

# Обучение с подкреплением для выявления недобросовестных пользователей на краудсорсинговой платформе

Филимохина А.И.

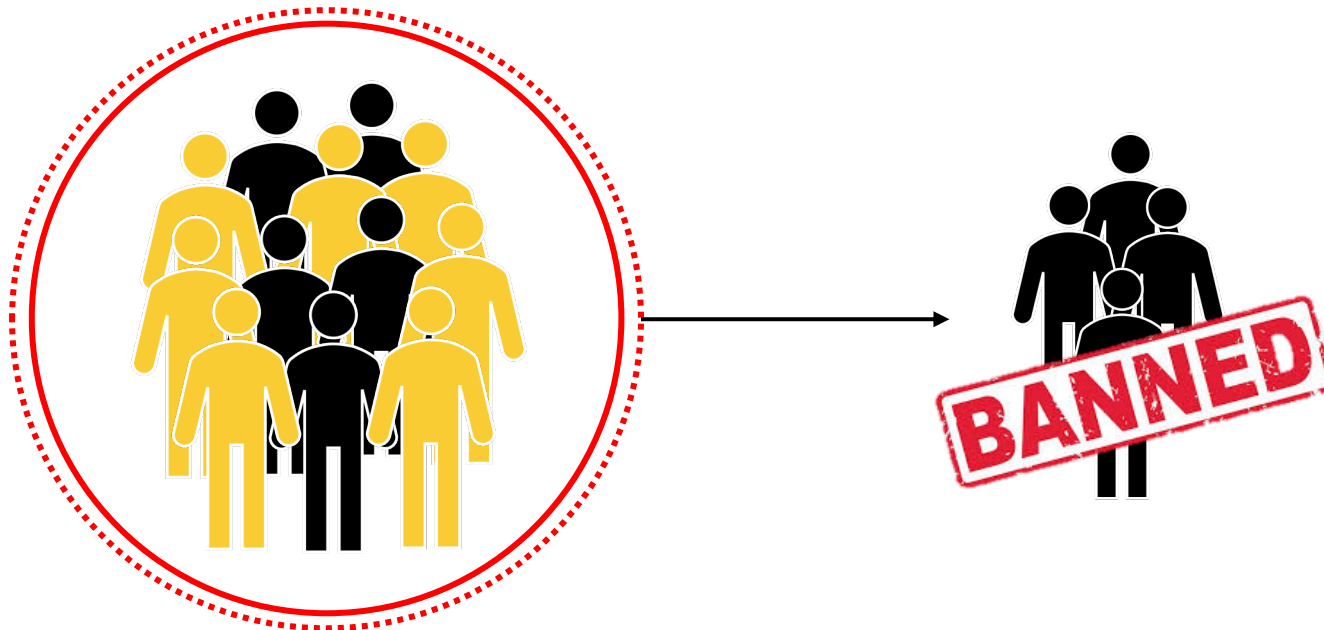
Научный руководитель: Шпильман А.А.

Научный консультант: Свидченко О.А.

Санкт-Петербургская школа физико-математических  
и компьютерных наук

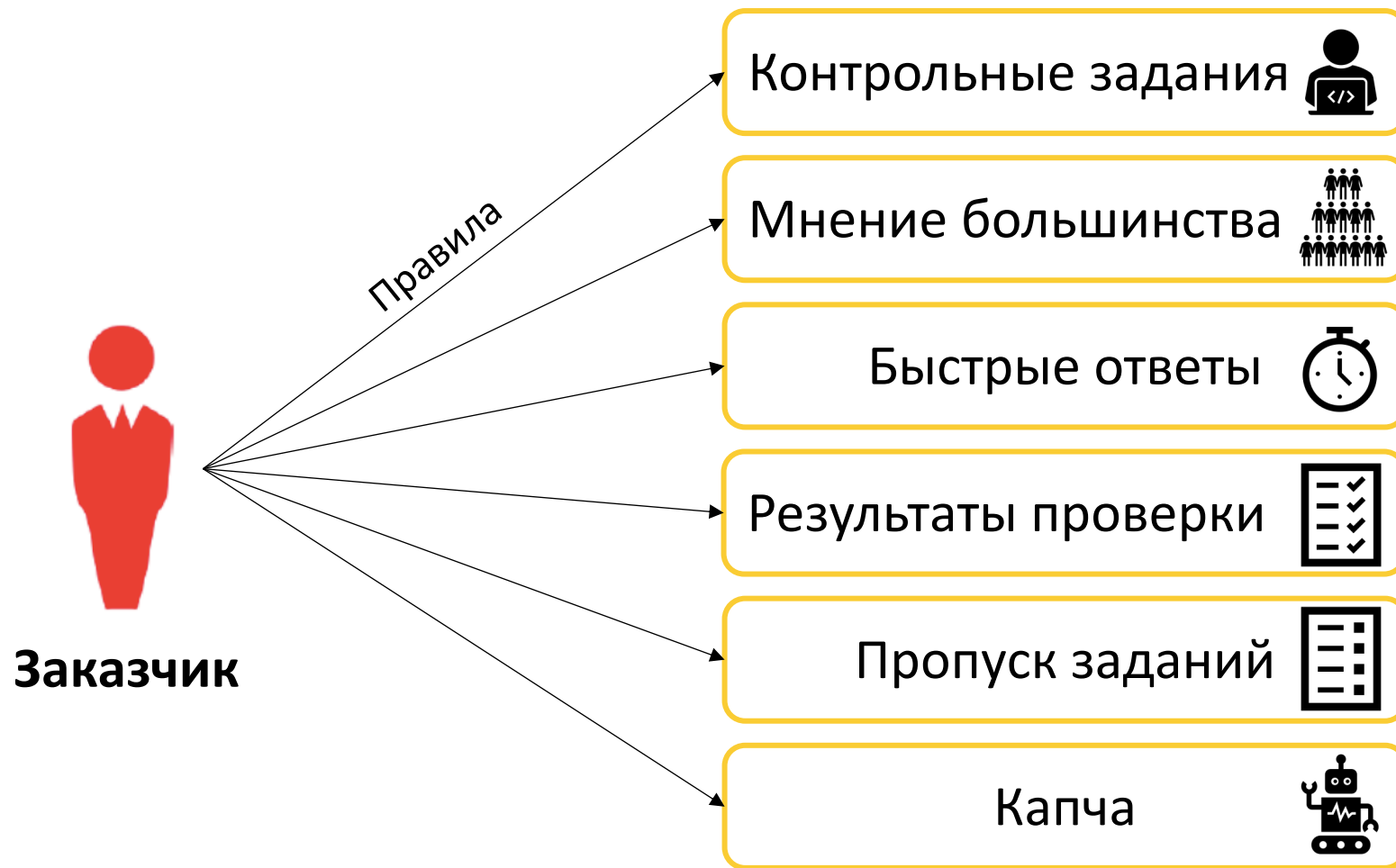
НИУ ВШЭ – Санкт-Петербург  
2020 год

# Задача определения недобросовестных пользователей



# Особенности существующих методов

## Яндекс Толока



## Недостатки



# Обучение с подкреплением



- **Diversity Is All You Need DIAYN<sup>1</sup>**
- **Variational Option Discovery Algorithm VALOR<sup>2</sup>**
- **Dynamic-Aware Unsupervised Discovery DADS<sup>3</sup>**

$\pi(a|s, z)$  - политика агента зависит от **навыка**  $z \in \{1..n\}$

Навык определяет распределение конечного состояния  $p(s_f|s_0, z)$



$p(s_f|s_0) \longrightarrow$  более разнообразное  
 $z$  — определяет достижимые конечные состояния

<sup>1</sup> Benjamin Eysenbach et al. Diversity is All You Need: Learning Skills without a Reward Function, 2018

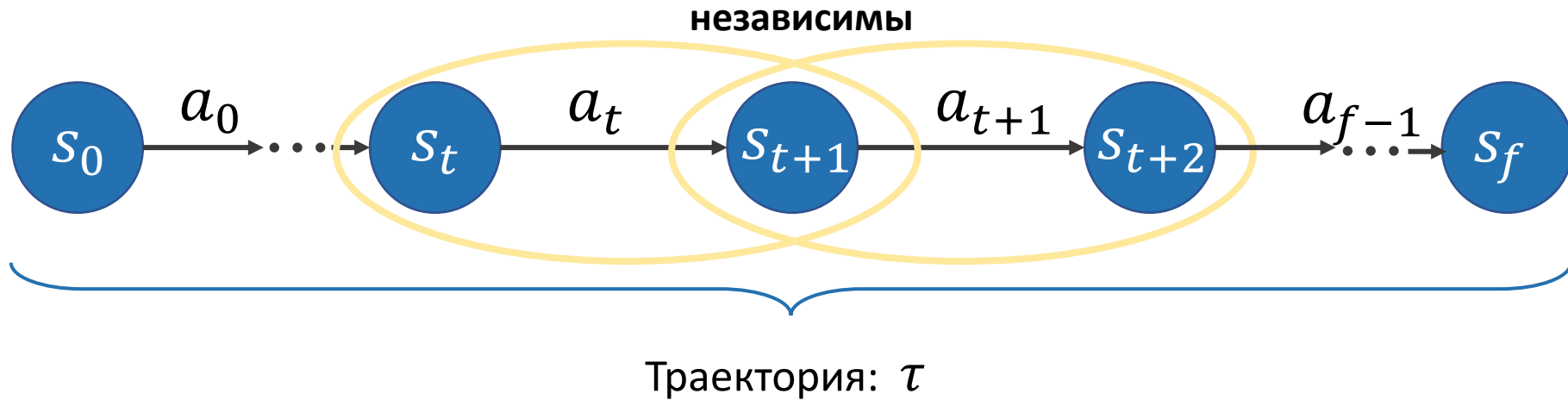
<sup>2</sup> Joshua Achiam et al. Variational Option Discovery Algorithms, 2018

<sup>3</sup> Archit Sharma et al. Dynamics-Aware Unsupervised Discovery of Skills, 2020

# Алгоритмы DIAYN и VALOR

DIAYN

VALOR



неверно

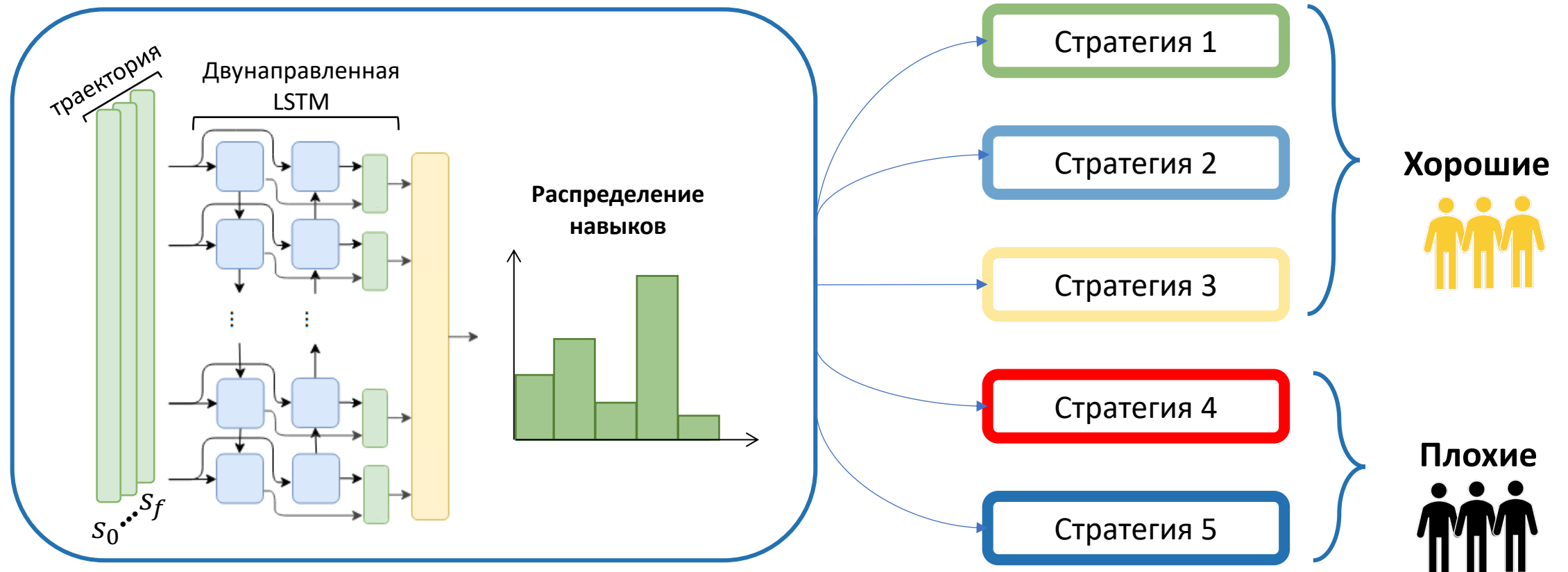
Предположение:  $\mathbb{P}(\tau) = \prod_{t=0}^{f-1} \mathbb{P}(s_t, a_t, s_{t+1})$

# Внутренняя награда и Дискриминатор

$$\text{Внутренняя награда: } r_{\text{внутренняя}} = \frac{\log q(z|s_{t+1})}{\log q(z|\tau)} - \log p^c(z)$$

1 DIAYN  
2 VALOR

## 2 VALOR: Дискриминатор



## 1 DIAYN

## 2 VALOR

- 1: Инициализируем  $\pi, q$
- 2: Повторяем до сходимости:
- 3:     Каждый эпизод семплируем навык  $z \sim p(z)$
- 4:     Проигрываем эпизод со стратегией:  $\pi(a|s, z)$
- 5:     Для
  - 1 → каждого перехода  $(s_t, a_t, s_{t+1})$
  - 2 → всей траектории  $\tau$
- 6:     Обновляем
  - 1 →  $q(z|s_{t+1})$
  - 2 →  $q(z|\tau)$
- 7:     Вычисляем награду:  $r_{\text{внутренняя}} =$ 
  - 1 →  $\log q(z|s_{t+1})$
  - 2 →  $\log q(z|\tau)$ $-\log p^c(z)$
- 8:     Обновляем:
  - 1 →  $\pi(a_t|s_t, z)$
  - 2 →  $\pi(a|s, z)$
 в соответствии с наградой




## Цель:

Создать модель, способную выявлять недобросовестных пользователей на краудсорсинговой платформе, применяя методы обучения с подкреплением



## Задачи:

- Реализовать алгоритмы обучения с подкреплением, нацеленные на исследование агентом различных навыков и стратегий поведения
- Проверить работу алгоритмов на простых средах
- Создать среду, имитирующую выполнения заданий на платформе Яндекс.Толока
- Протестировать алгоритмы на своей среде

# Модификация награды

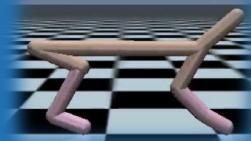
**Внутренняя награда:**  $r_{\text{внутренняя}} \sim \log(\mathbb{P}_{\text{Дискриминатор}})$  

**Комбинирование награды:**  $r = \alpha \cdot r_{\text{внутренняя}} + \beta \cdot r_{\text{среды}}$

$\alpha, \beta$  – настраиваемые гиперпараметры

# Тестирование алгоритма VALOR на HalfCheetah-v2



- Использование комбинированной награды
- Использование внутренней награды

График награды от среды  $r_{\text{среды}}$

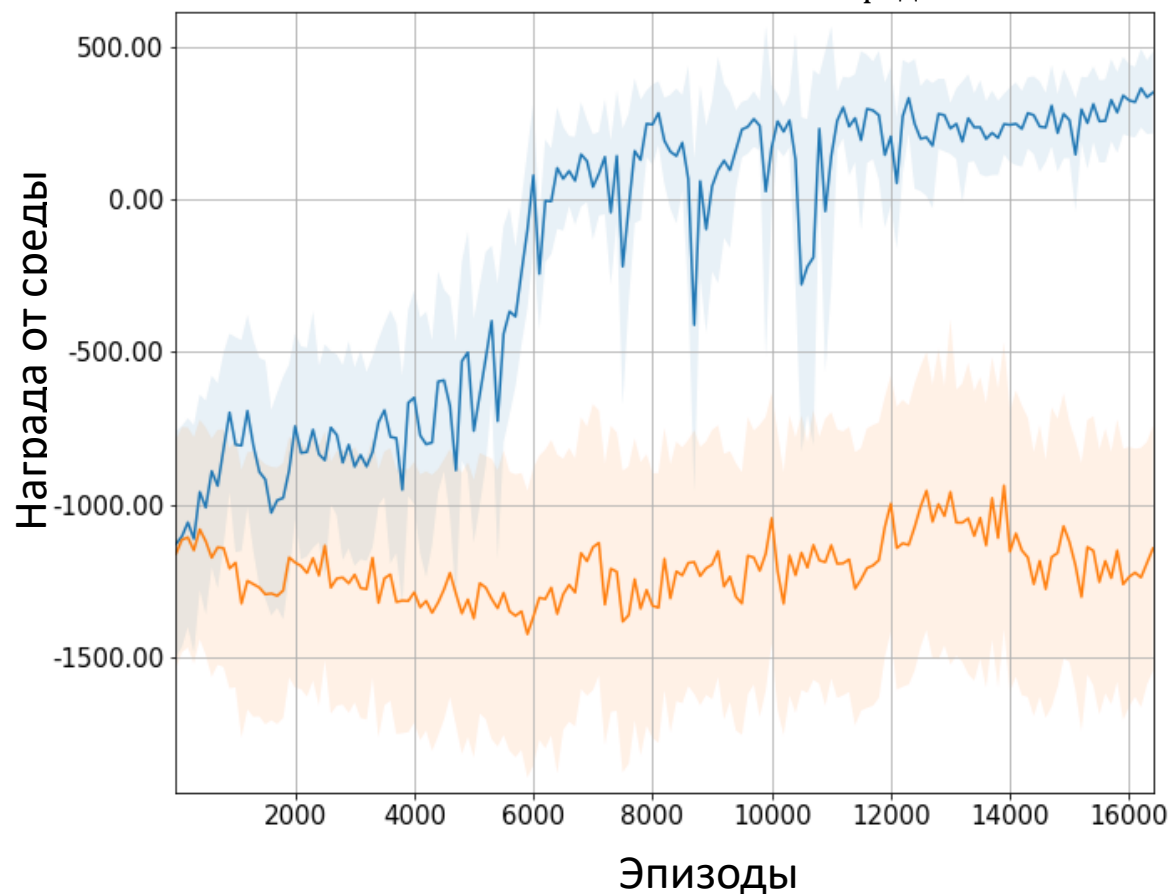
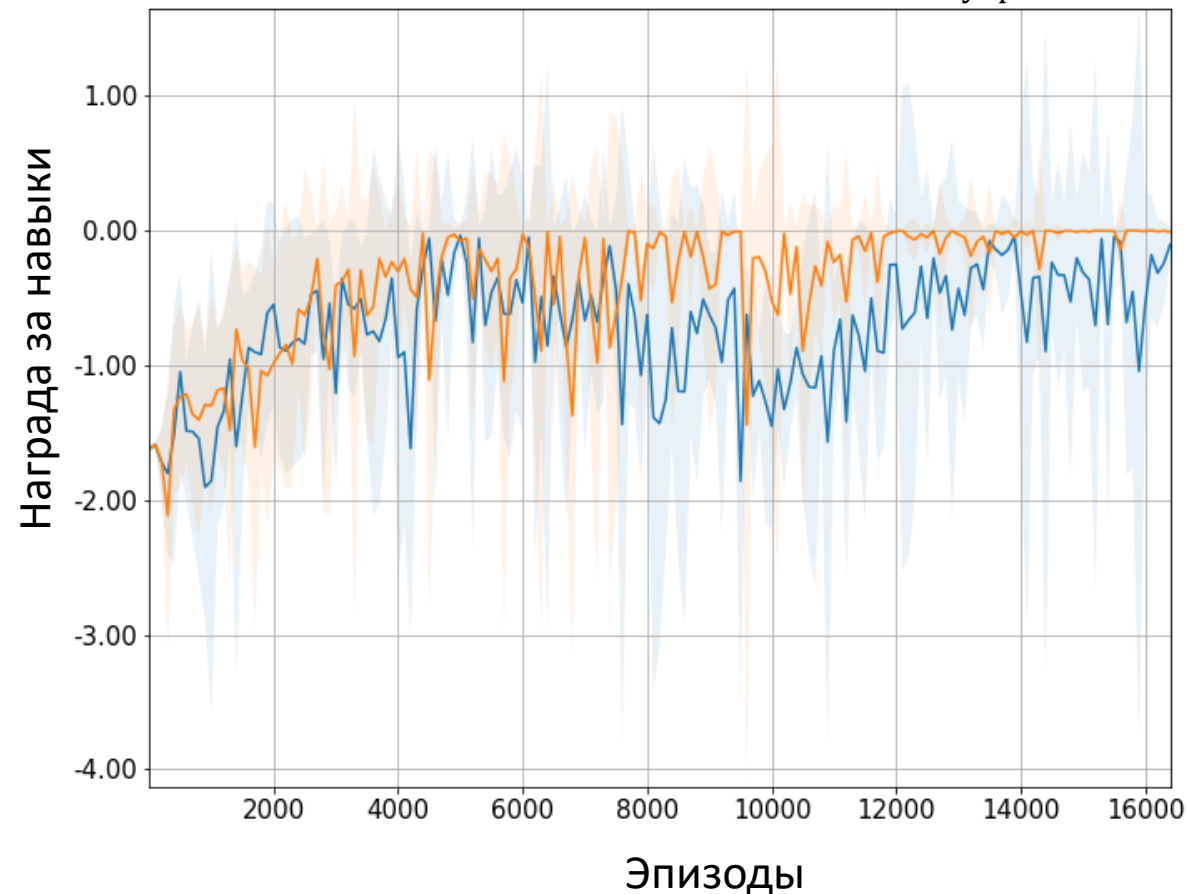


График внутренней награды  $r_{\text{внутренняя}}$



# Тестирование алгоритма DIAYN на MountainCar-v0

Навык 1

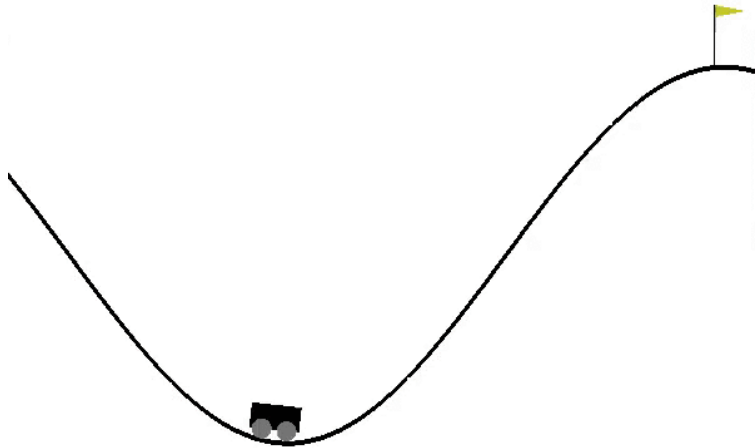
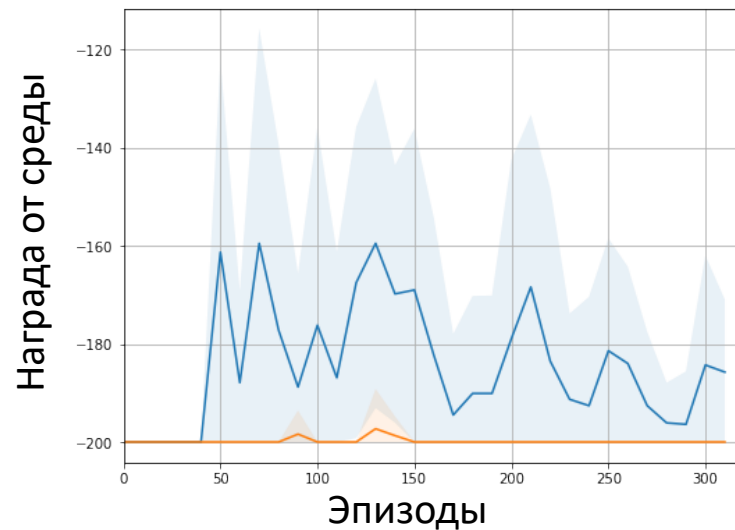
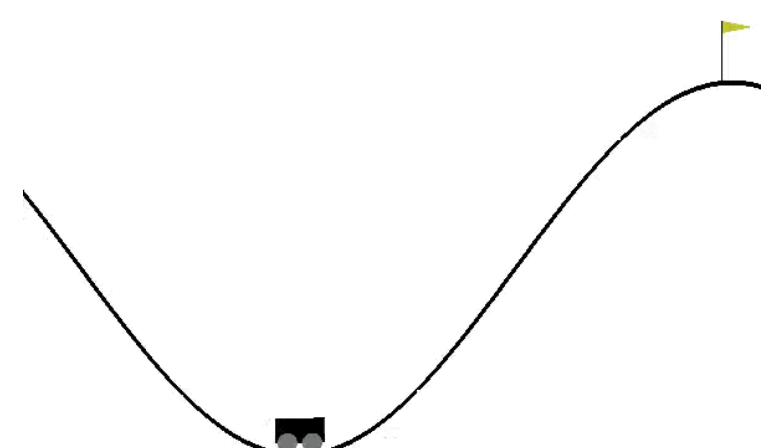


График награды от среды  $r_{\text{среды}}$



Навык 3



Навык 2

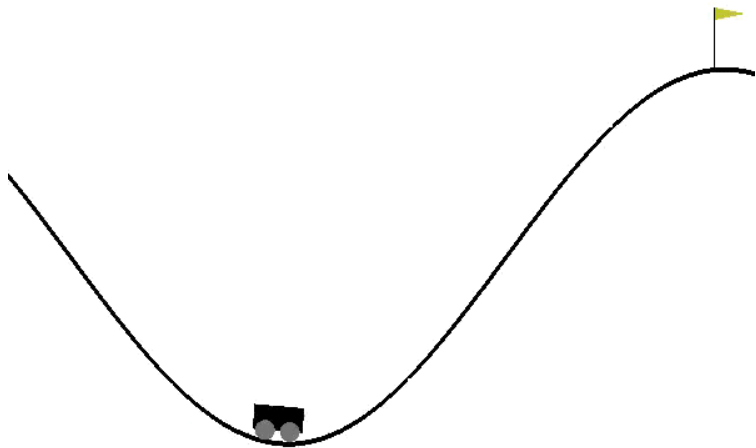
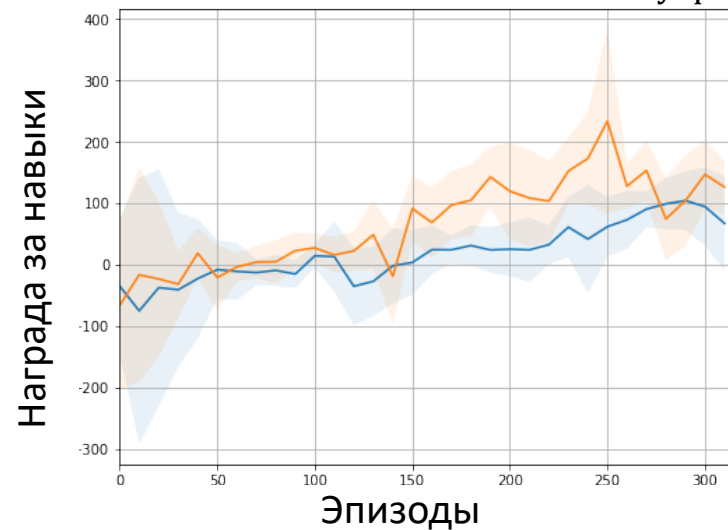
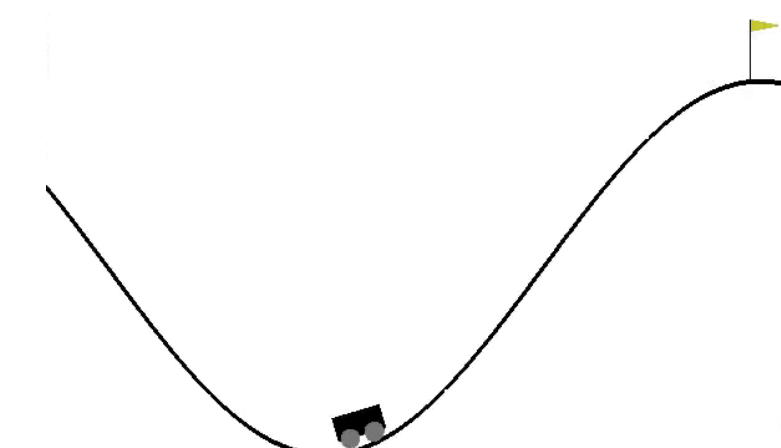


График внутренней награды  $r_{\text{внутренняя}}$



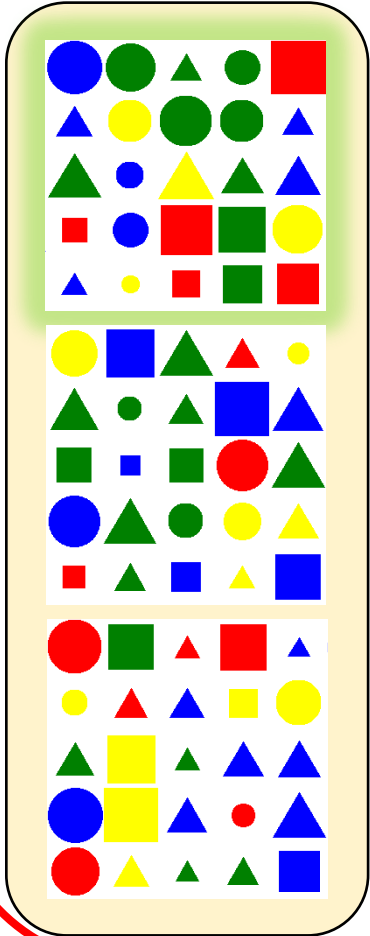
Навык 4



# Создание среды Толока

**Задание:** Выберите картинку на которой расположен самый маленький кружок

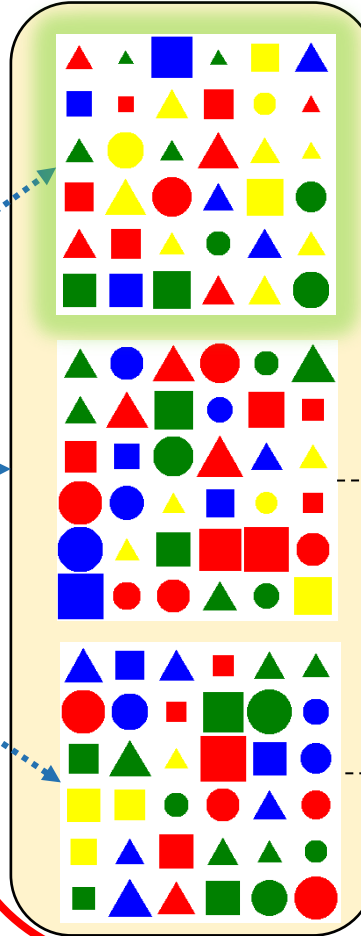
УРОВЕНЬ : 5



Среда  
Яндекс Толока

Рандомно  
генерируется  
новые картинки

УРОВЕНЬ: 6



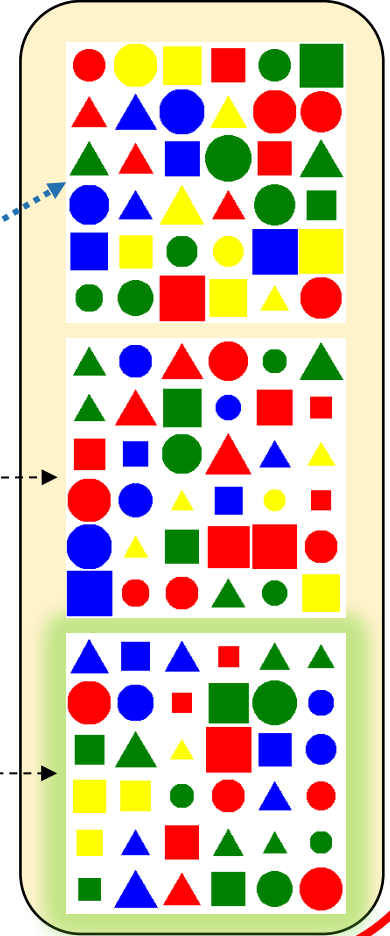
Модифицированная  
среда

Рандомно  
генерируется новая  
картинка

Остается та же картинка

Остается та же картинка

УРОВЕНЬ: 6



# Тестирование алгоритма VALOR на среде Толока

- Использование комбинированной награды
- Использование внутренней награды

График награды от среды  $r_{\text{среды}}$

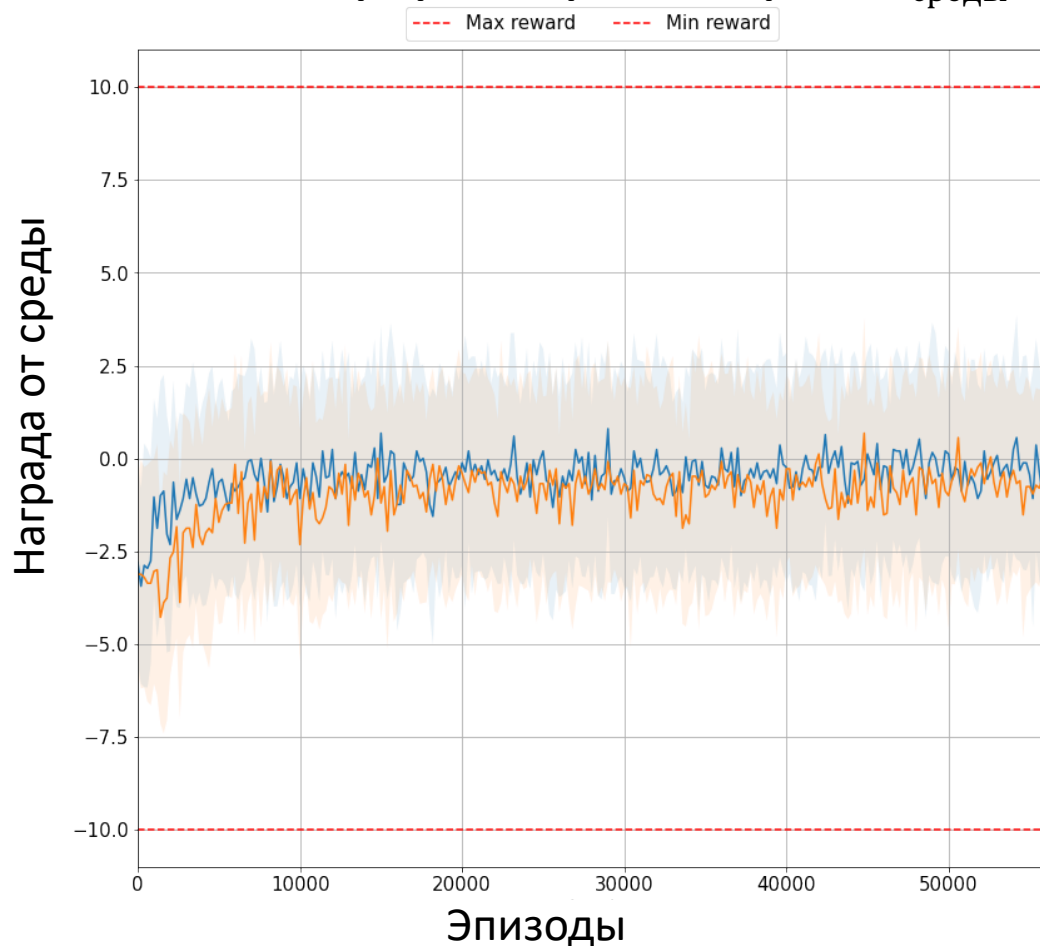
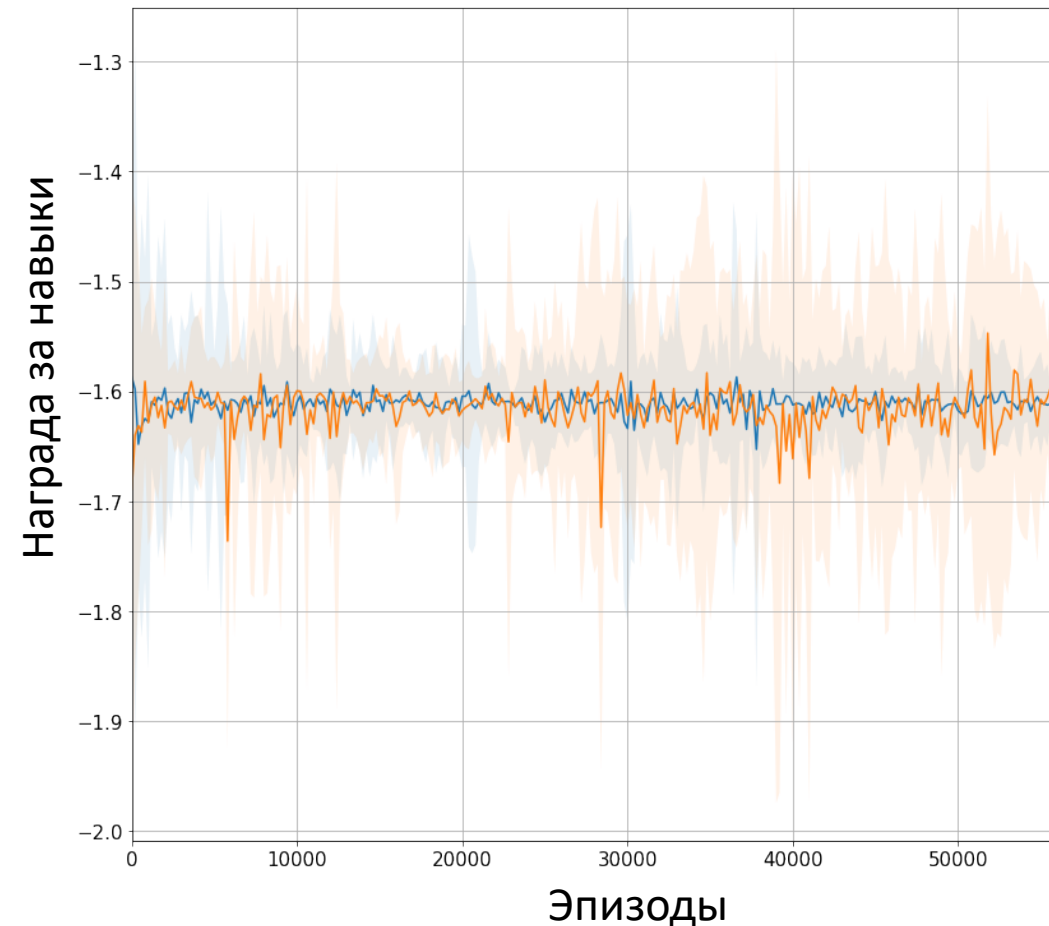


График внутренней награды  $r_{\text{внутренняя}}$



# Тестирование алгоритма DIAYN на среде Толока

- Использование комбинированной награды
- Использование внутренней награды

График награды от среды  $r_{\text{среды}}$

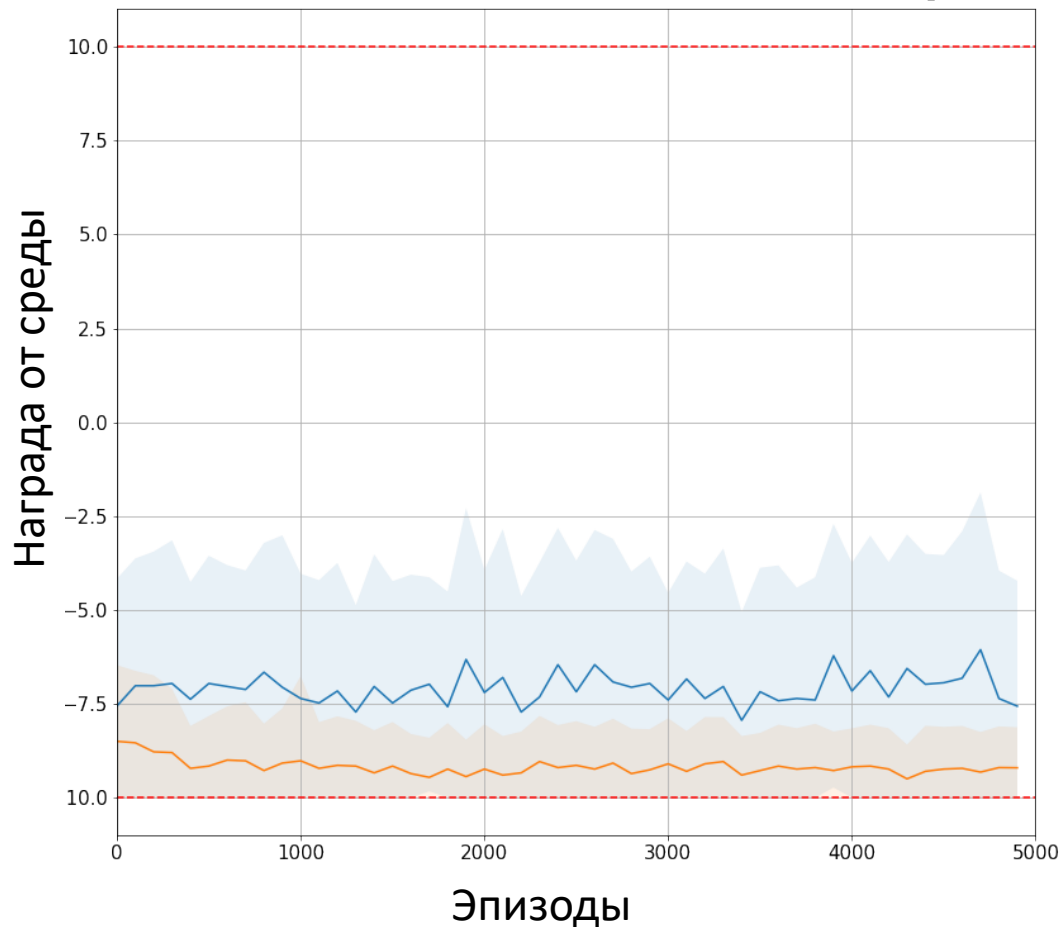
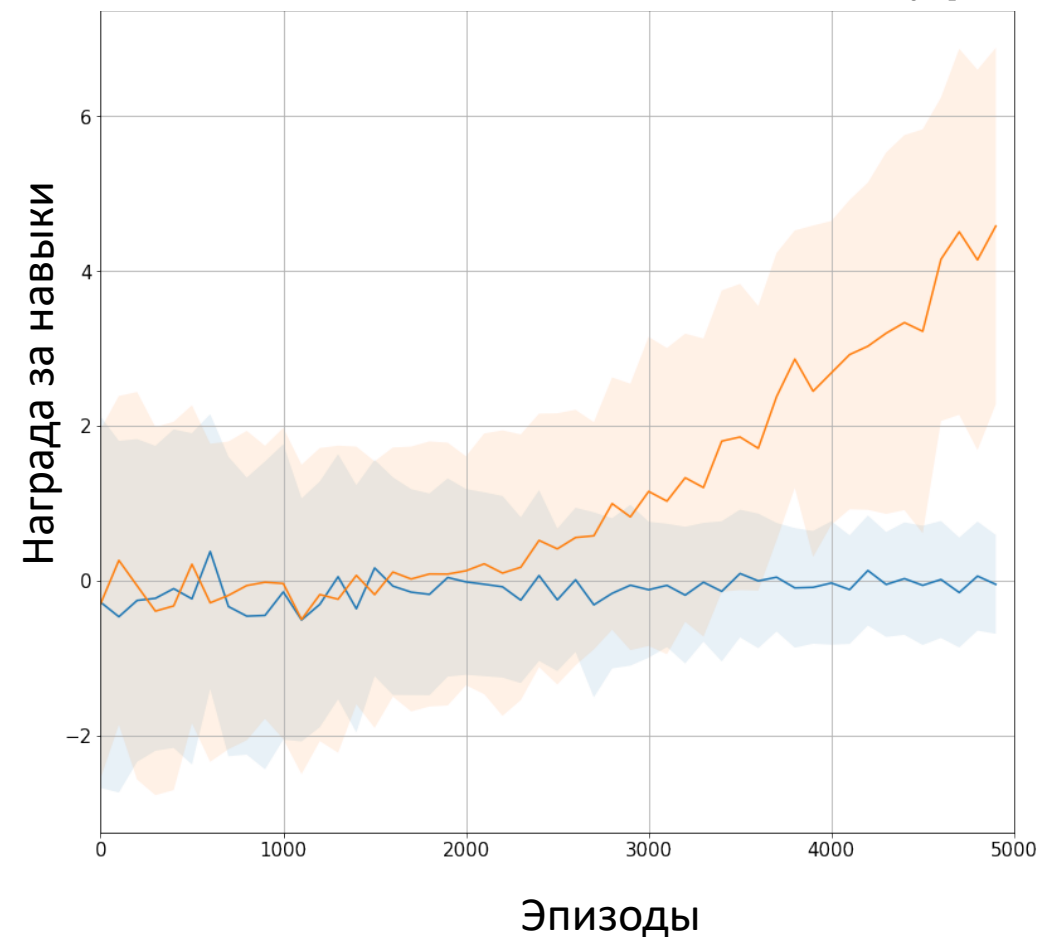


График внутренней награды  $r_{\text{внутренняя}}$



- Реализованы алгоритмы VALOR и DIAYN
- Проверена работа алгоритмов на стандартных средах
- Создана среда, имитирующая взаимодействие исполнителей с Яндекс.Толокой
- Имеющаяся среда адаптирована под особенности каждого из алгоритмов
- Алгоритмы протестированы на среде Толока
- По полученным результатам введены некоторые модификации моделей



- Понять из-за чего алгоритмы не могут выучить различные навыки на нашей среде и устранить эту проблему
- Определить как лучше применять комбинирование награды
- Оценить возможность использования алгоритмов на реальных задачах
- Применить к реальным задачам