**2.2 Общая постановка задачи обучения с подкреплением**

**2.2.1 Значение оценочной обратной связи**

Как уже было сказано, цель обучения с подкреплением заключается в максимизации выгоды для интеллектуального агента. Здесь мы можем рассматривать две задачи: прогнозирование и управление.

В задаче прогнозирования обучение с подкреплением используется только для вычисления некой политики (стратегии), которая описывает для каждого посещаемого состояния ожидаемую в будущем награду. [8]

Задача управления ставит перед собой поиск набора стратегий, который максимизирует награду в режиме реального времени, т.е. во время переходов от одного состояния к другому. [8]

Для поиска оптимального маршрута используются алгоритмы, упомянутые в п. 2.1. При этом существует два вида нейрональных архитектур для прогнозирования и контроля: с обратной связью и без обратной связи. [8]

Классический пример обучения с подкреплением без обратной связи – собака Павлова. [42] В данном случае собака является интеллектуальным агентом, пища – безусловным раздражителем, звонок – условным. Условный раздражитель предсказывает безусловный, после обучения у агента появляется реакция на условный раздражитель. В данном случае нет обратной связи, т.к. реакция на условный раздражитель не влияет на представление стимулов. [8]

В случае существования обратной связи архитектура обучения представляет собой замкнутую схему. Замкнутость необходима для решения задачи управления, т.к. действия агента влияют на свои собственные входные значения. Эта схема изображена на рисунке 1. Точка X0 является сигналом обратной связи. [8, 35]

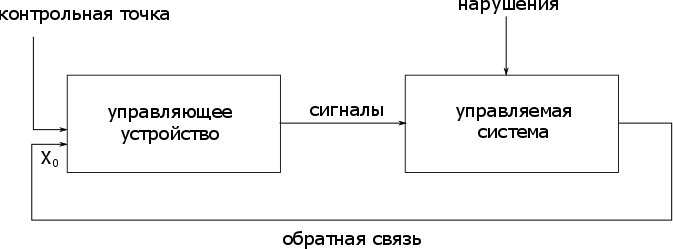


Рисунок 1 – Архитектура обучения с подкреплением с обратной связью

На рисунке 2 показана эта же архитектура, расширенная до так называемой модели Актёр-Критик (Actor-Critic). Эта модель получила такое название благодаря аналогии с актёром, который выбирает действие, и с критиком, который эти действия критикует. Критик представляет собой функцию оценочной стоимости. [8, 35]

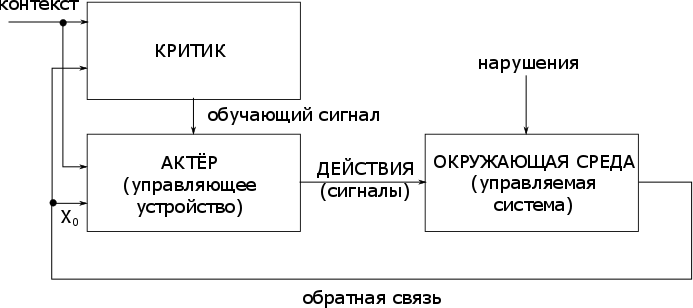


Рисунок 2 – Архитектура Актёр-Критик

Таким образом, обучение с подкреплением предполагает взаимодействие между некой средой и некой обучающейся системой, то есть наличие обратной связи.

**2.2.2 Постановка задачи**

Задачу обучения с подкреплением можно привести к математической модели. Рассмотрим, как она будет выглядеть в общем виде.

Пусть S – множество всех принимаемых состояний, и оно конечно.

В процессе обучения с подкреплением агент взаимодействует со средой. Другими словами, происходит игра между ними. Ниже приведен порядок действий в этой игре.

1. Инициализация стратегии и состояния среды
2. Для каждого момента времени
   1. Агент выбирает действие
   2. Среда генерирует награду
   3. Среда генерирует новое состояние
   4. Агент корректирует стратегию [43]

Эта игра будет называться марковским процессом принятия решений, если вероятность получения нового состояния и награды в нем зависит от текущего состояния и действия, совершенного агентом. [44]

Так как нам важна максимизация в долгосрочном периоде, то это требуется учесть.

Суммарная награда:

В общем случае суммарная награда является дисконтированной, т.е. формула (1) принимает вид:

где – коэффициент дисконтирования. Этот коэффициент был введен для того, чтобы обеспечить конечность суммы будущих наград, но он так же показывает, что люди и животные предпочитают ранние награды поздним. [29] Другими словами, чем больше тем агент будет более дальновидным.

Если – математическое ожидание при условии, что агент придерживается стратегии π, то [43]:

Функция ценности состояния при стратегии :

Функция ценности действия в состоянии при стратегии :

Таким образом, задача состоит в том, чтобы найти такую стратегию , при которой суммарная награда будет максимальна.