ТюмГУ

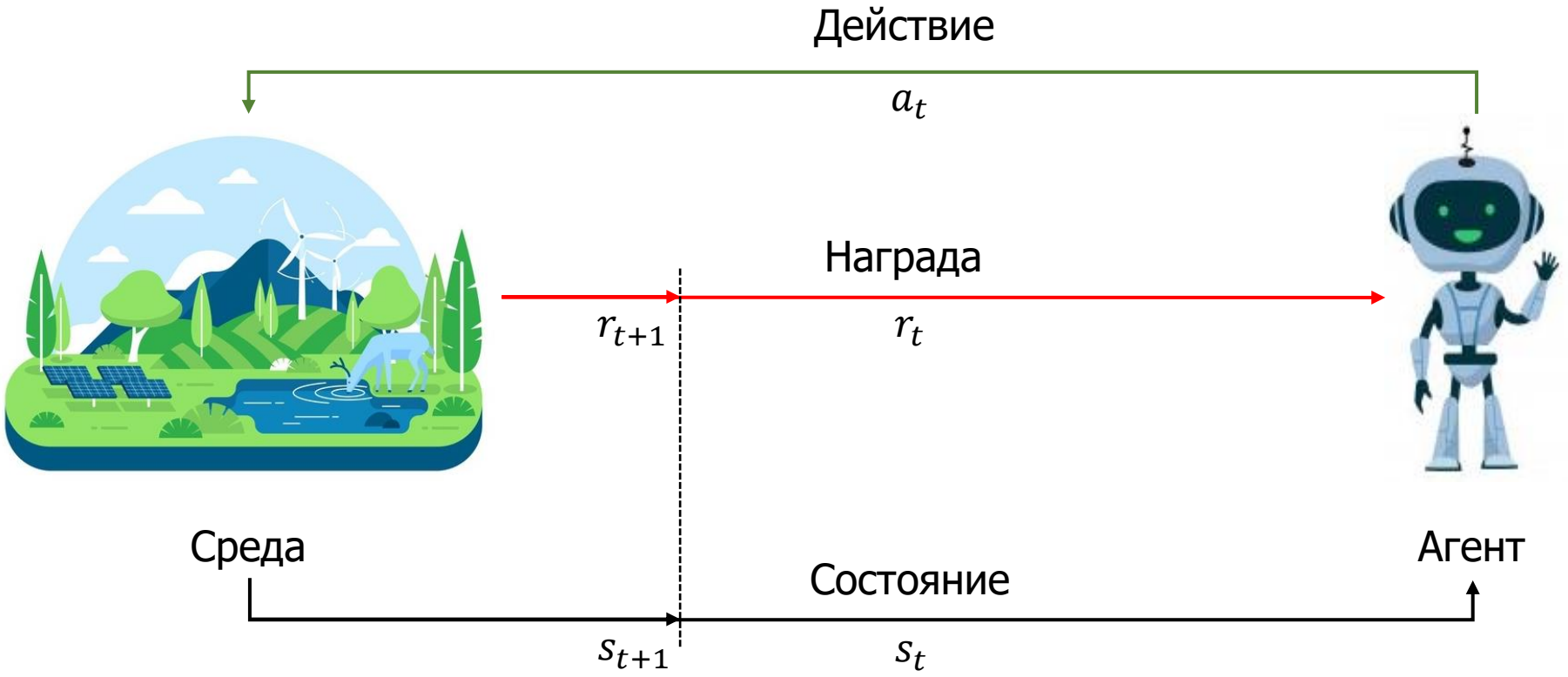
Обучение с подкреплением.

Введение. Математические основы.

Сергей Аксёнов,
Доцент отделения информационных технологий
Инженерной школы информационных технологий и робототехники
Томский политехнический университет

Тюмень-2024

Обучение с подкреплением

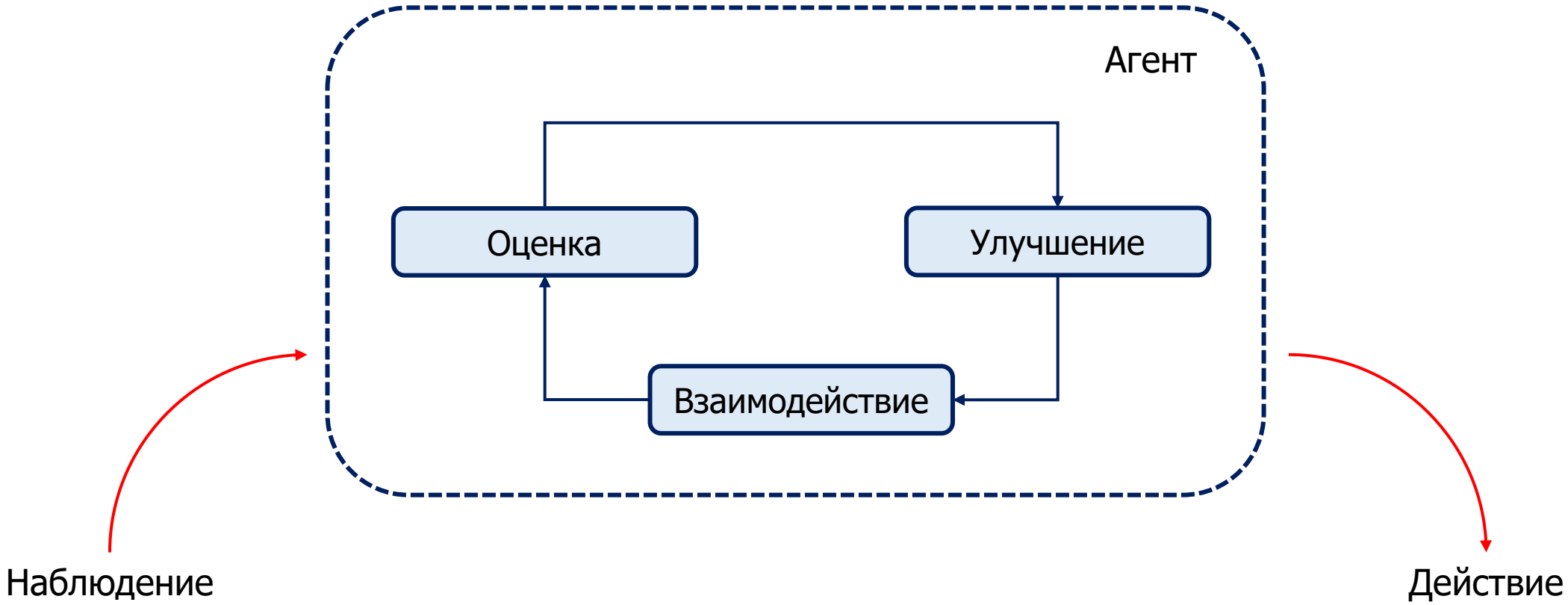


Использованы изображения из freepik

Компоненты обучения с подкреплением

Компонент	Объяснение
Агент	Программа, обучающаяся делать интеллектуальные решения
Среда	Мир агента. Представление проблемы
Состояние	Положение или момент в среде, в котором может находиться агент
Действие	Агент взаимодействует со средой, выполняя действие и переходя из одного состояния в другое.
Награда	Числовое значение, которое агент получает за своё действие
Пространство состояний	Набор всех возможных состояний
Политика (стратегия)	Агент принимает решение на основе политики. Политика указывает агенту, какое действие выполняется в каждом состоянии
Наблюдение	Часть состояния, которое агент может обозревать
Эпизод	Взаимодействие агента и среды с начального состояния до терминального
Горизонт	Временной шаг, до которого агент взаимодействует со средой
Выгода	Сумма наград, получаемая агентом в эпизоде

Работа агента



Алгоритм обучения с подкреплением

1. Сначала агент взаимодействует со средой, выполняя действия.

2. Агент делает действие и переходит с одного состояния в другое.

3. Далее агент получает вознаграждение в зависимости от выполненного им действия.

4. По вознаграждению агент понимает, хорошее или плохое действие он совершил.

5. Если действие было хорошим, то есть если агент получил положительное вознаграждение, то агент предпочтет выполнить это действие, в противном случае агент попытается выполнить другие действия, которые могут привести к положительному вознаграждению. Таким образом, обучение с подкреплением — это, по сути, процесс обучения методом проб и ошибок.

Ожидание

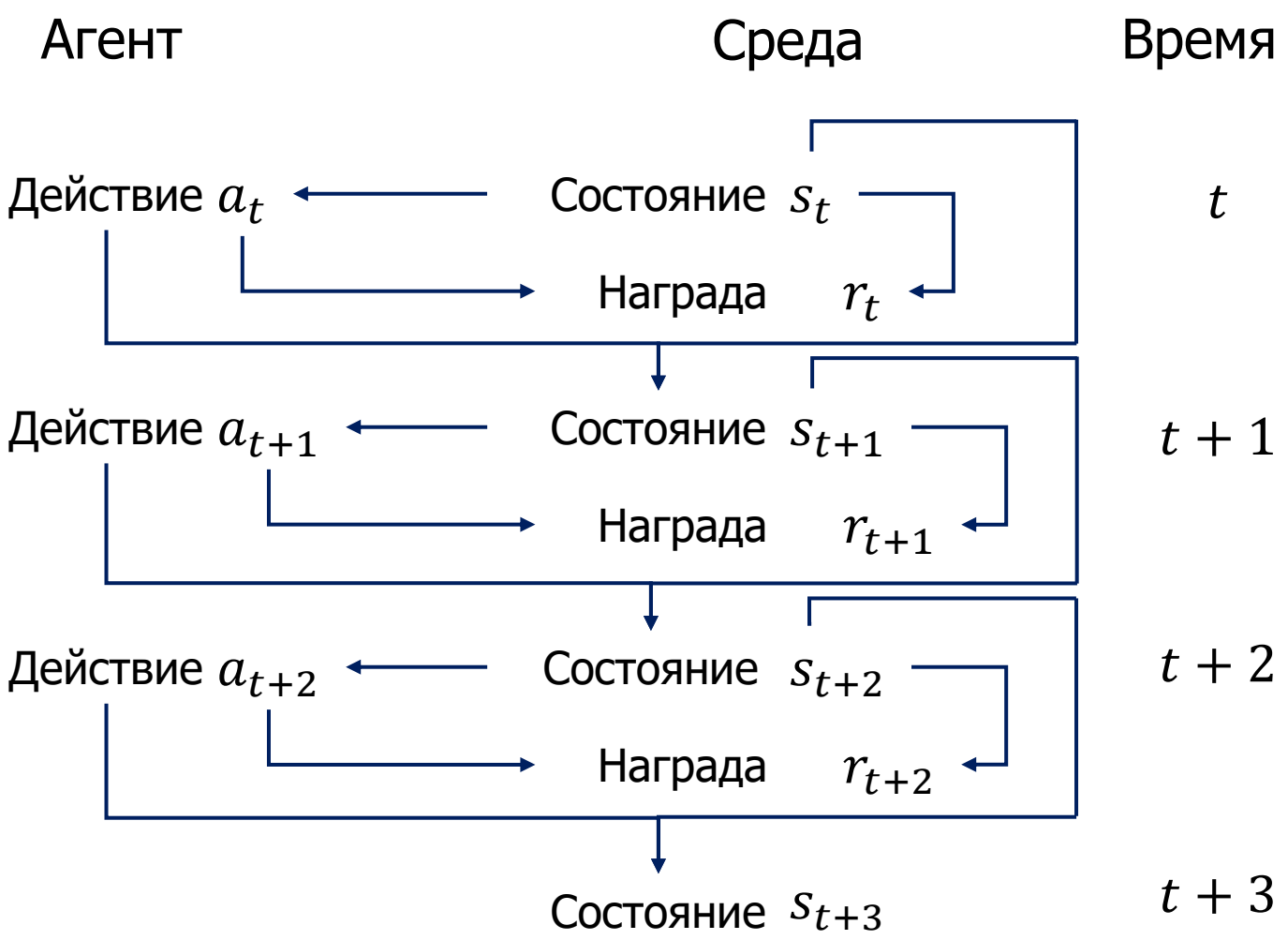
Вероятности выпадения кубика с шестью гранями

X	1	2	3	4	5	6
P(X)	1/6	1/6	1/6	1/6	1/6	1/6

$$E(X) \stackrel{\text{def}}{=} \sum_{i=1}^{N} x_i p(x_i) \qquad E(X) = 1 * \frac{1}{6} + 2 * \frac{1}{6} + 3 * \frac{1}{6} + 4 * \frac{1}{6} + 5 * \frac{1}{6} + 6 * \frac{1}{6} = 3.5$$

$$\mathbb{E}_{x \sim p(x)}[f(X)] \stackrel{\text{def}}{=} \sum_{i=1}^{N} f(x_i) p(x_i)$$

Кортежи опыта



Опыт:

$t, (s_t, a_t, r_t, s_{t+1})$

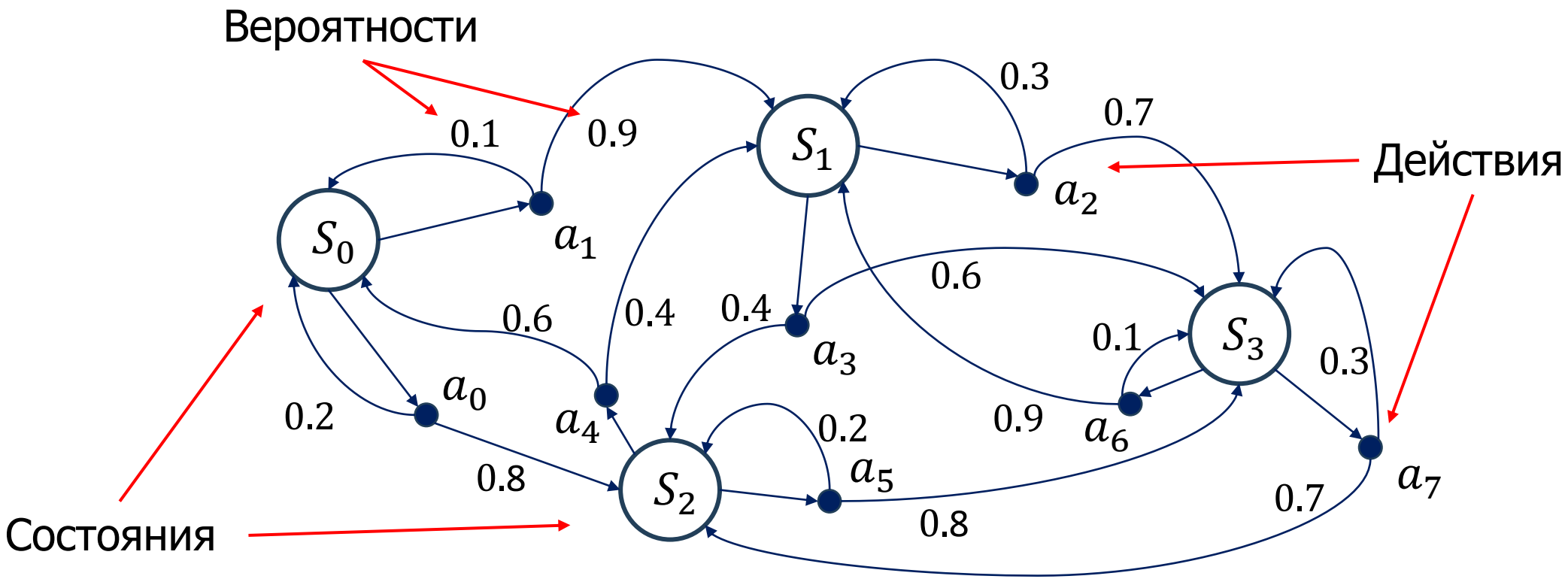
$t+1, (s_{t+1}, a_{t+1}, r_{t+1}, s_{t+2})$

$t+2, (s_{t+2}, a_{t+2}, r_{t+2}, s_{t+3})$

Немного терминов

Термин	Объяснение
Эпизодическая задача	Задача, имеющая терминальное состояние
Продолжающаяся задача	Задача, не имеющая терминального состояния
Коэффициент дисконтирования	Фактор определяет, хотим ли мы придавать значение немедленному вознаграждению или будущим вознаграждениям
Функция ценности состояния	Ожидаемая прибыль, которую агент получит, начиная с состояния s, следуя политике π
Q функция (ценности действия)	Ожидаемая прибыль, которую агент получит, начиная с состояния s и выполняя действие a в соответствии с политикой π
Детерминированная среда	Среда, в которой агент, выполняя действие a в состоянии s, переходит всегда в одно и то же состояние s' каждый раз
Стохастическая среда	Среда, в которой агент, выполняя действие a в состоянии s, может перейти в разные состояния каждый раз, основываясь на некотором распределении вероятностей

Марковский процесс принятия решений (MDP)



S - Набор состояний (Пространство состояний)

A - Набор действий (Пространство действий)

$A(s)$ - Набор действий, доступный из состояния s

Марковское свойство

Вероятность следующего состояния:

$$P(S_{t+1}|S_t, A_t) = P(S_{t+1}|S_t, A_t, S_{t-1}, A_{t-1}, \dots)$$

S_t - Состояние в момент времени t (текущее состояние)

A_t - Действие, выбранное в момент времени t (текущее действие)

Немного побольше терминов

Термин	Объяснение
Модель среды	Набор функций перехода и вознаграждения
Опыт	Набор состояние, действие, награда и новое состояние
Компромисс между сбором информацией и эксплуатацией	Баланс между сбором информации и использованием текущей информацией
Политико-ориентированные агенты	Агенты, аппроксимирующие политики
Ценностно-ориентированные агенты	Агенты, аппроксимирующие функции ценности
Модельно-ориентированные агенты	Агенты, аппроксимирующие модели среды
Актеры-критики	Агенты, аппроксимирующие функции ценности и политики

Функция перехода

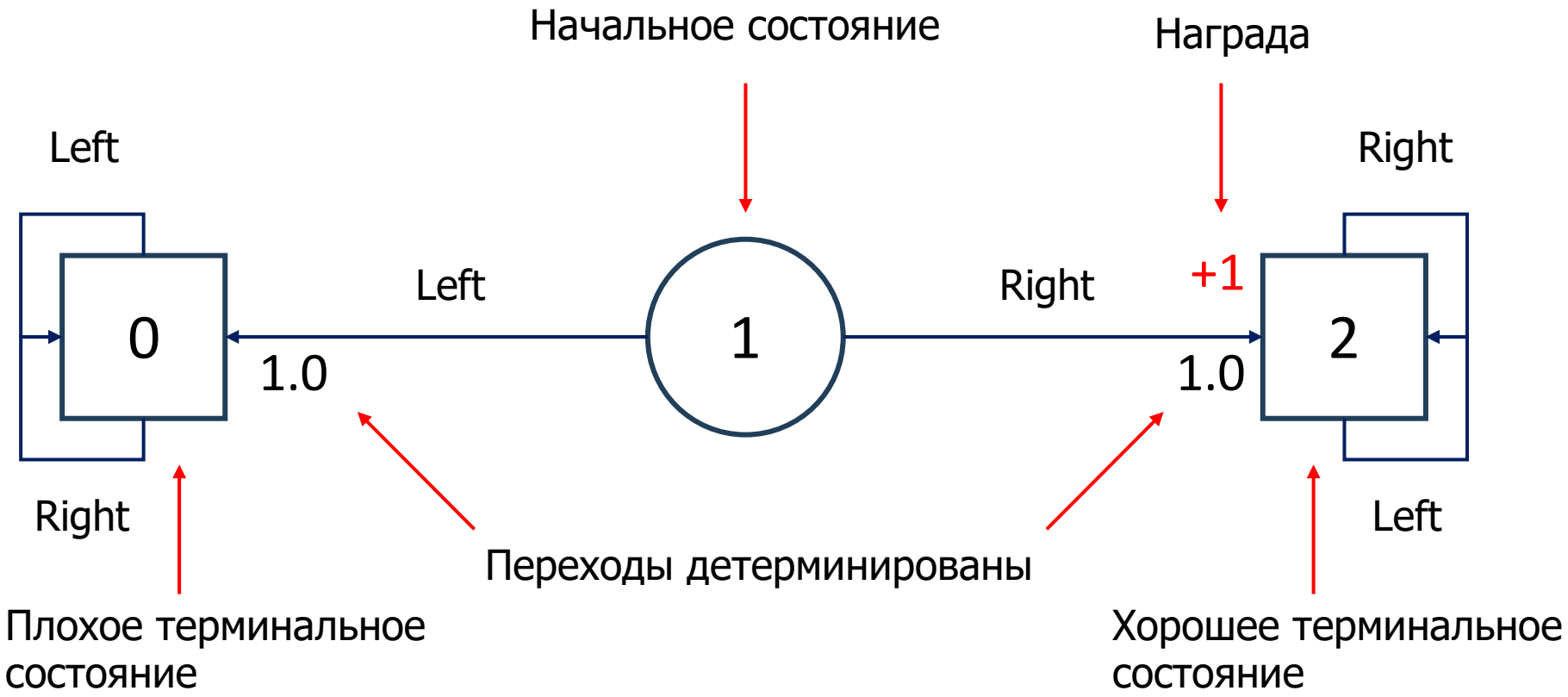
$$p(s'|s,a) \stackrel{\text{def}}{=} P(S_t = s'|S_{t-1} = s, A_{t-1} = a)$$

$$\sum_{s' \in S} p(s'|s,a) = 1, \forall s \in S, \forall a \in A(s)$$

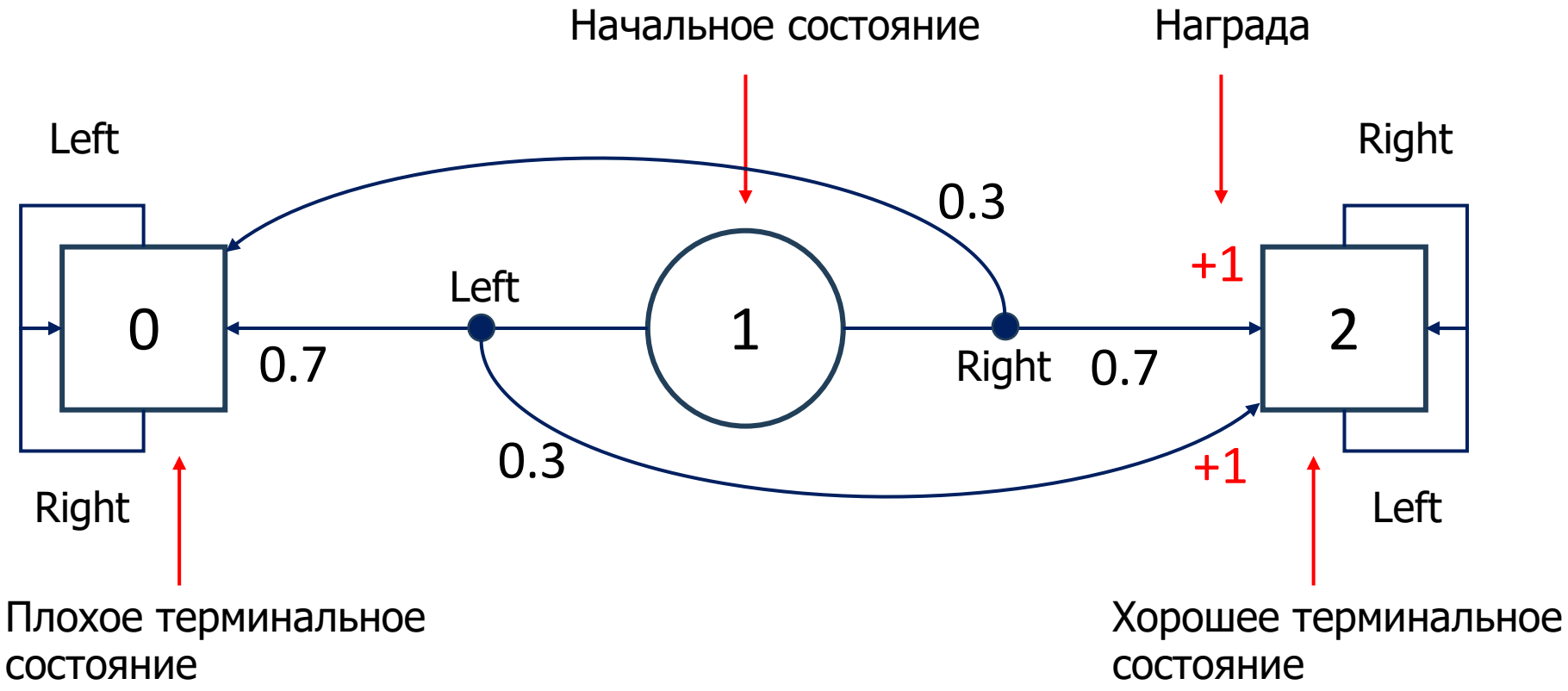
S_t - Состояние в момент t (текущее состояние)

A_t - Действие, выбранное в момент t (текущее действие)

Пример: Детерминированные переходы



Пример: Стохастические переходы



Функция вознаграждения

$$\mathrm{r}(\mathrm{s},\mathrm{a}) \stackrel{\text{def}}{=} \mathbb{E}(R_t | S_{t-1} = s, A_{t-1} = a)$$

$R_t \in \mathcal{R} \subset \mathbb{R}$ $\mathcal{R}$ - Набор всех вознаграждений

$$\mathrm{r}(\mathrm{s},\mathrm{a},s') \stackrel{\text{def}}{=} \mathbb{E}(R_t | S_{t-1} = s, A_{t-1} = a, S_t = s')$$

S_t - Состояние в момент t (текущее состояние)

A_t - Действие, выбранное в момент t (текущее действие)

Коэффициент дисконтирования γ и выгода G

Выгода

$G_t = R_{t+1} + R_{t+2} + R_{t+3} + \cdots + R_T$ Выгода после временного шага t $R_t \in \mathbb{R}$

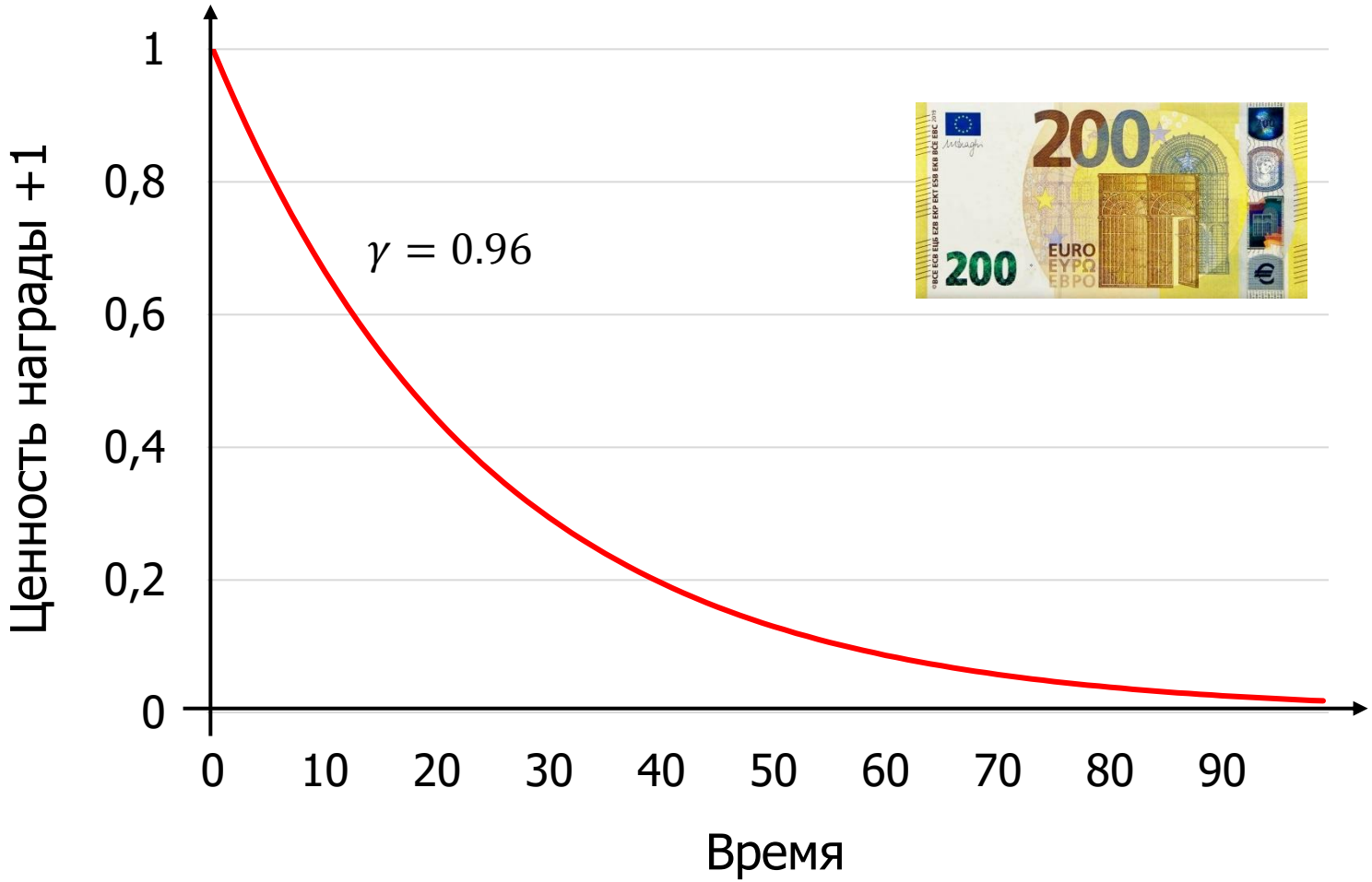
Дисконтированная выгода

$$G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \cdots = \sum_{k=0}^{\infty} \gamma^k R_{t+k+1}$$

$G_t = R_{t+1} + \gamma G_{t+1}$ Коэффициент дисконтирования $0 \leq \gamma \leq 1$

$$G_t = \lim_{T \to \infty} \frac{1}{T} \sum_{t=0}^{T-1} R_{t+i}$$

Влияние дисконтирования и времени на выгоду



Политика (стратегия) π

Политика (стратегия) – отображение пространства состояний на пространство действий.

Детерминированная политика

$$a = \pi(s), s \in S$$

Стохастическая политика

$$\pi(a|s) = p\{a_t = a|s_t = s\}, s \in S$$

Параметризуемая политика

$$\pi_\theta(s) \stackrel{\text{def}}{=} \pi(s|\theta)$$

$$\pi_\theta(a|s) \stackrel{\text{def}}{=} \pi(a|s;\theta)$$

Функция ценности состояния

Функция ценности состояния s в рамках политики π

$$v_\pi(s) \stackrel{\text{def}}{=} \mathbb{E}_\pi[G_t|S_t = s] = \mathbb{E}_\pi\left[\sum_{k=0} \gamma^{k+1} R_{t+k+1}|S_t = s\right]$$

Уравнение Беллмана для ценности состояния s в рамках политики π:

$$v_\pi(s) \stackrel{\text{def}}{=} \sum_a \pi(a|s) \sum_{s',r} p(s', r|s, a)[r + \gamma v_\pi(s')], \forall s \in S$$

Функция ценности действия Q

Функция ценности действий согласно политике π (ценность действия a в состоянии s согласно политике π)

$$q_\pi(s,a) \stackrel{\text{def}}{=} \mathbb{E}_\pi[G_t|S_t = s, A_t = a] = \mathbb{E}_\pi\left[\sum_{k=0} \gamma^{k+1} R_{t+k+1}|S_t = s\,, A_t = a\right]$$

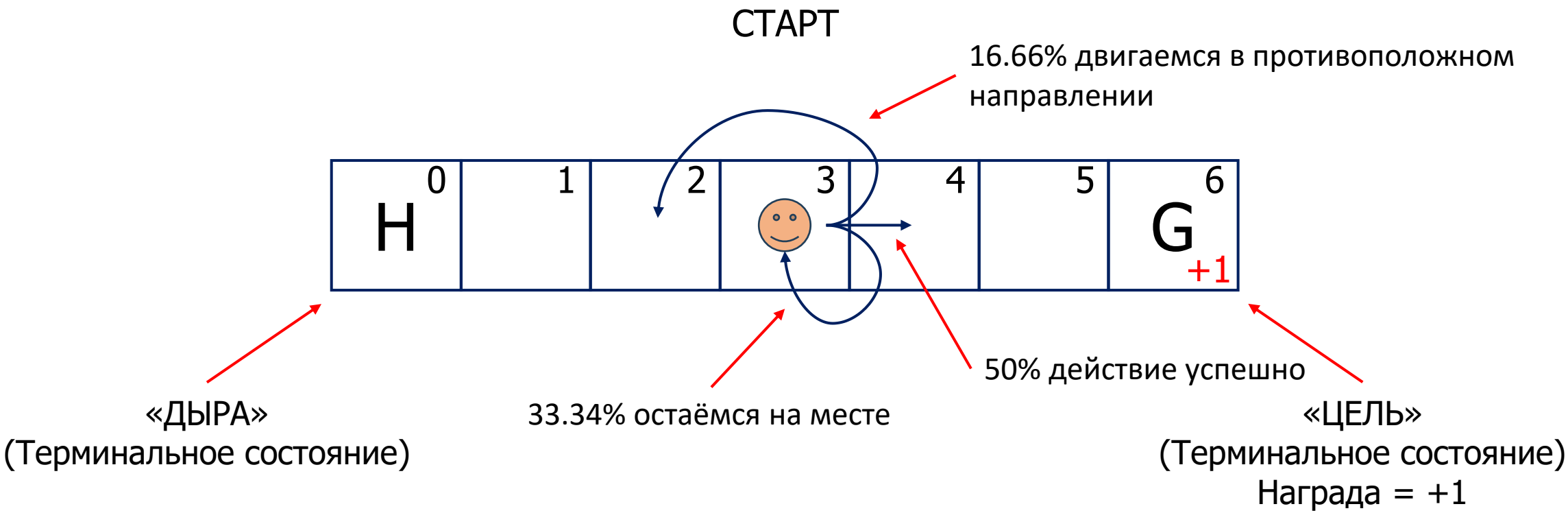
Уравнение Беллмана для ценности действий:

$$q_\pi(s,a) \stackrel{\text{def}}{=} \sum_{s',r} p(s',r|s,a)[r + \gamma v_\pi(s')]\,, \forall s \in S, \forall a \in A(s)$$

Функция преимущества действий A или $A^\pi(s,a)$. Преимущество действия a в состоянии s согласно политике π

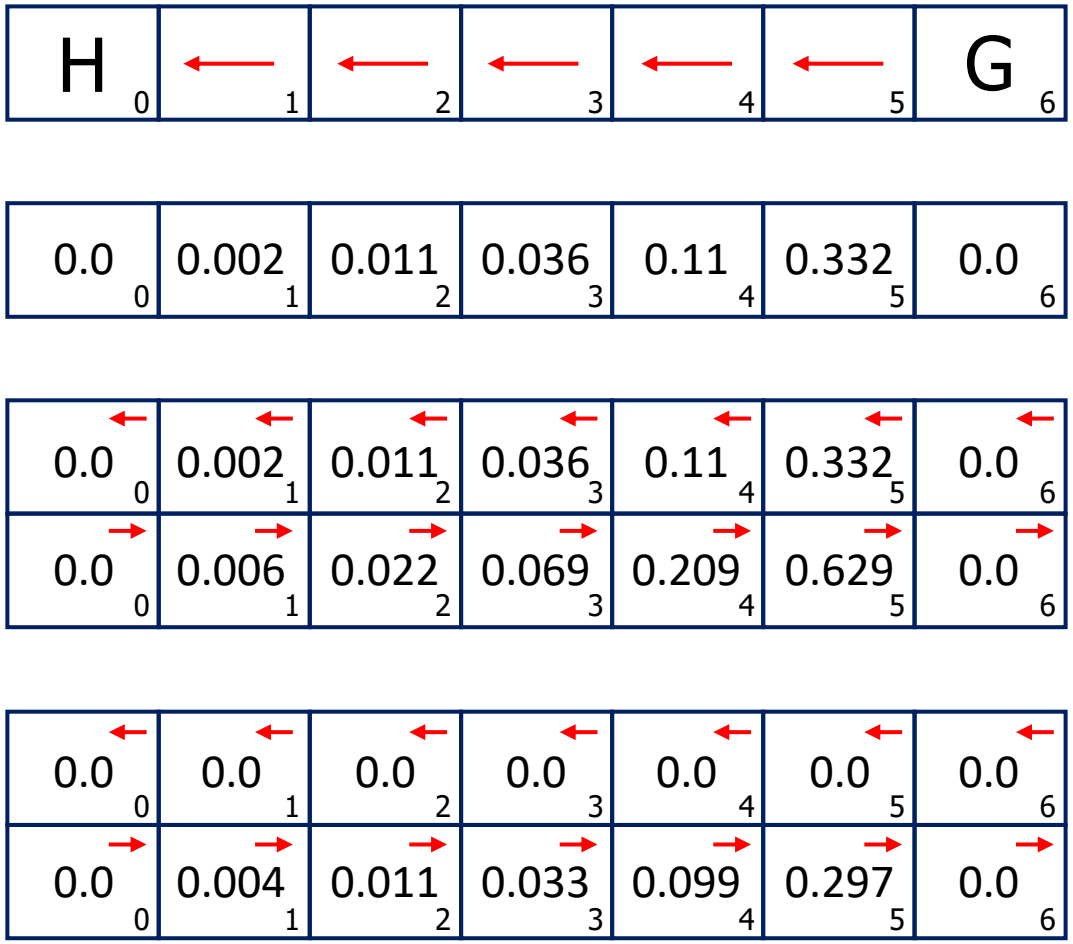
$$a_\pi(s,a) \stackrel{\text{def}}{=} q_\pi(s,a) - v_\pi(s)$$

Пример



Действия: LEFT, RIGHT

Пример: Функции состояния, действий и преимущества действий



Уравнения оптимальности Беллмана

Оптимальная функция оценки состояния

$$v_*(s) = \max_\pi v_\pi(s), \forall s \in S$$

$$v_*(s) = \max_a \sum_{s',r} p(s', r|s, a)\,[r + \gamma v_*(s')]$$

Оптимальная функция оценки действий

$$q_*(s, a) = \max_\pi q_\pi(s, a), \forall s \in S, \forall a \in A(s)$$

$$q_*(s, a) = \sum_{s',r} p(s', r|s, a)\left[r + \gamma \max_{a'} q_*(s', a')\right]$$

Обучение на основе модели и без модели

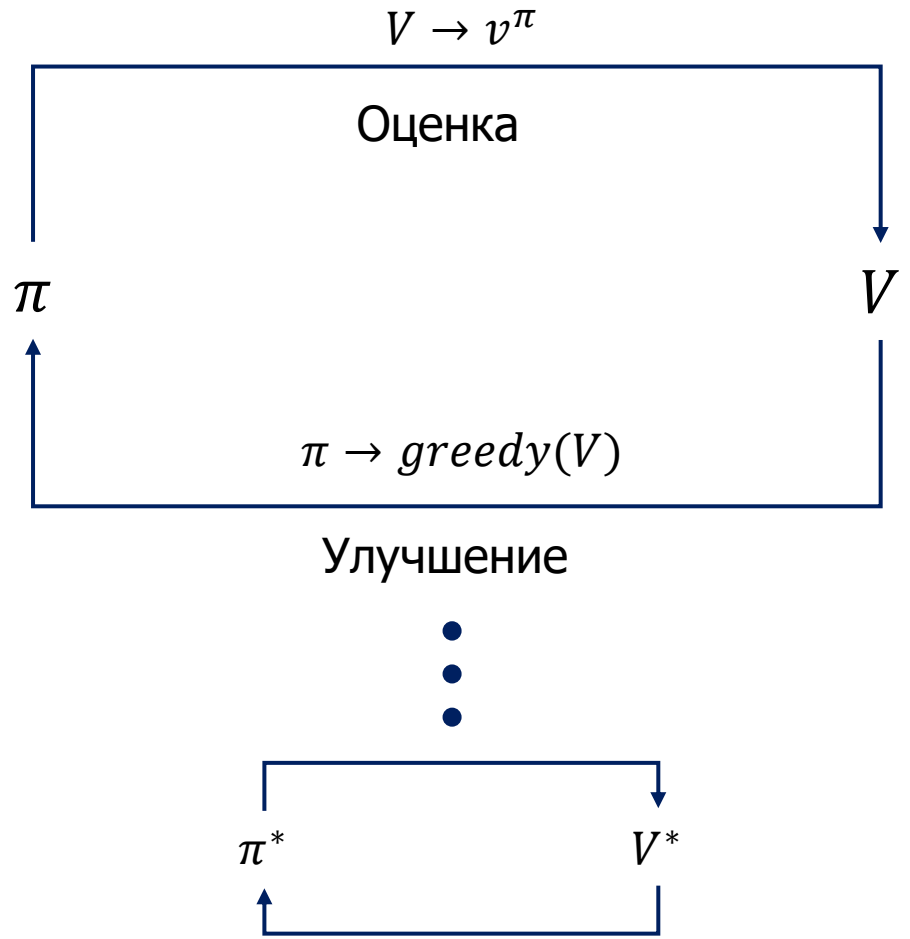
Обучение на основе модели

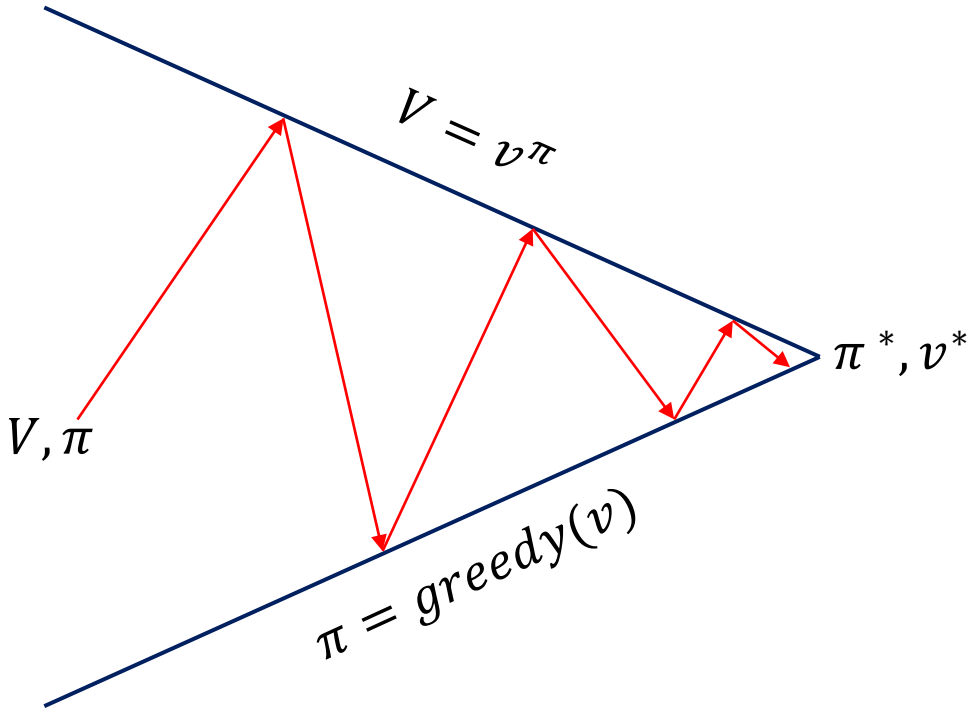
Агент имеет полное описание среды. Он знает модельную динамику модели среды, то есть функцию перехода и функцию вознаграждения. Агент использует динамику модели для поиска оптимальной политики.

Обучение без модели

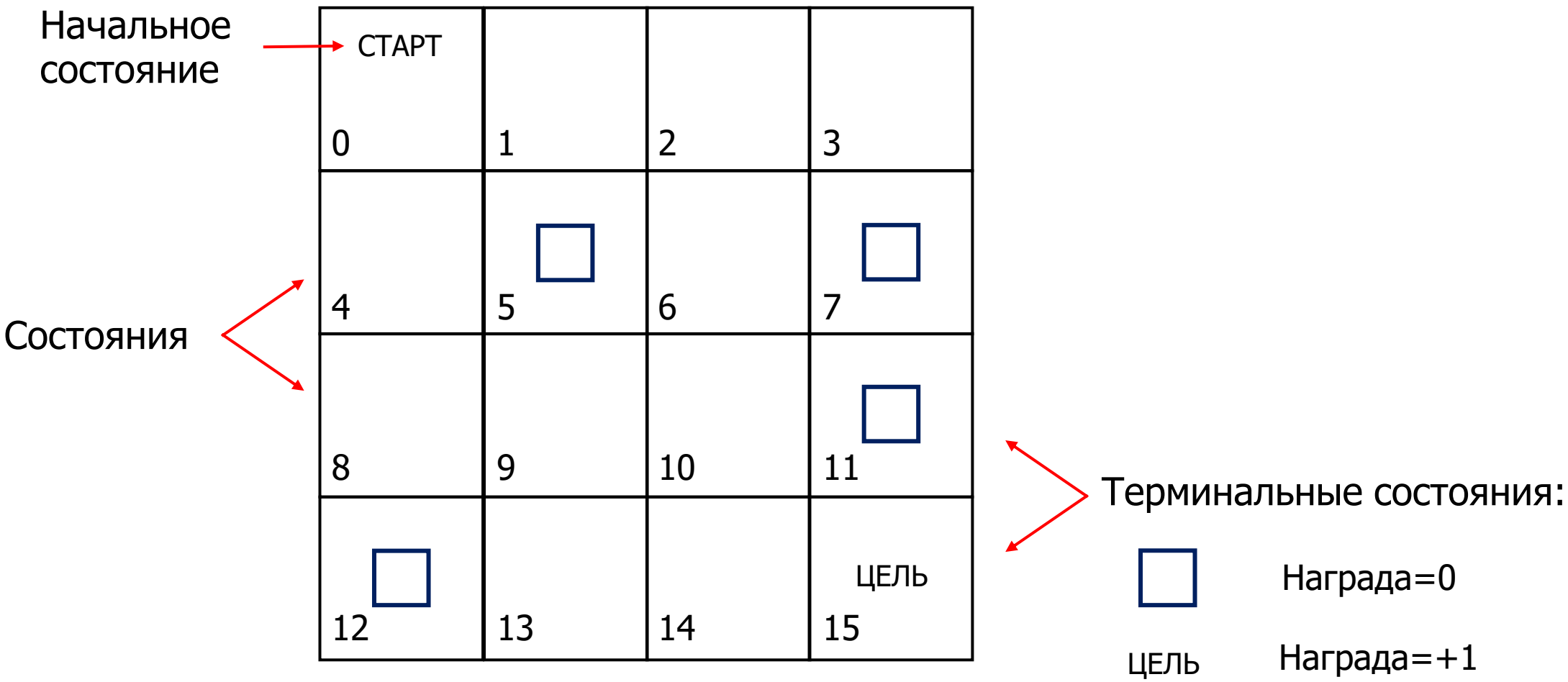
Агент не знает динамики модели своей среды. Он пытается найти оптимальную политику без динамики модели.

Обобщенная итерация политик (GPI)



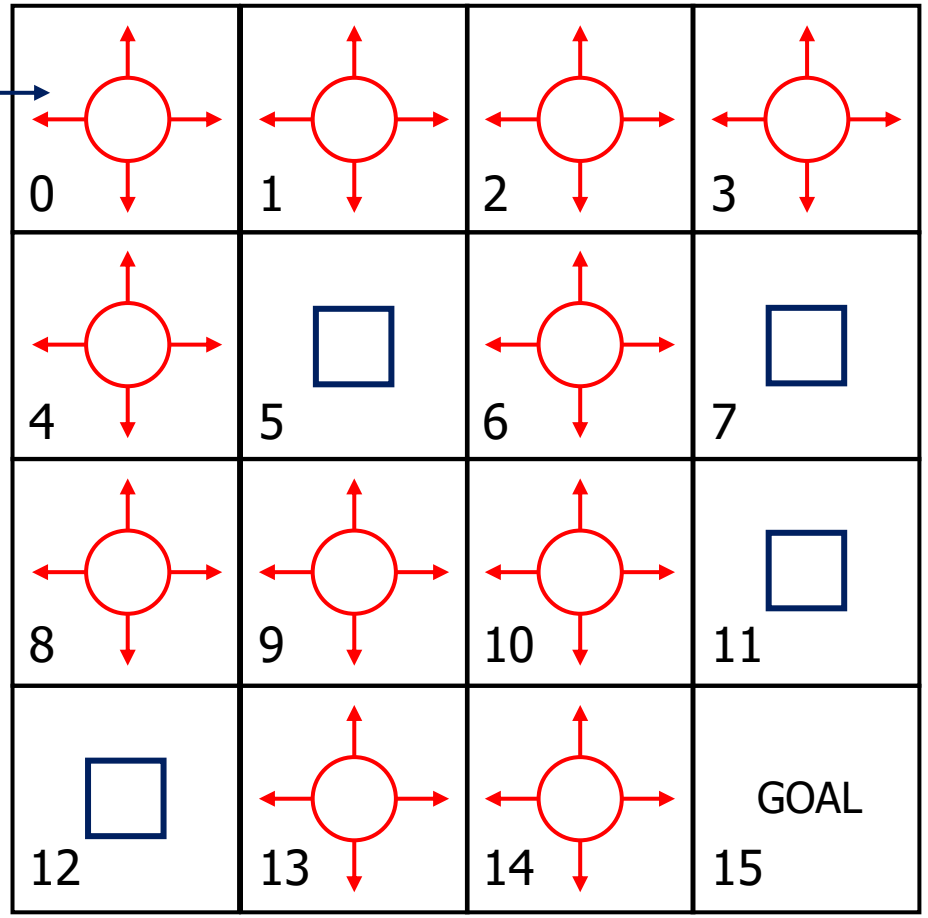


Среда Grid World: состояния

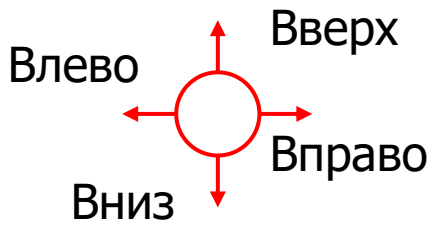


Среда Grid World: Действия

Начальное состояние



Действия:



Терминальные состояния:

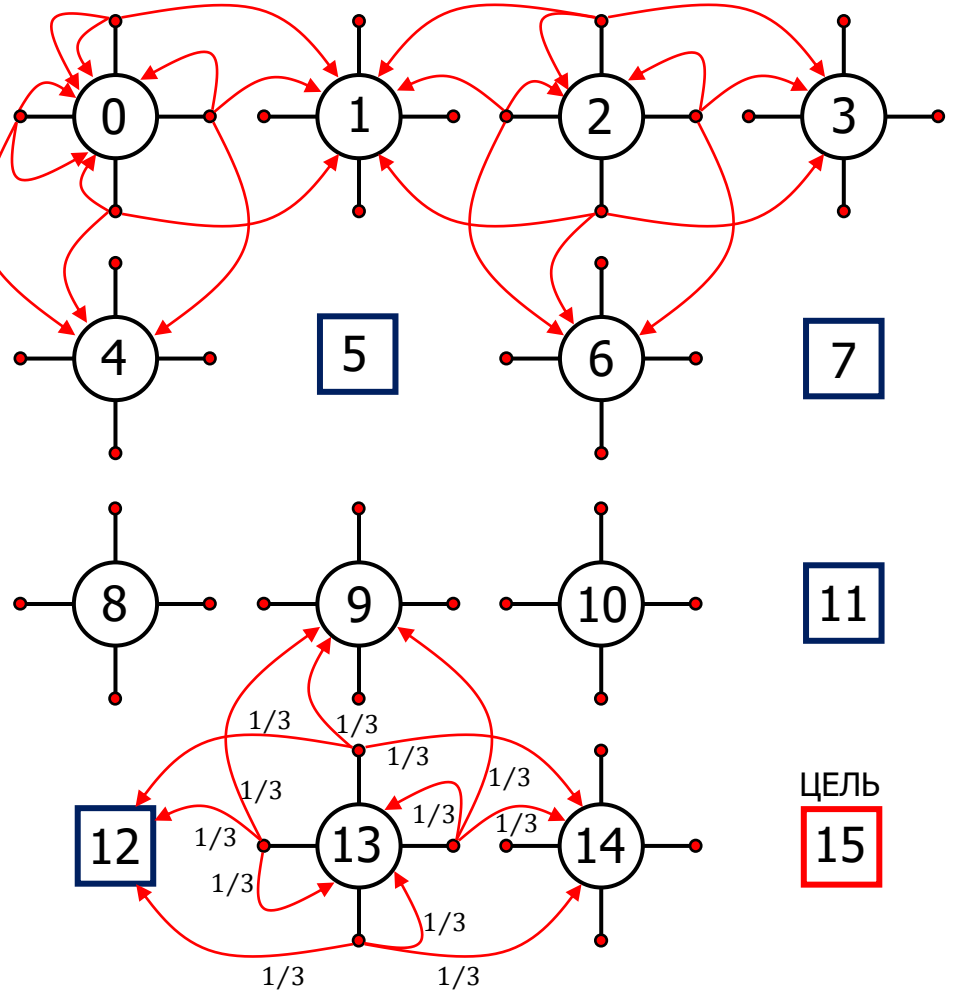
Награда=0

ЦЕЛЬ

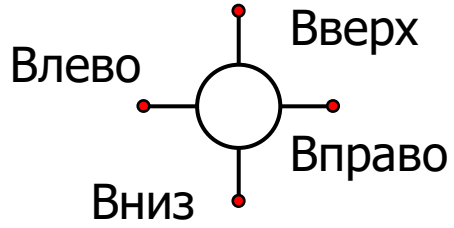
Награда=+1

Среда Grid World: Функция перехода

Начальное состояние

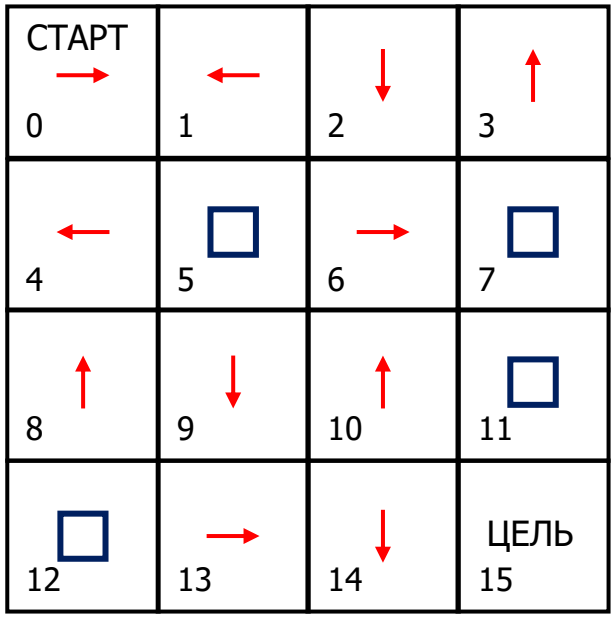


Действия:

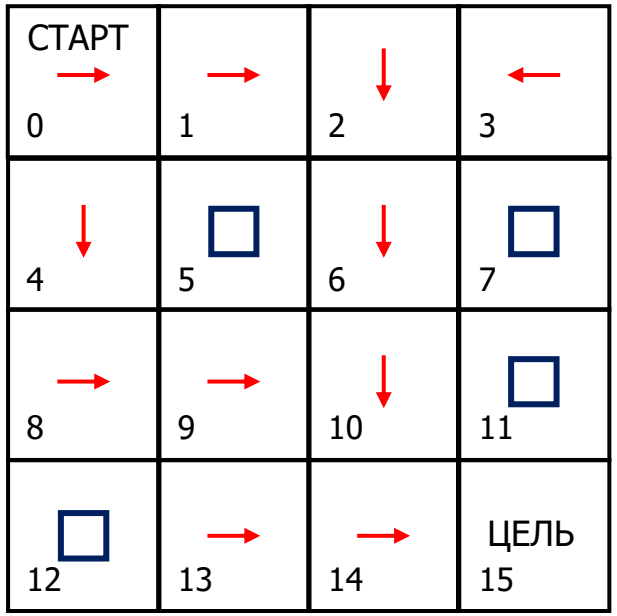


Различные политики (стратегии)

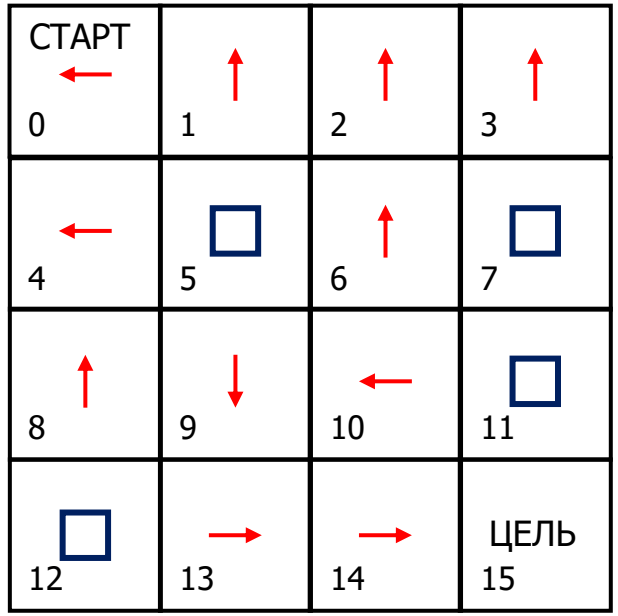
«Случайная»



«Быстрее к цели»



«Осторожная»



Алгоритм оценки политики (стратегии)

Итеративная аппроксимация функции ценности оцениваемой политики

1. Инициализировать $v_0(s)$ для всех s в S произвольными значениями, и 0 для всех s , являющимися терминальными.
2. С увеличением k итеративно улучшать оценки:

$$v_{k+1}(s) = \sum_a \pi(a|s) \sum_{s',r} p(s', r|s, a)[r + \gamma v_\pi(s')]$$

Алгоритм сходится, когда k приближается к бесконечности.

Пример 1: Оценка политики

S

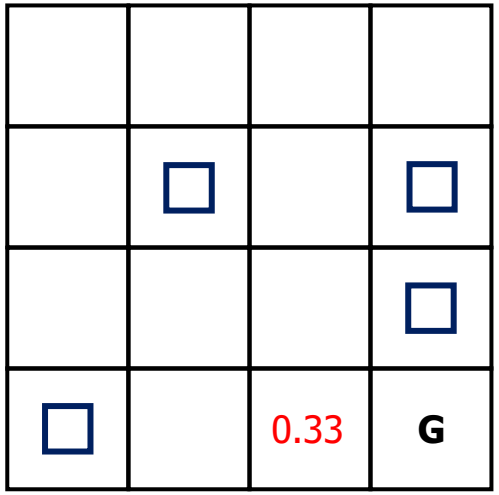
	H	←	←	←	←	←	G

π

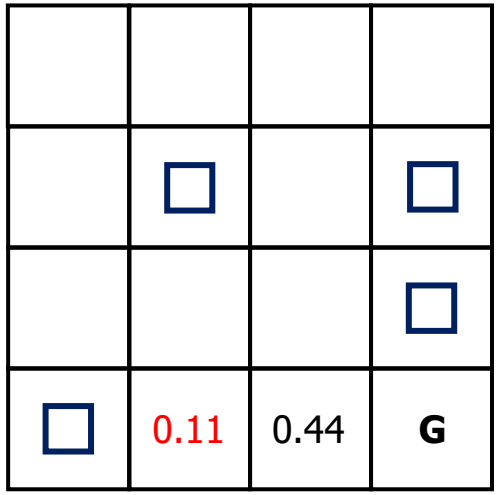
$k = 0$	0	0	0	0	0	0	0
$k = 1$	0	0	0	0	0	0.166	0
$k = 2$	0	0	0	0	0.027	0.222	0
$k = 3$	0	0	0	0.004	0.046	0.254	0
$k = 4$	0	0	0.000	0.009	0.060	0.274	0
$k = 5$	0	0.000	0.001	0.013	0.070	0.288	0
$k = 6$	0	0.000	0.002	0.017	0.078	0.298	0

V^{π}

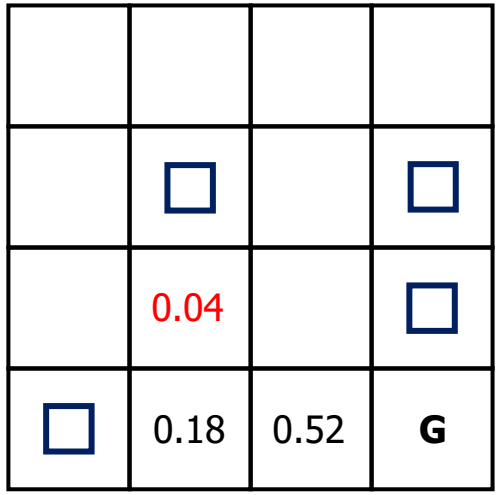
Пример 2: Оценка политики («Случайная»)



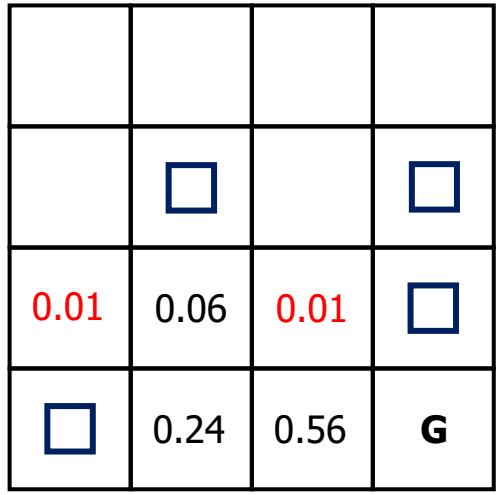
$k = 1$



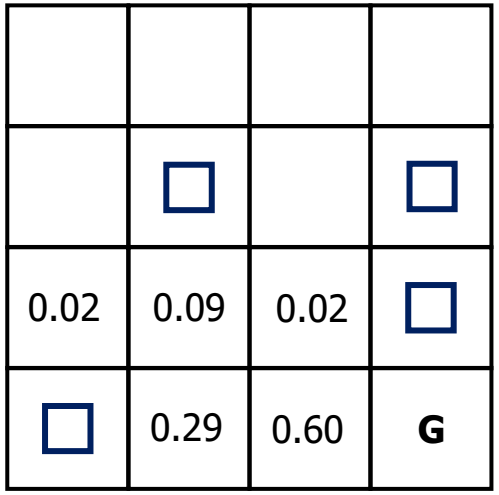
$k = 2$



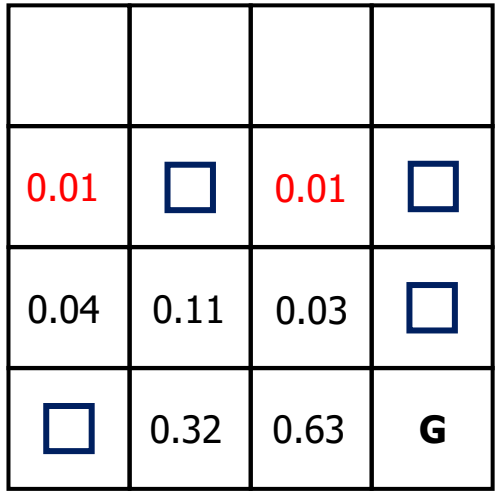
$k = 3$



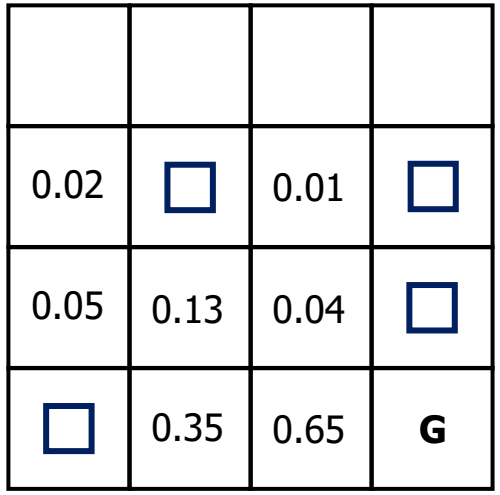
$k = 4$



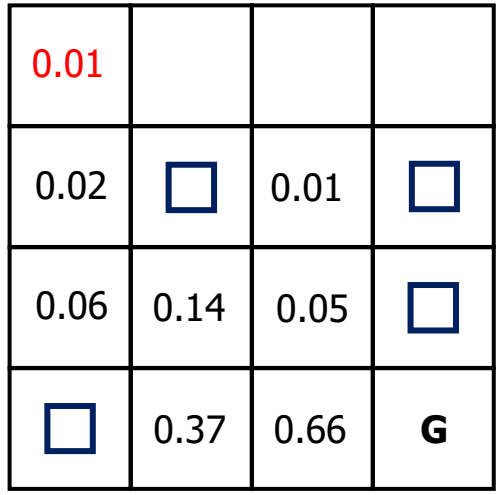
$k = 5$



$k = 6$



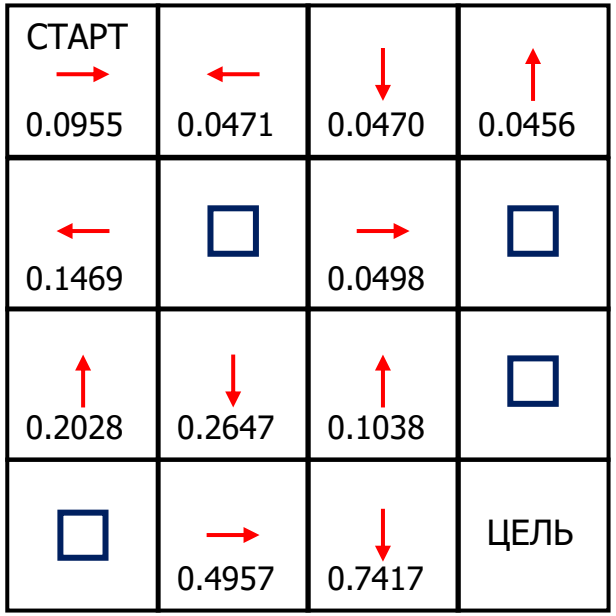
$k = 7$



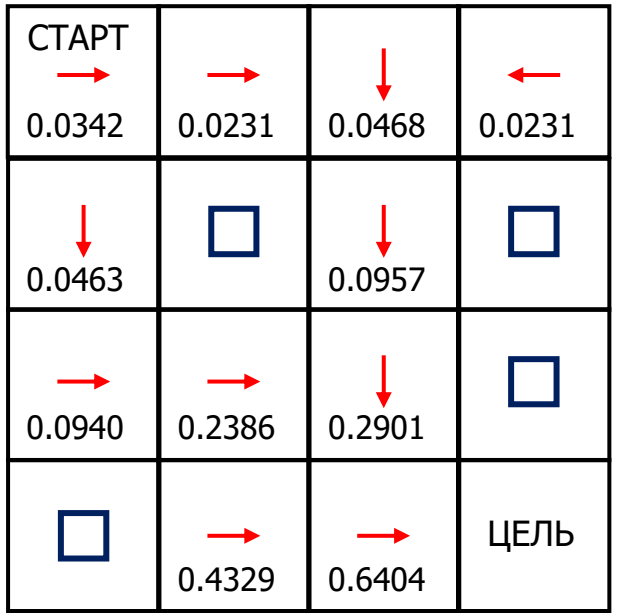
$k = 8$

Пример 3: Оценка политик (3 политики)

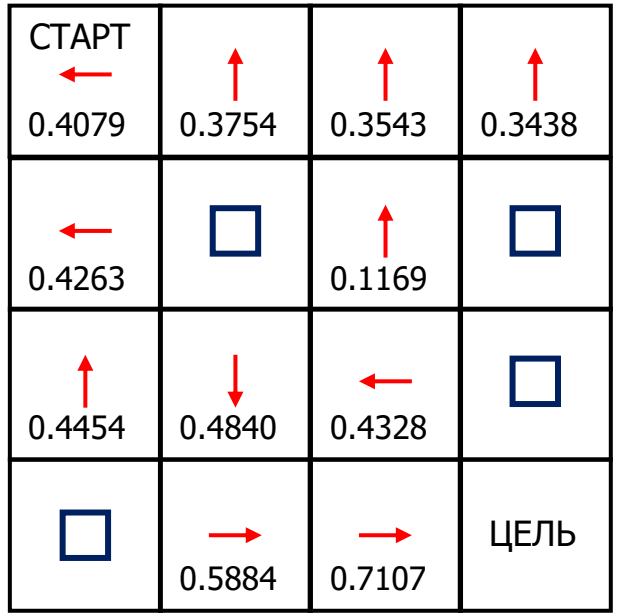
«Случайная»



«Быстрее к цели»

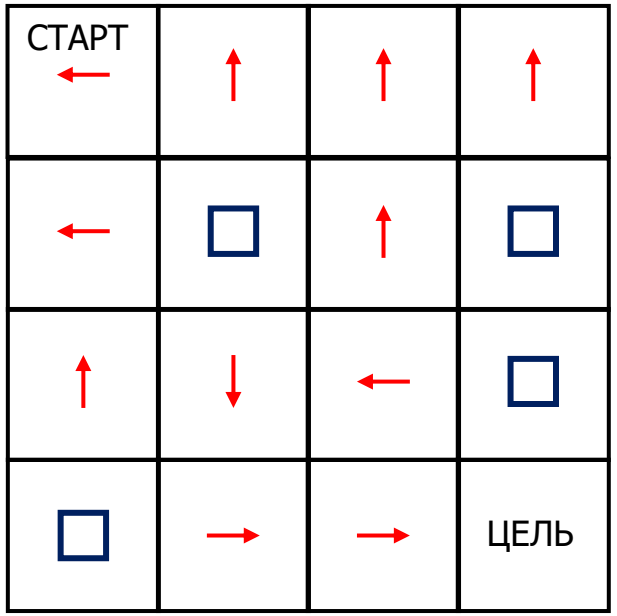


«Осторожная»



Q-функция улучшает политику

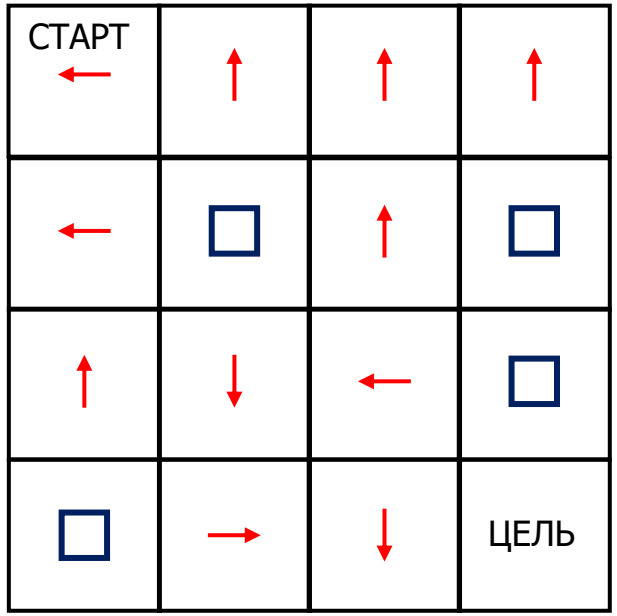
«Осторожная»



Функция ценности действий

0.39 0.41 0.40 0.40	0.38 0.26 0.24 0.25	0.35 0.28 0.27 0.28	0.34 0.23 0.23 0.23
0.27 0.42 0.28 0.29		0.12 0.26 0.26 0.14	
0.45 0.29 0.30 0.31	0.29 0.34 0.34 0.48	0.20 0.43 0.27 0.39	
	0.39 0.35 0.59 0.43	0.67 0.57 0.71 0.76	ЦЕЛЬ

«Осторожная+»



Оптимизация политик

Уравнение оптимизации политик:

$$\pi'(s) = \underset{a}{\operatorname{argmax}} \sum_{s',r} p(s', r|s, a)[r + \gamma v_\pi(s')]$$

«Осторожная»

СТАРТ ← 0.4079	↑ 0.3754	↑ 0.3543	↑ 0.3438
← 0.4263		↑ 0.1169	
↑ 0.4454	↓ 0.4840	← 0.4328	
	→ 0.5884	→ 0.7107	ЦЕЛЬ

Политика после конвергенции

Функция оценки состояния

СТАРТ ← 0.5420	↑ 0.4988	↑ 0.4707	↑ 0.4569
← 0.5585		← 0.3583	
↑ 0.5918	↓ 0.6431	← 0.6152	
	→ 0.7417	↓ 0.8628	ЦЕЛЬ

Разность между V функциями

СТАРТ ← +0.1341	↑ +0.1234	↑ +0.1164	↑ +0.1130
← +0.1381		← +0.2414	
↑ +0.1464	↓ +0.1591	← +0.1824	
	→ +0.1533	↓ +0.1521	ЦЕЛЬ

Алгоритм итерации политик

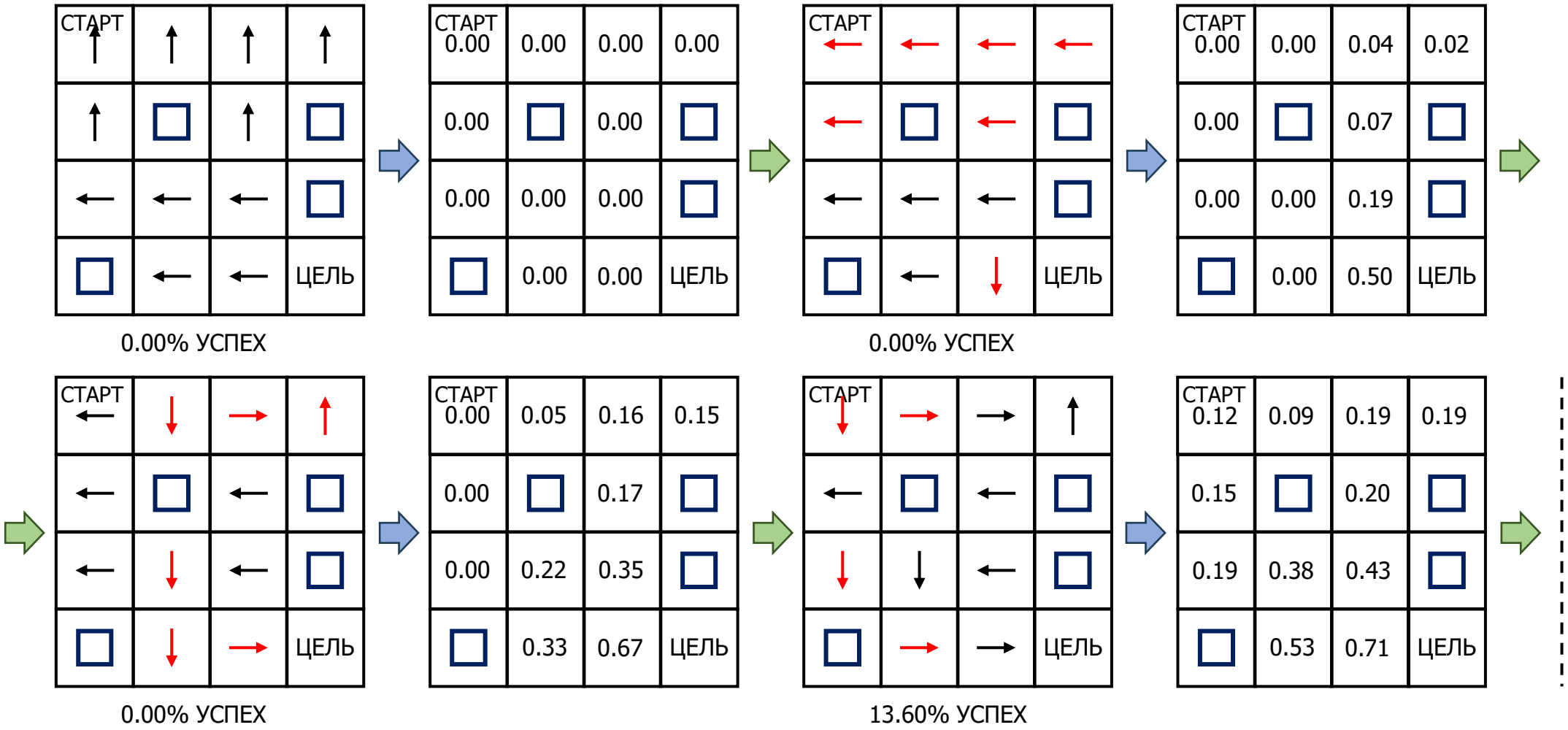
1. Инициализировать случайную политику.

2. Рассчитать функцию состояний для выбранной политики.

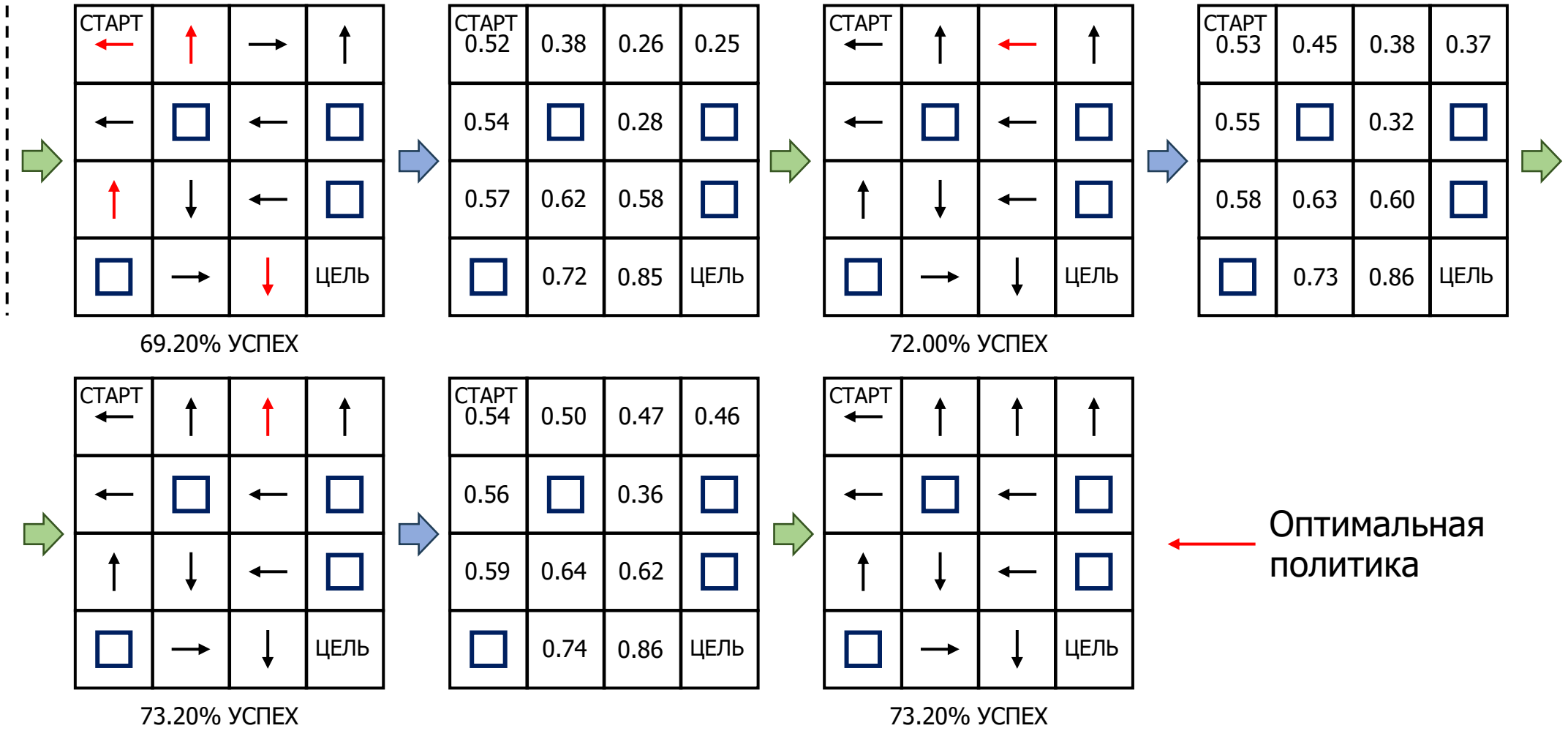
3. Получить новую политику , используя значения функции ценности, полученной на шаге 2.

4. Если новая политика совпадает с предыдущей (использованной на шаге 2), то остановка, в противном случае заменить предыдущую политику новой и перейти на шаг 2.

Итерация политик: Пример. Начало



Итерация политик: Пример. Окончание



Алгоритм итерации ценности

1. Вычисление оптимальной функции ценности путём расчёта максимума по Q функции

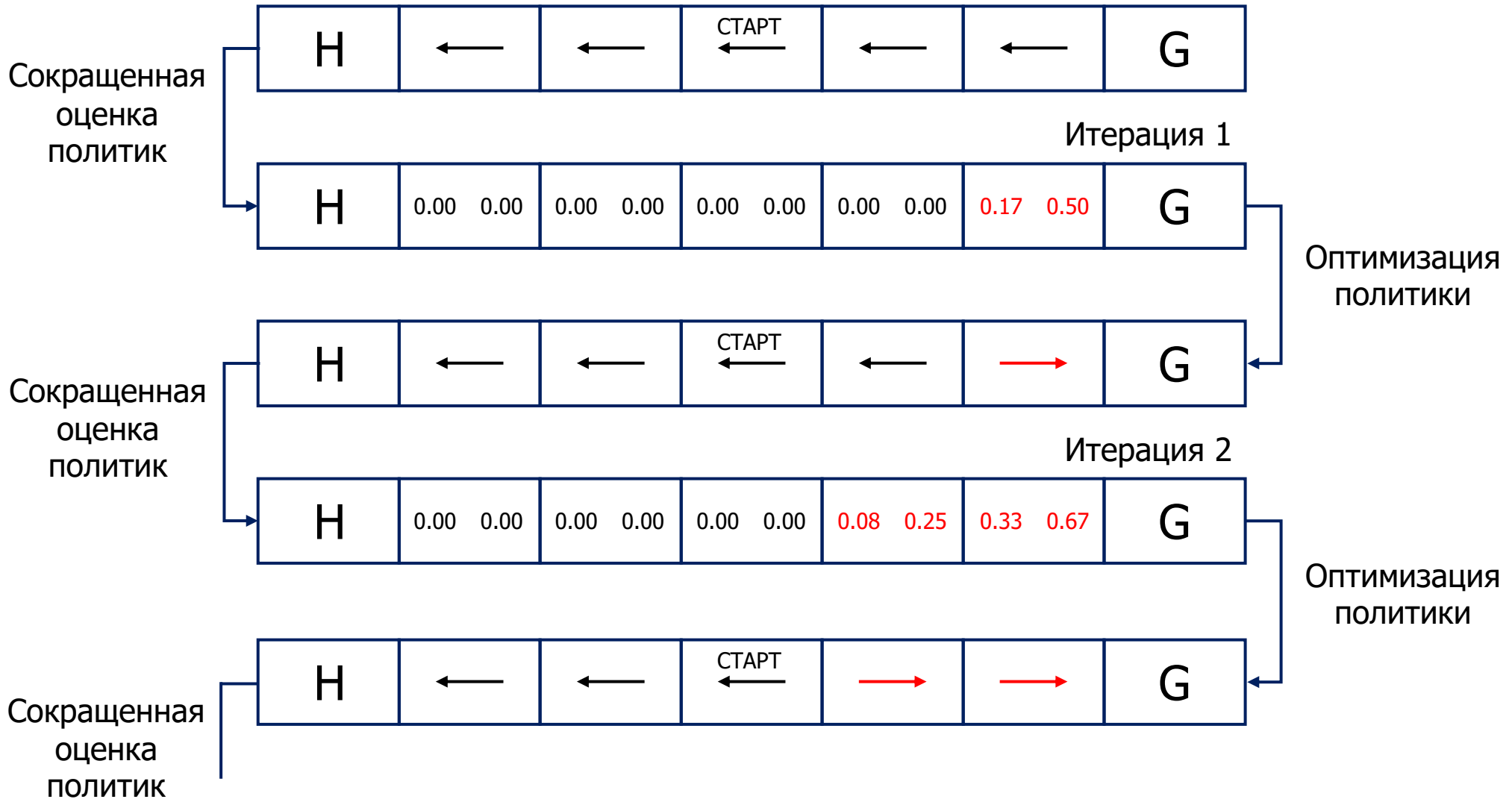
$$V^*(s) = \max_a Q^*(s, a)$$

2. Получить оптимальную политику на основе вычисленной оптимальной функции ценности.

Уравнение итерации ценности:

$$v_{k+1}(s) = \max_a \sum_{s',r} p(s', r|s, a)[r + \gamma v_k(s')]$$

Итерация ценностей: Пример. Начало



Итерация ценностей: Пример. Окончание