Введение в обучение с подкреплением

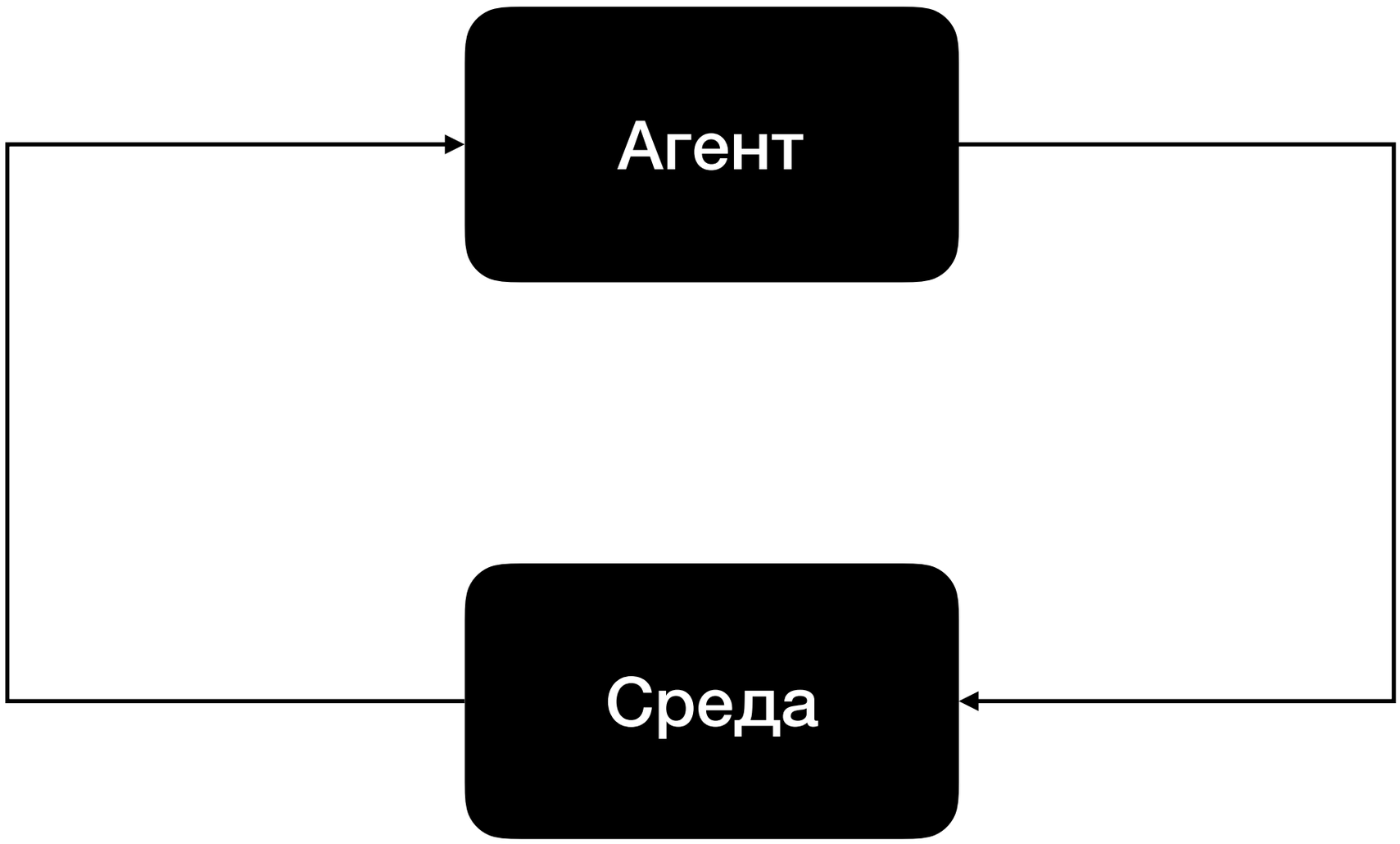
Левина Александра, БПМИ171
03.02.2020

Вдохновение

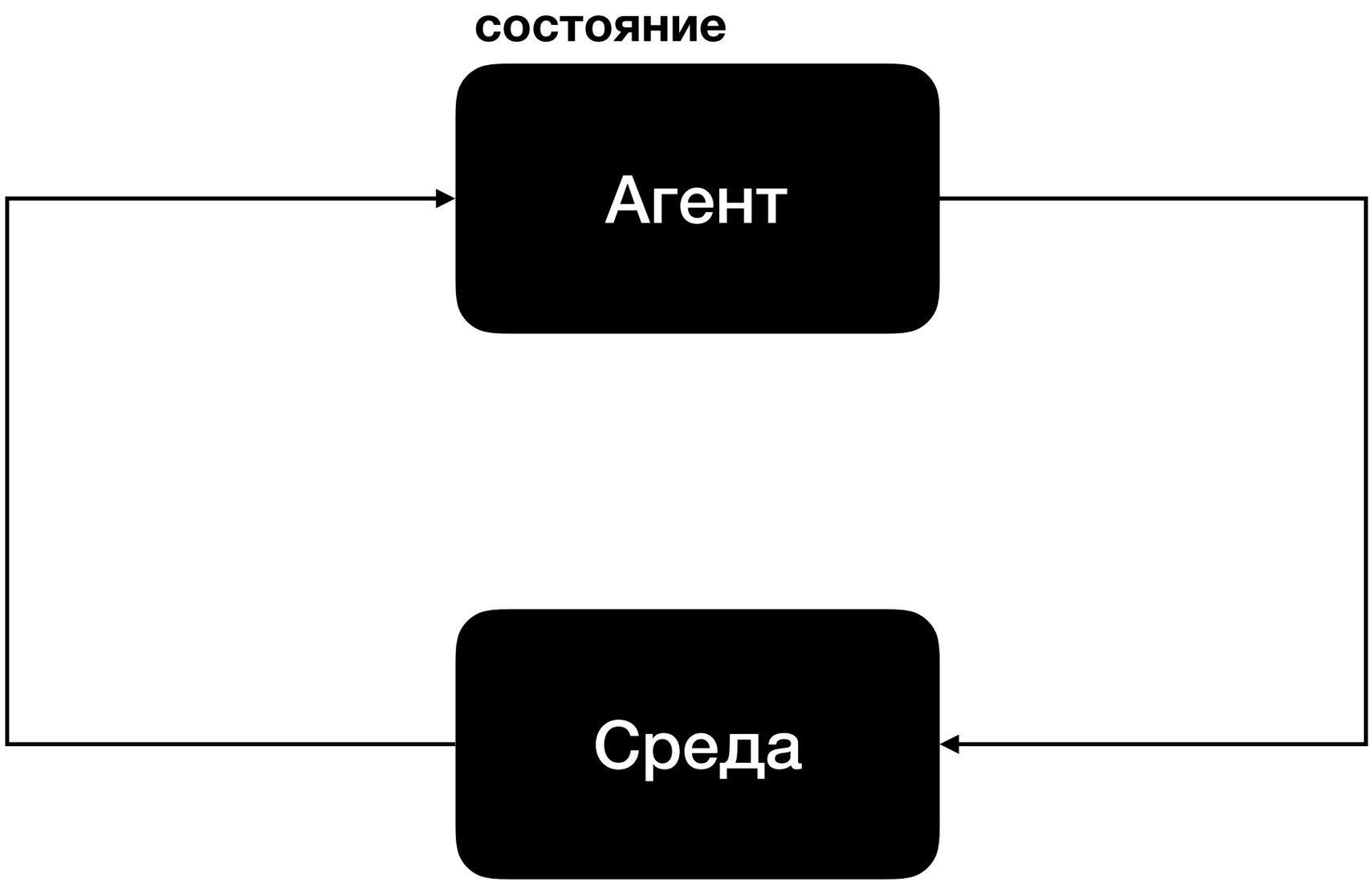
учитель — окружающий мир

учитель — дрессировщик

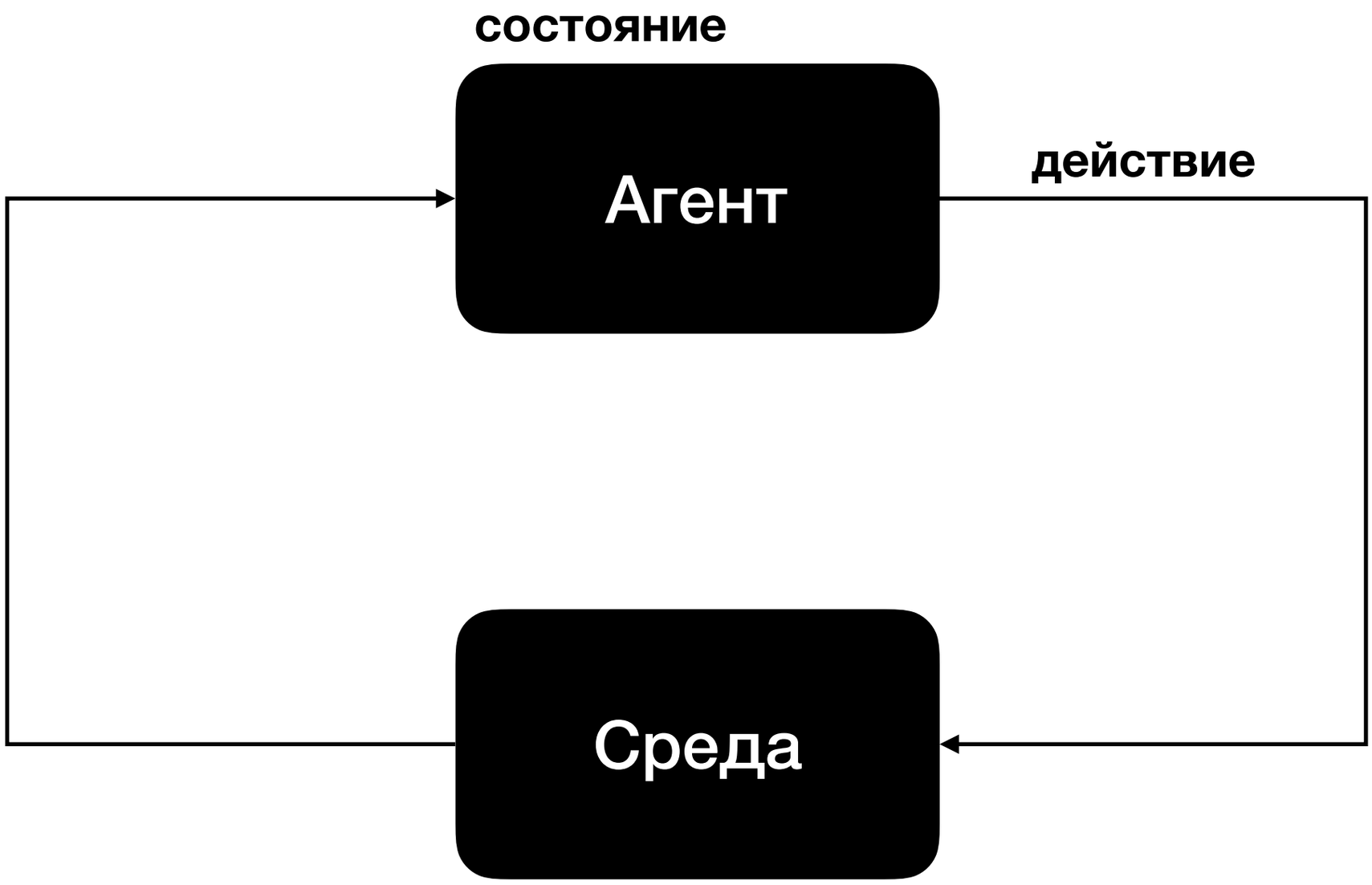
Постановка задачи



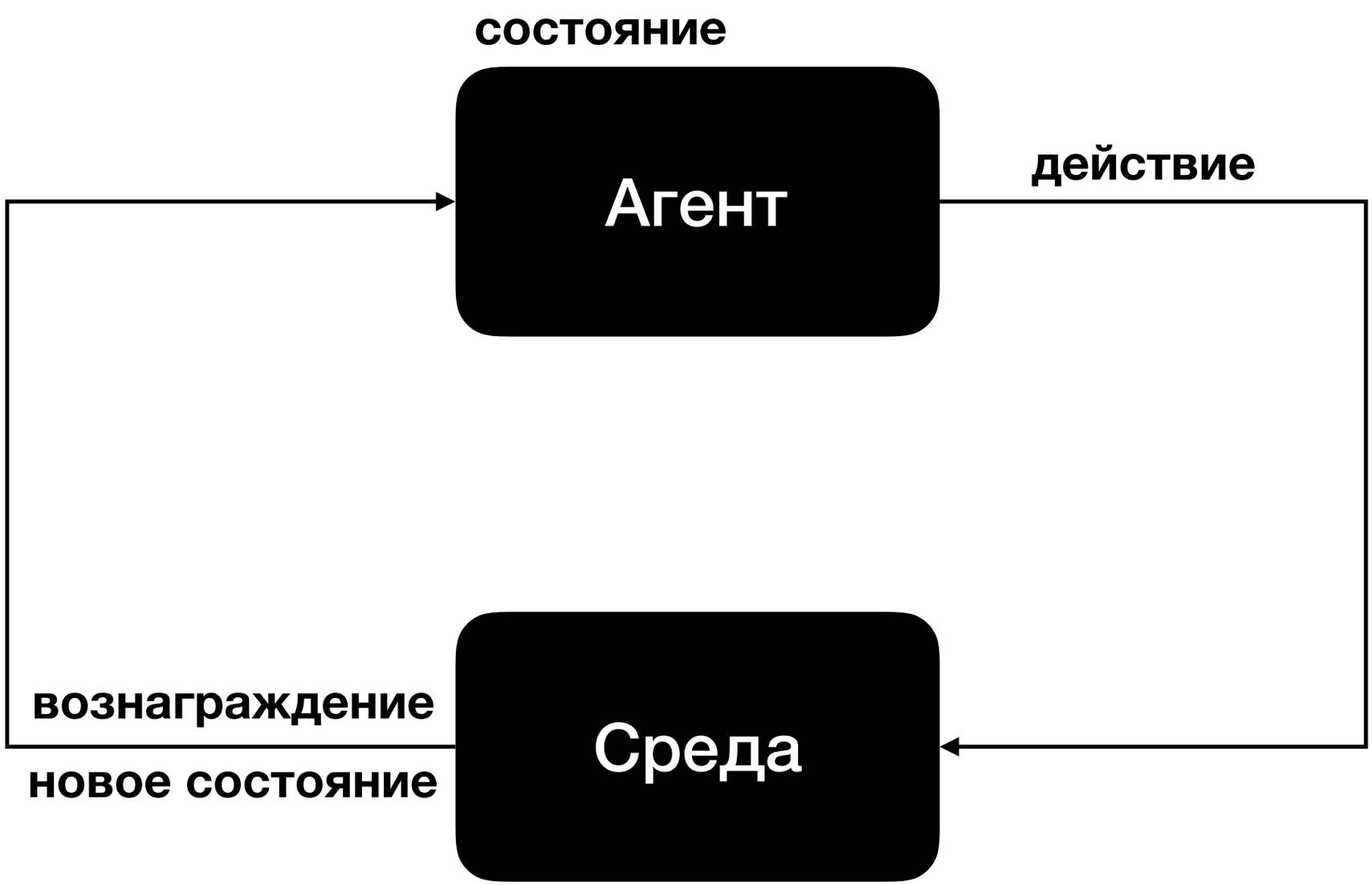
Постановка задачи



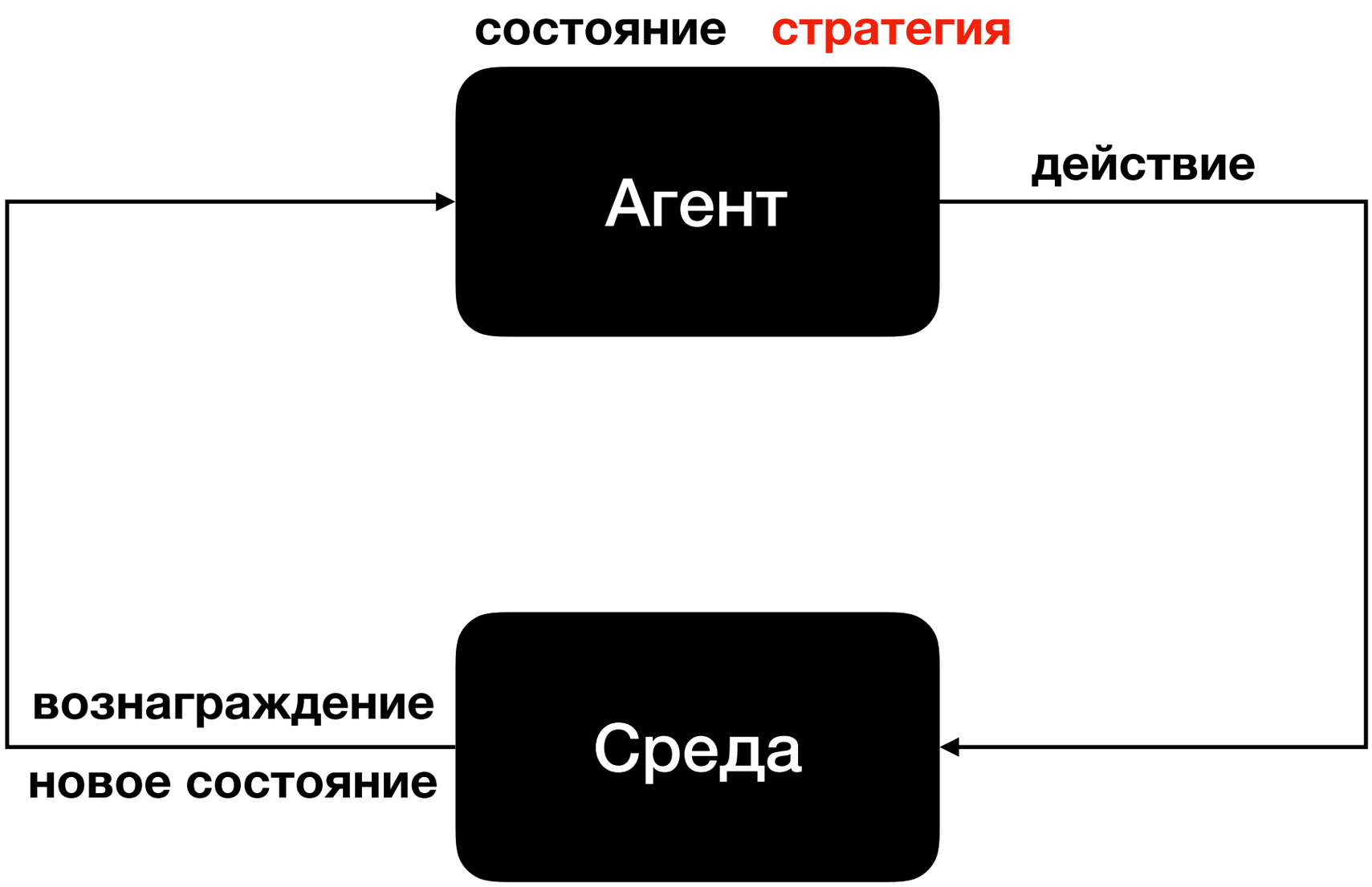
Постановка задачи



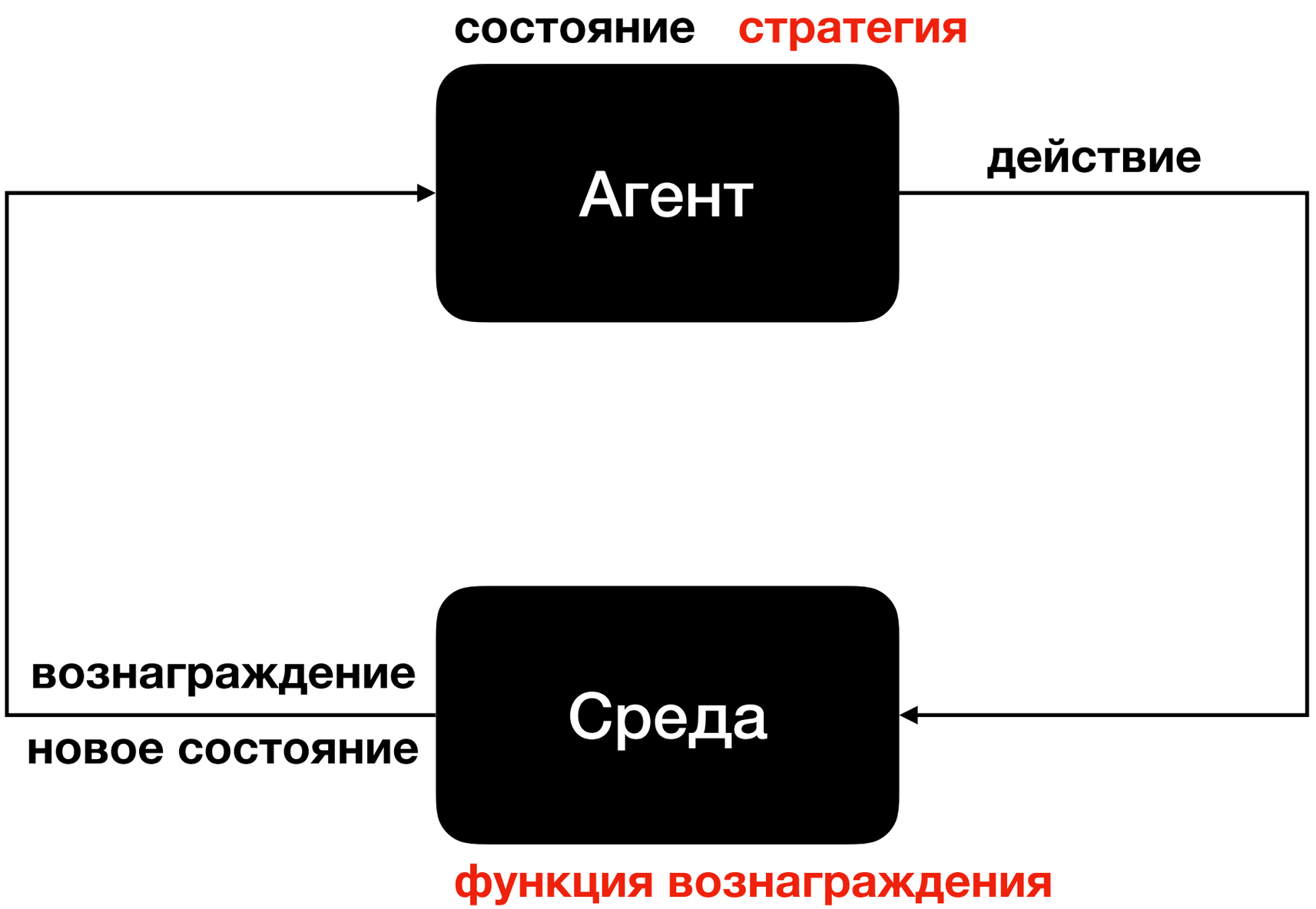
Постановка задачи



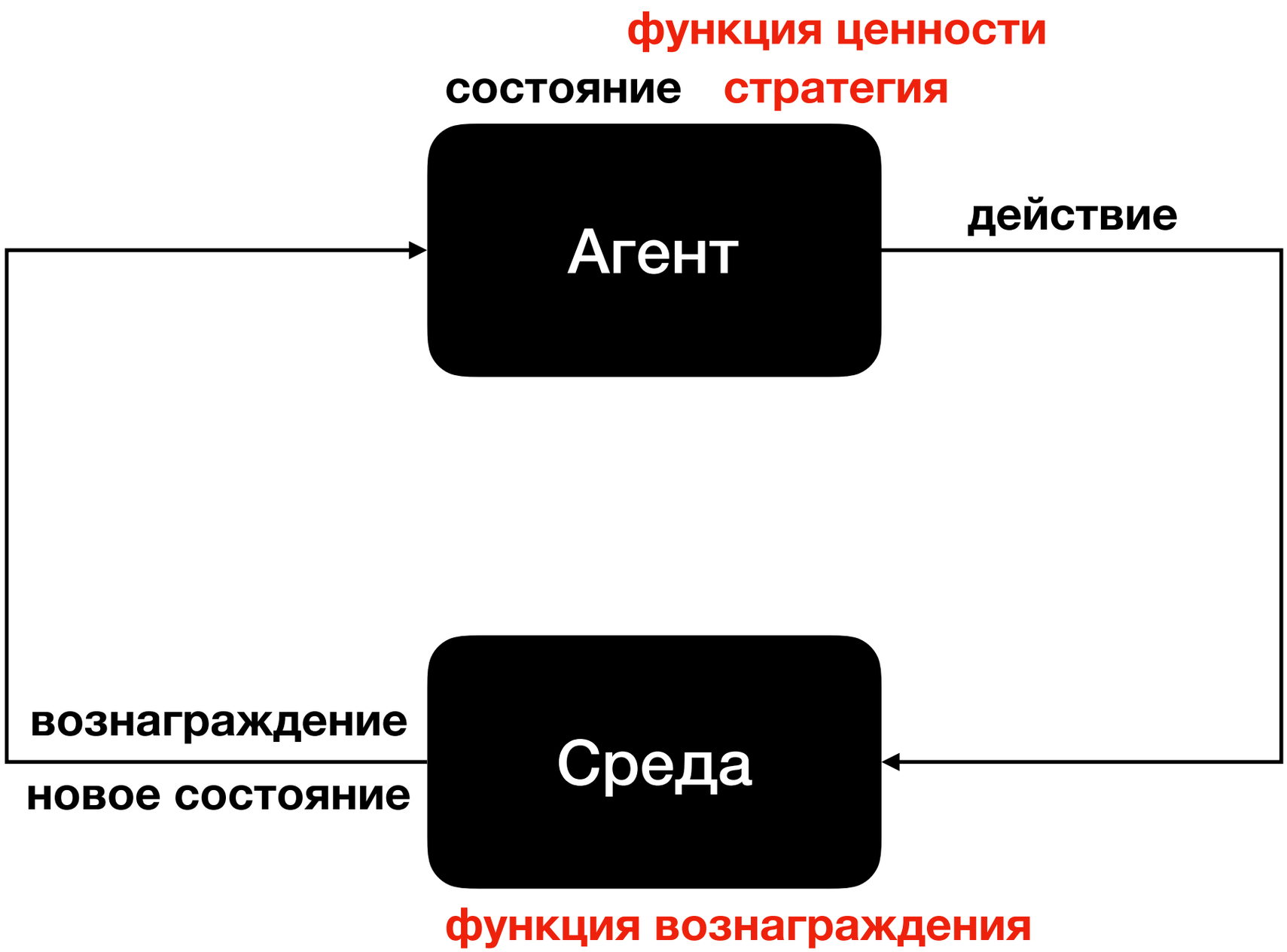
Постановка задачи



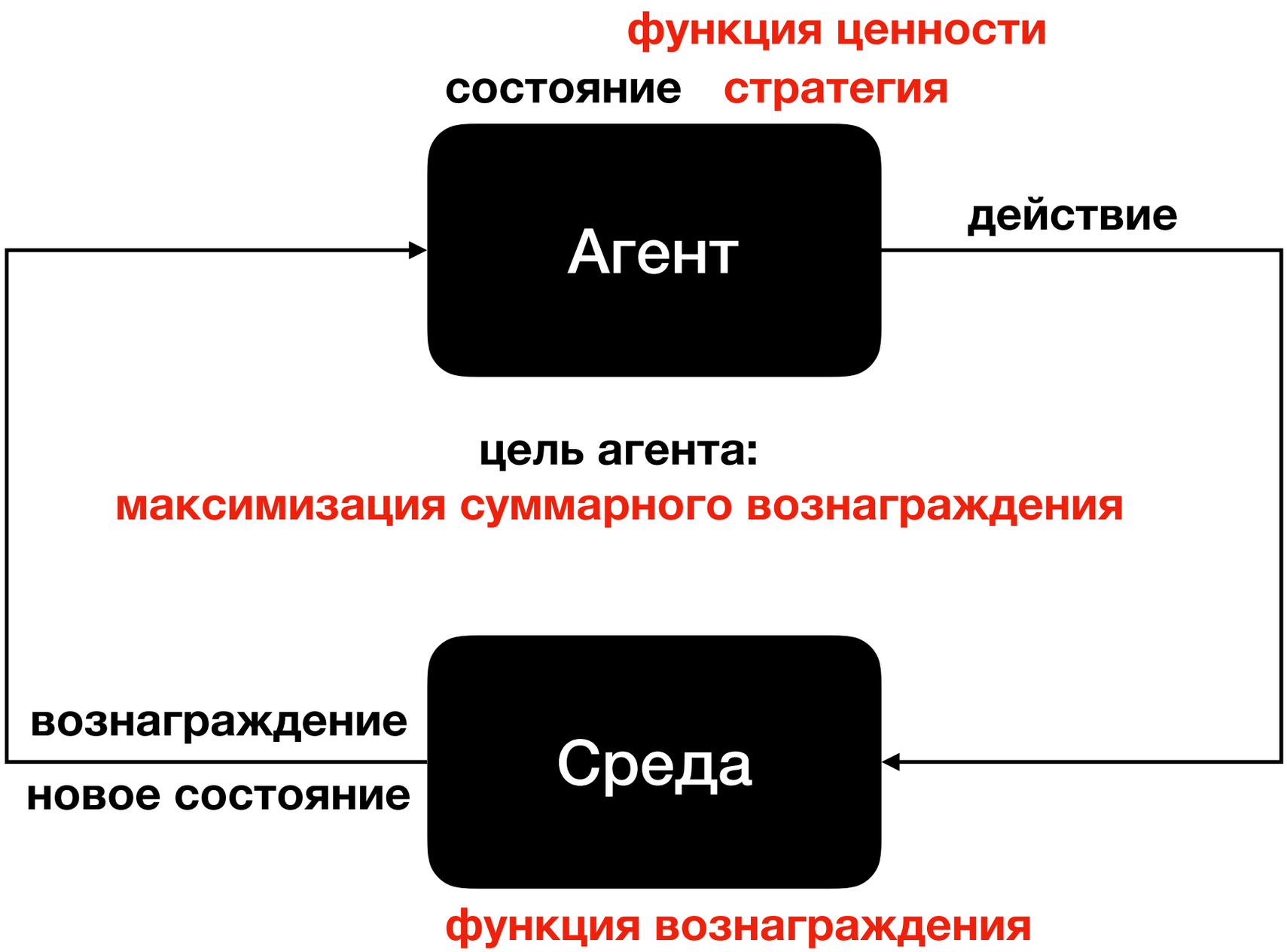
Постановка задачи



Постановка задачи



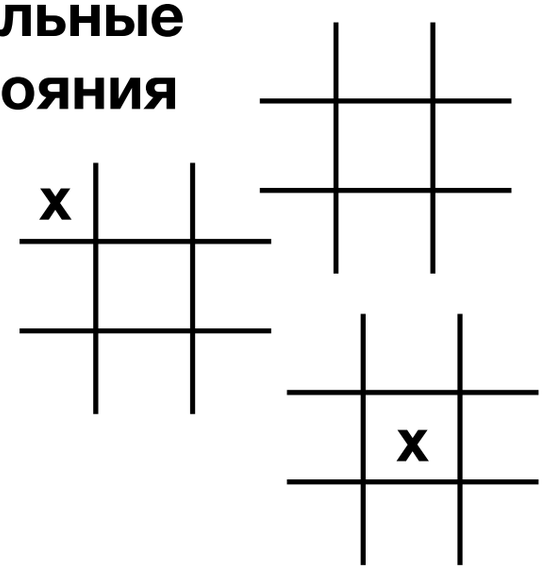
Постановка задачи



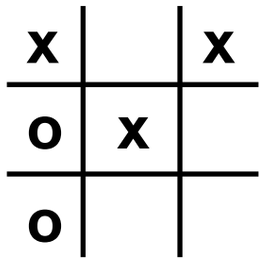
Эпизодическая задача

Крестики-нолики

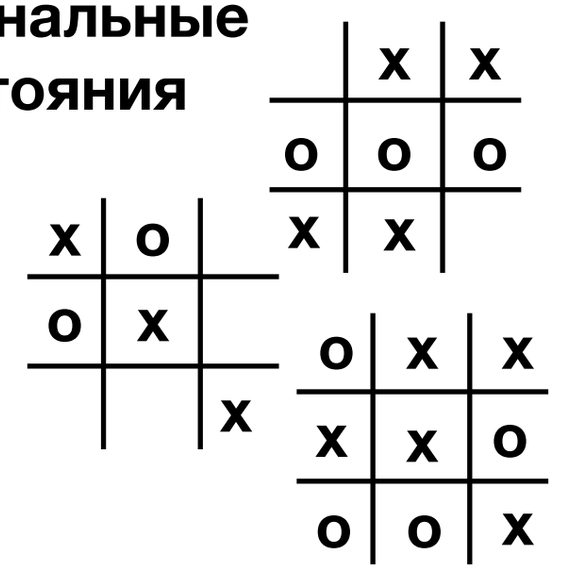
начальные состояния



текущее состояние

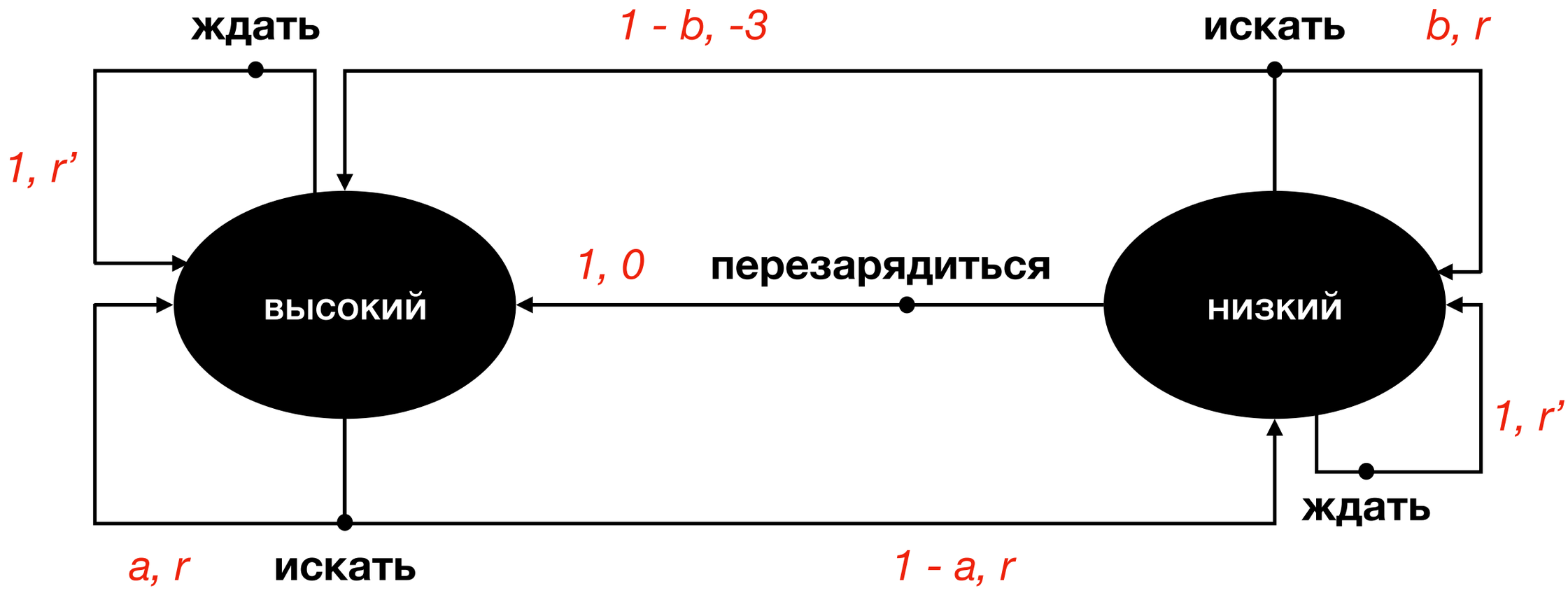


терминальные состояния



Не эпизодическая задача

Не эпизодическая задача



ε-жадный алгоритм

$\varepsilon \in (0, 1)$

random_num $\in$ **[0, 1]**

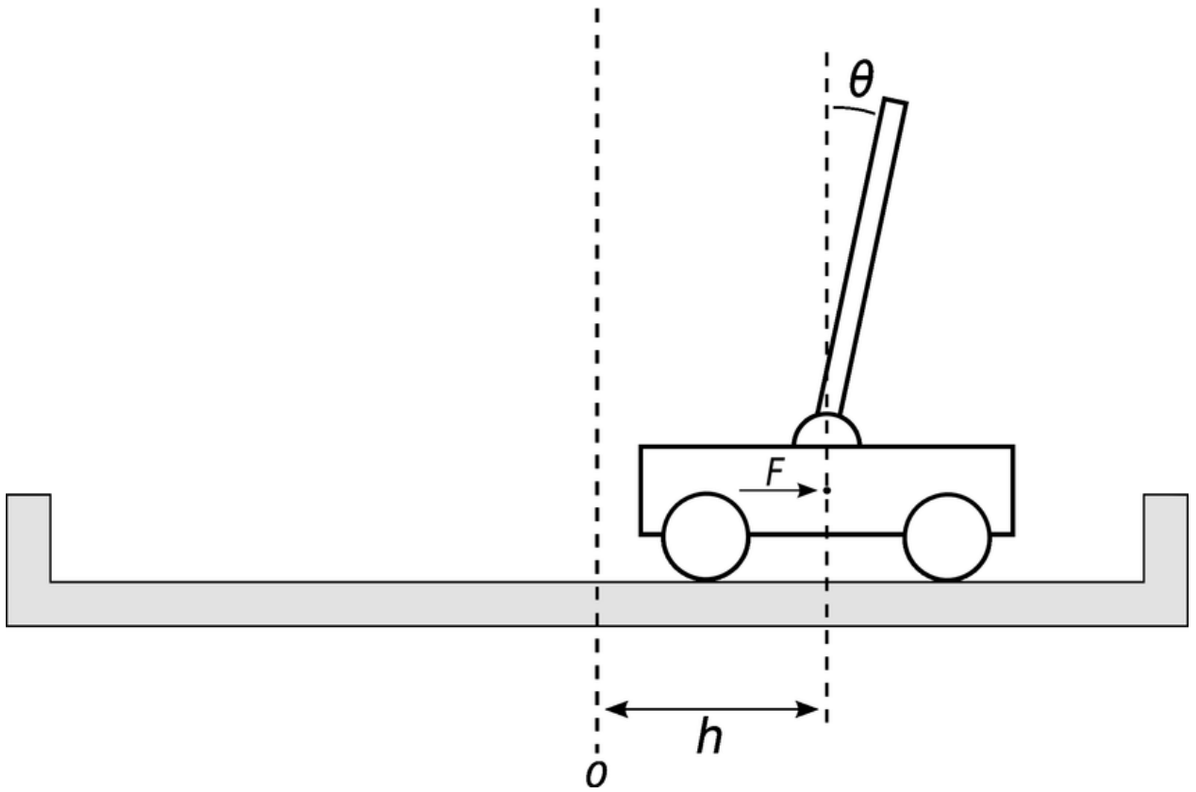
```
if random_num < ε:
    explore()

if random_num >= ε:
    exploit()
```

Табличные методы

функция ценности

действия / состояния	A1	A2	A3	A4	…
S1	0.1	0	1	0	
S2	0.2	0.32	0	0.15	
S3	0.34	0	0.88	0	
S4	0.01	0	0	0	
…					

Не табличные задачи



величина угла?

приложенная сила?

SARSA

state - action - reward - state - action

цикл:

агент находится в состоянии *s*

агент совершает действие *a* **согласно стратегии**

стратегия ε-жадная

среда возвращает вознаграждение *r* **и новое состояние** *s'*

агент выбирает действие *a'* **из состояния** *s'* **согласно стратегии,**

не совершая его

скорость обучения

$$Q(s, a) := Q(s, a) + alpha(r + gamma * Q(s', a') - Q(s, a))$$

функция ценности

коэффициент дисконтирования

Табличный метод кросс-энтропии

подходит для эпизодических задач

цикл:

производится N эпизодов игры
выбирается M лучших эпизодов

на лучших эпизодах считается:
P[*s*, *a*] = (сколько раз было принято действие *a* из состояния *s*) /
(сколько раз были в состоянии *s*)
стратегия = argmax (P[s, :])

Табличный метод кросс-энтропии

Недостаток метода:

— **если за все эпизоды мы были в состоянии s всего один раз (или очень мало раз)**

Табличный метод кросс-энтропии

Недостаток метода:

— **если за все эпизоды мы были в состоянии** s **всего один раз (или очень мало раз)**

сглаживание:

P[s**,** a**] = (сколько раз было принято действие** a **из состояния** $s + \lambda$**) / (сколько раз были в состоянии** $s + \lambda$ *** количество действий)**

Важно правильно задать функцию вознаграждения

Вопросы

1. Чем отличается функция вознаграждения от функции ценности?
2. Что такое эпсилон-жадная стратегия?
3. В чем заключается табличный метод кросс-энтропии?

Источники

1. Sutton and Barto «Reinforcement Learning», главы 1, 3 (общие идеи обучения с подкреплением и хорошие примеры)
2. https://www.coursera.org/lecture/practical-rl/crossentropy-method-TAT8g (кратко метод кросс-энтропии)
3. https://disk.yandex.ru/i/dPsWYMK13EDJj7 (семинад ШАД про метод кросс-энтропии)
4. http://www.machinelearning.ru/wiki/images/3/35/Voron-ML-RL-slides.pdf (общие идеи, табличные методы)