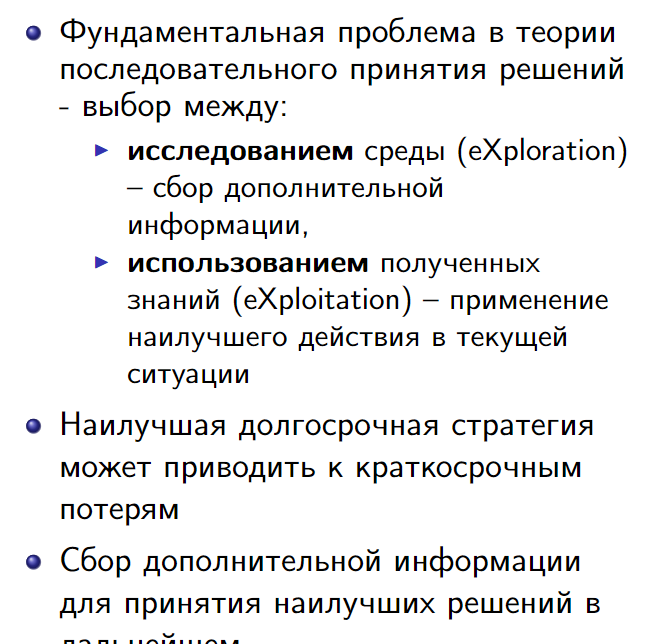


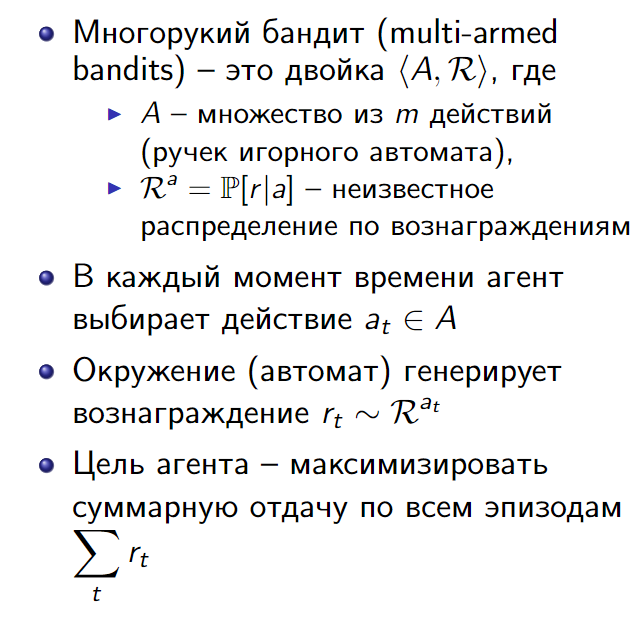
Выводы

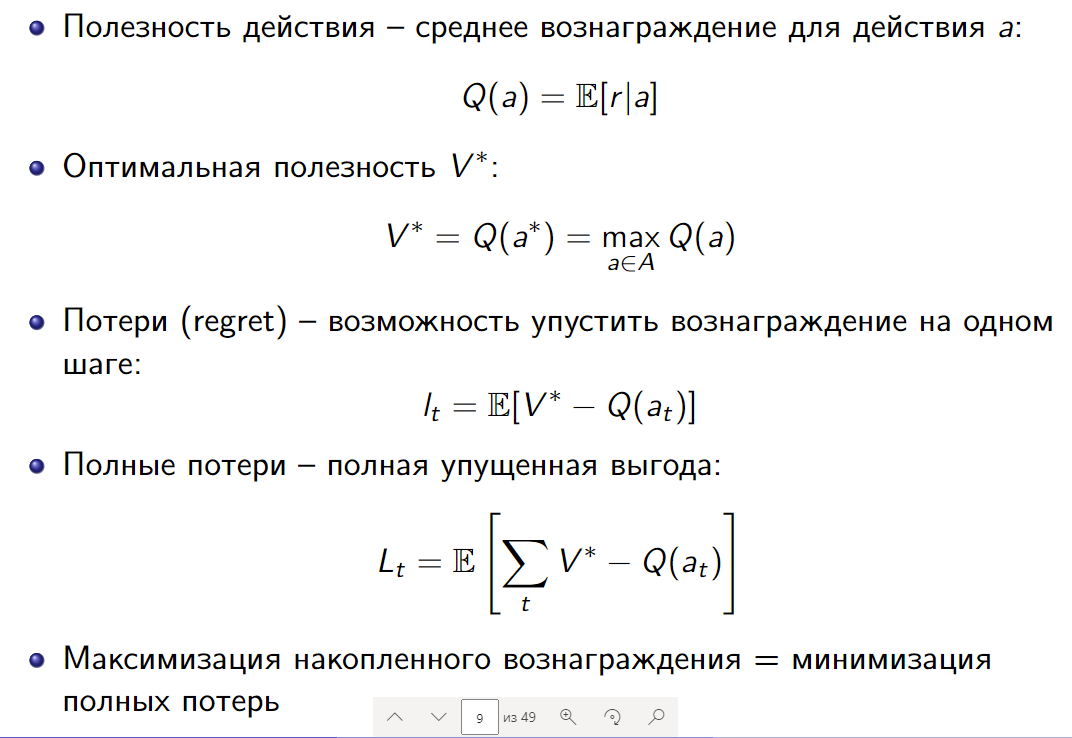


# 26. Многорукие бандиты. Наивный подход к исследованию среды. Оптимизм в неопределённости.

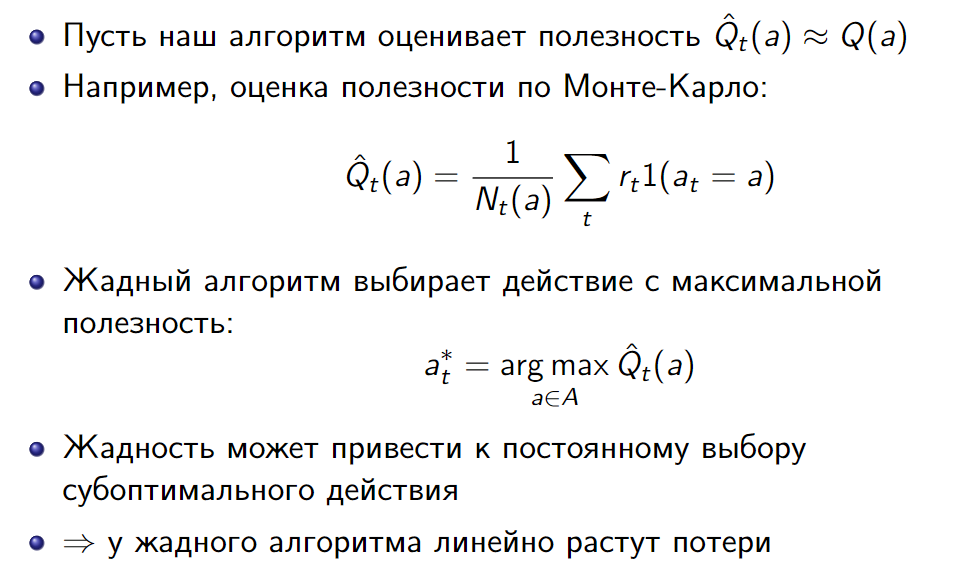
Задача

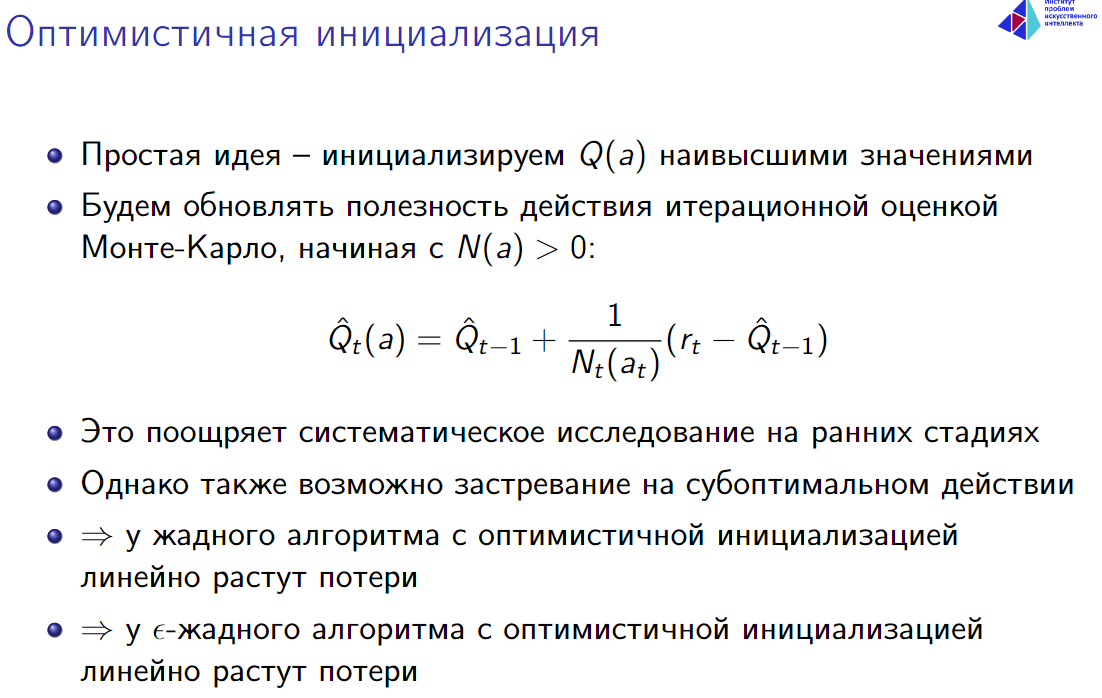


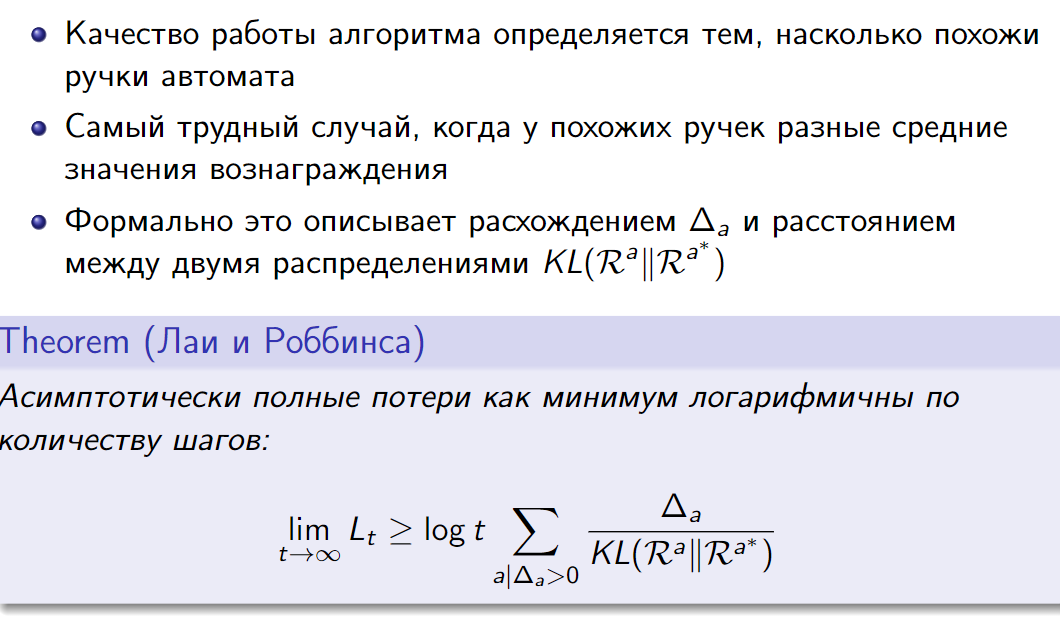


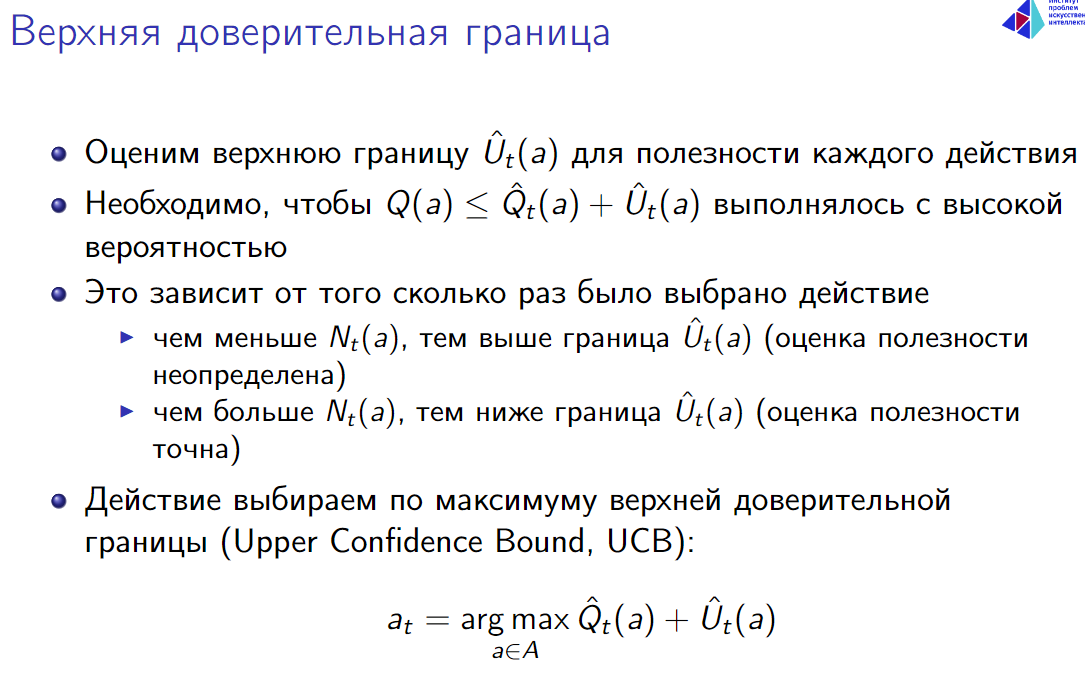


Жадный алгоритм



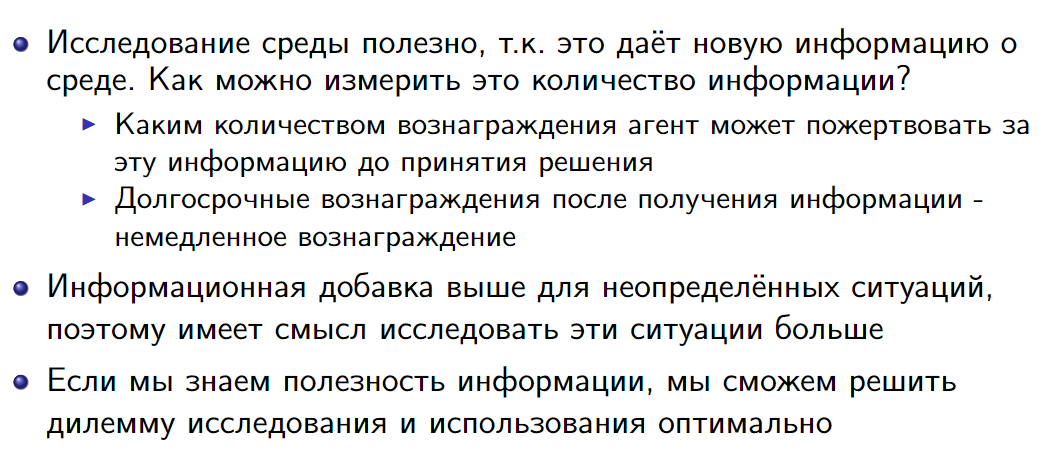




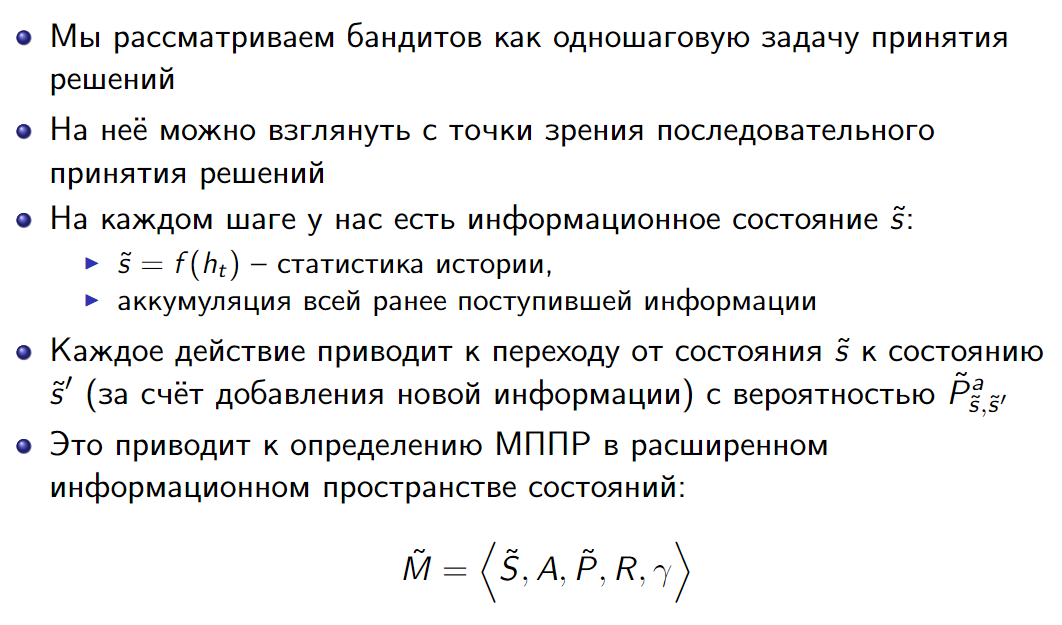


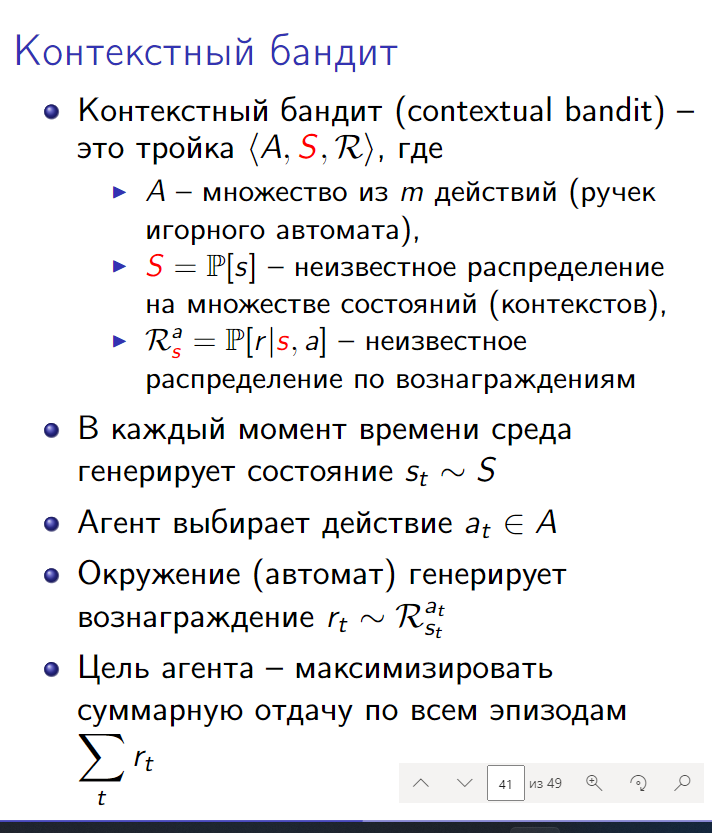
# 27. Несоответствие вероятностей в задаче исследования среды. Поиск в информационном пространстве. Контекстные бандиты

Полезность



Пространство информационных состояний





# 28. Обучение с подкреплением, основанное на модели. Интеграция на уровне эпизодов (Dyna).

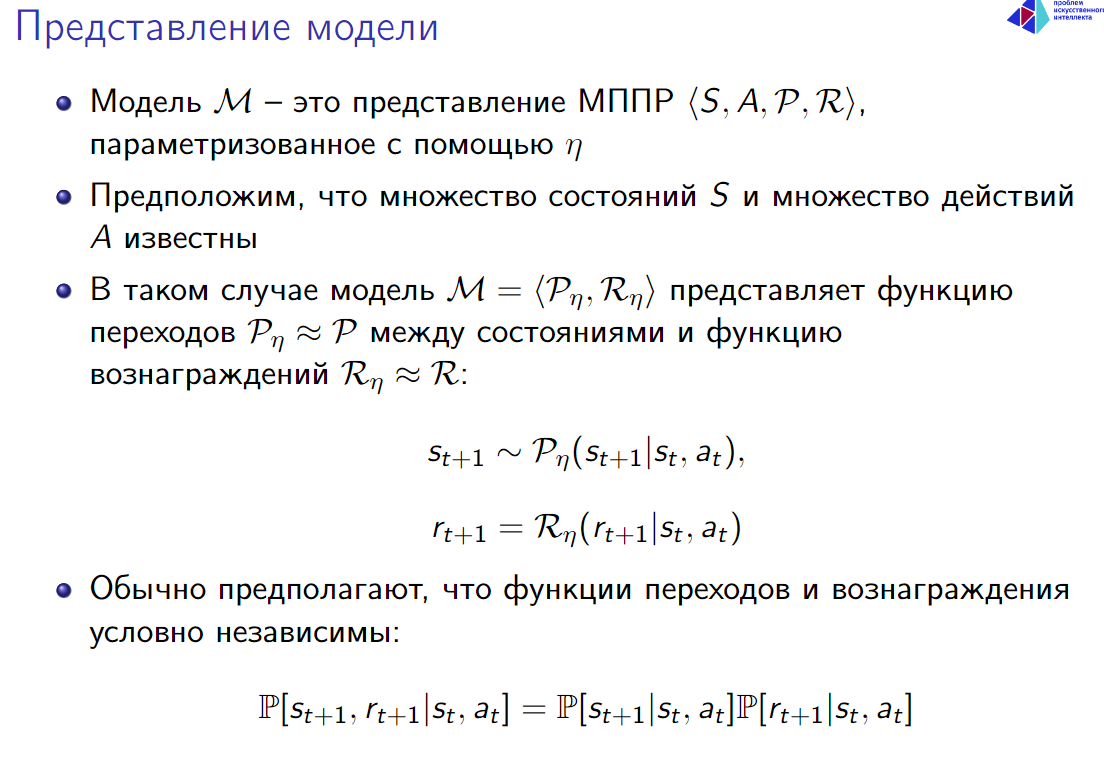
В этой лекции мы рассмотрим построение модели по опыту и

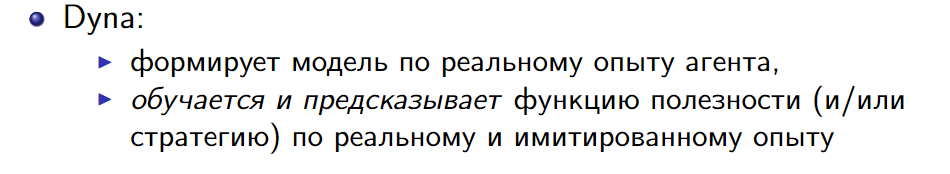
использование планирования для построения функции полезности или стратегии

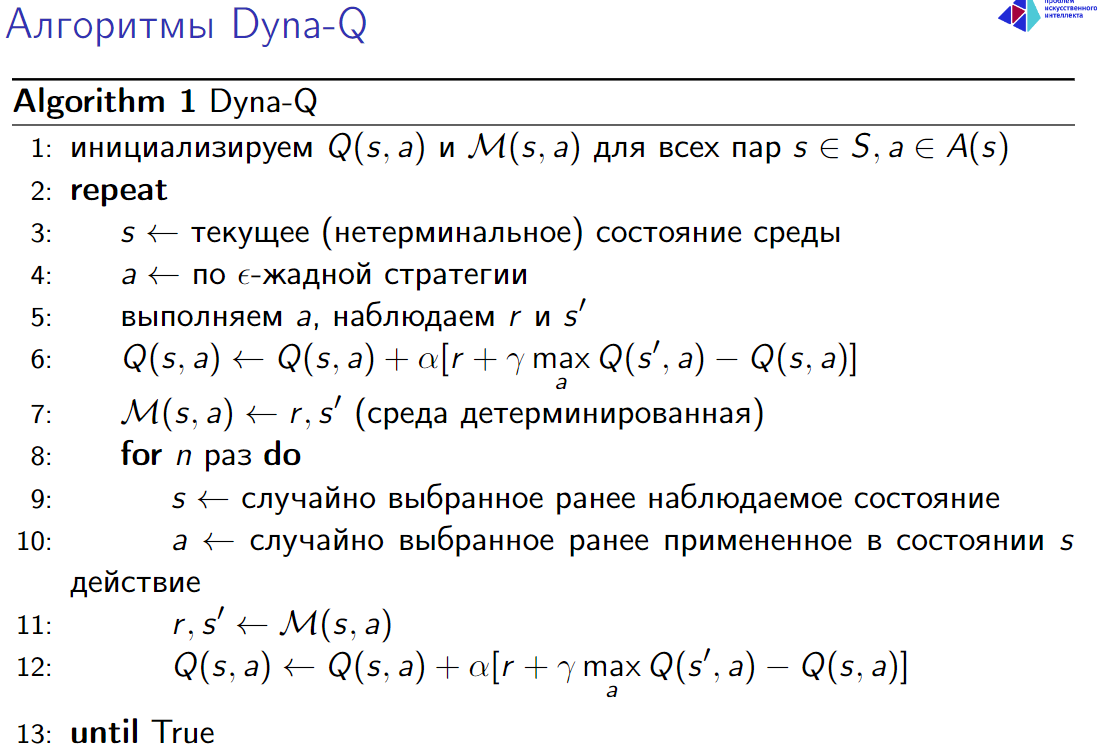
Наша задача – объединить обучение с подкреплением и

планирование в единую архитектуру

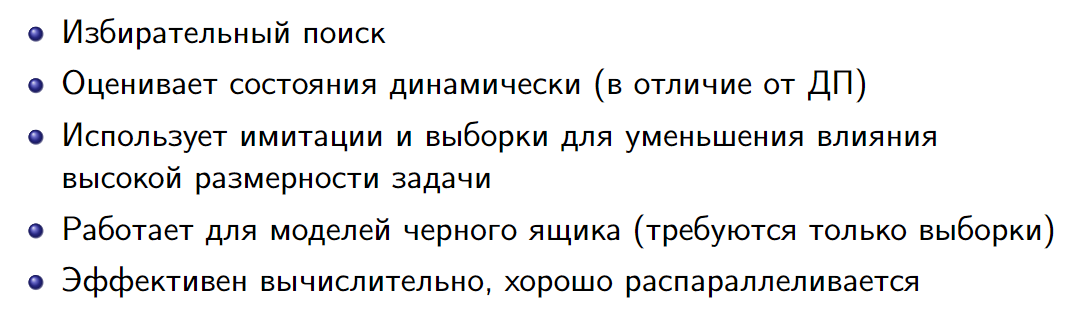


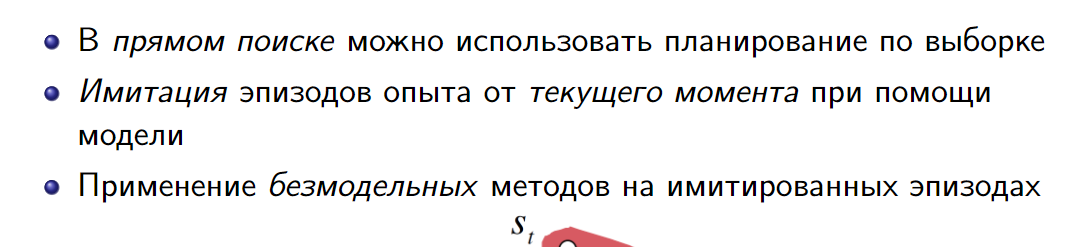


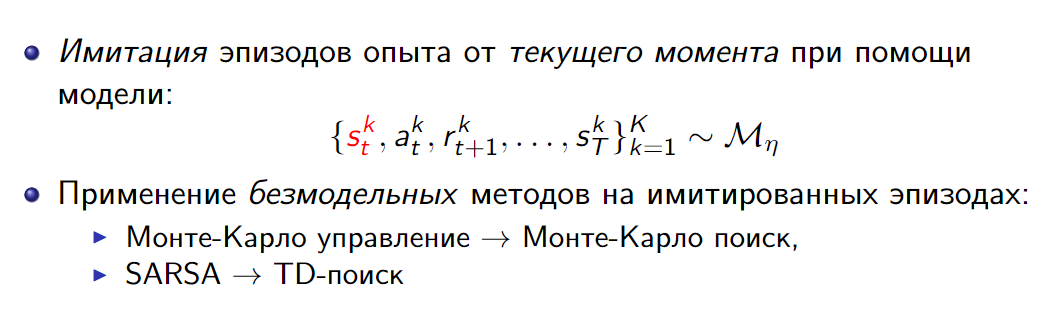


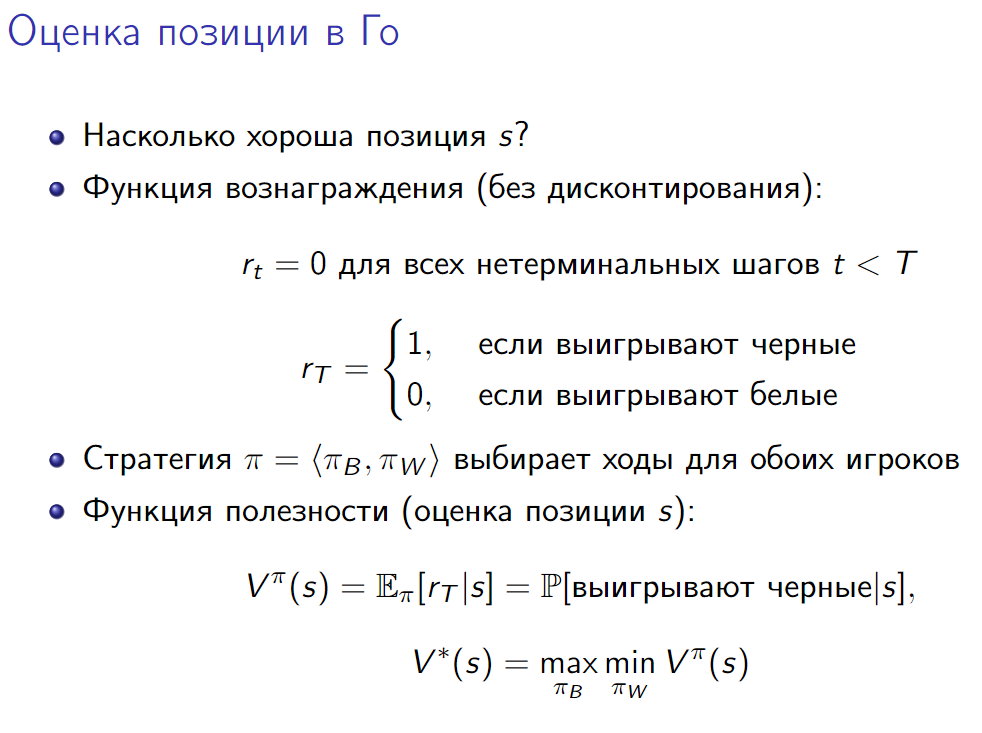


# 29. Имитационные поиск по дереву (MCTS). Применение MCTS в игре Го. Обновление модели в MuZero.









# 30. Внутренняя мотивация. Определение новизны, алгоритм RND. - ЛЕКЦИЯ 11 И СЕМИНАР

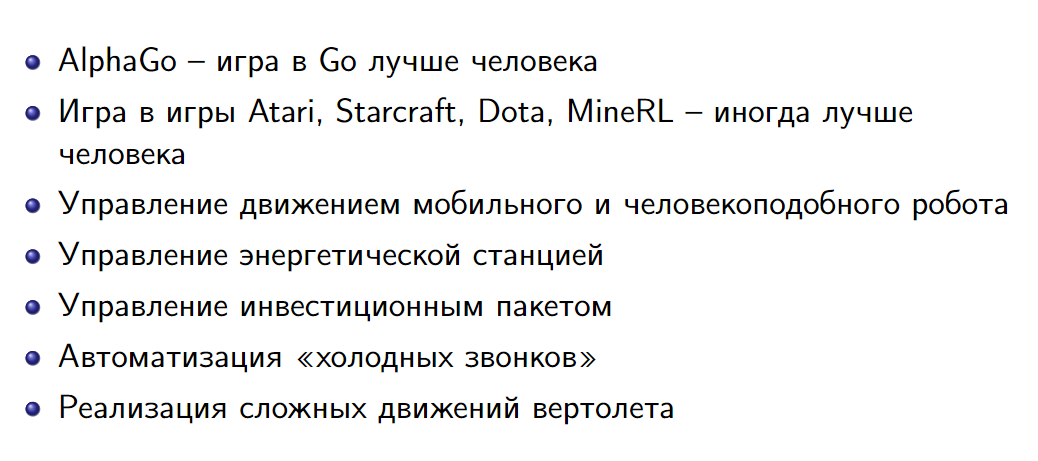
# 31. Любопытство при исследовании среды. Алгоритмы ICM и Go-explore.

# 32. Копирование поведения. Обратное обучение с подкреплением. Подражательное обучение.

# 33. Глубокое Q-обучение на основе демонстраций (DQfD). Генеративное соревновательное имитационное обучение (GAIL). Мягкое имитационное Q-обучение (SQIL).

# 34. Автономное обучение с подкреплением. Пакетное обучение с выборкой по значимости. Неявное Q-обучение. Консервативное Q-обучение.

# 36. Актуальные проблемы в обучении с подкреплением. Перспективные направления в обучении с подкреплением. Области применения методов обучения с подкреплением

****