
УТВЕРЖДЕНА

Решением Ученого совета
АНО ВО «Центральный университет»
«25» мая 2026 г.
Протокол №2

**Рабочая программа дисциплины (модуля)
«Обучение с подкреплением»**

Направление подготовки: 02.03.01 Математика и компьютерные науки

Направленность (профиль) подготовки: Прикладной искусственный интеллект

Квалификация (степень) выпускника: бакалавр

Форма обучения: очная

Срок освоения программы: 4 года

Год набора: 2026

**Москва
2026**

Содержание

1. Краткая характеристика дисциплины (модуля)	3
2. Перечень планируемых результатов обучения	5
3. Тематический план	10
4. Содержание дисциплины (модуля)	10
5. Учебно-методическое обеспечение	12
6. Материально-техническое обеспечение	12
7. Методические и оценочные материалы	14

1. Краткая характеристика дисциплины (модуля)

Рабочая программа дисциплины (модуля) «Обучение с подкреплением» составлена в соответствии с федеральным государственным образовательным стандартом высшего образования – бакалавриат по направлению 02.03.01 «Математика и компьютерные науки», профиль «Прикладной искусственный интеллект», утвержденный приказом Министерства науки и высшего образования Российской Федерации № 807 от 23.08.2017 года.

Изучение дисциплины (модуля) «Обучение с подкреплением» играет ключевую роль в разработке автономных систем, способных обучаться в сложных и неопределённых средах, таких как робототехника, игры и адаптивные интерфейсы. Полученные знания и навыки позволяют эффективно применять методы обучения с подкреплением в современных задачах искусственного интеллекта, включая управление, оптимизацию и выравнивание генеративных моделей.

Место дисциплины (модуля) в структуре образовательной программы

Настоящая дисциплина (модуль) включена в учебный план по программе подготовки бакалавриата по направлению 02.03.01 «Математика и компьютерные науки», профиль «Прикладной искусственный интеллект» и входит в вариативную часть Блока 1 формируемую участниками образовательных отношений как дисциплина по выбору.

Дисциплина (модуль) изучается на 3 курсе в 5 или 6 семестре на выбор. Доступна к изучению после успешного освоения дисциплины (модуля) «Глубокое обучение».

Цель изучения дисциплины (модуля): сформировать системное понимание теоретических основ и практических методов обучения с подкреплением, а также развить навыки проектирования, реализации и анализа агентов, способных принимать оптимальные решения в динамических средах.

Задачи изучения дисциплины (модуля):

— изучить основные понятия обучения с подкреплением: агент, среда, состояние, наблюдение, действие, награда, эпизод, стратегия и функция ценности, а также формализовать задачи в виде марковского процесса принятия решений;

— освоить уравнения беллмана и принцип оптимальности, лежащие в основе методов динамического программирования и временных разностей;

— научиться различать и применять методы на основе текущей стратегии и другой стратегии, а также реализовывать алгоритмы, основанные на функции ценности и на стратегии, включая метод подкрепления и архитектуру актор-критик;

— изучить архитектуру и методы стабилизации глубоких алгоритмов, такие как сеть с воспроизведением опыта, целевая сеть, двойное q-обучение и сеть с разделённой оценкой, и применять их в средах с высокой размерностью состояний;

— освоить современные подходы к обучению с подкреплением, включая методы с доверительной областью, мягкий актор-критик и их применение в задачах непрерывного управления, а также выравнивание больших языковых моделей по предпочтениям человека.

В результате освоения дисциплины (модуля) обучающийся должен:

знать:

— основные понятия обучения с подкреплением (reinforcement learning, rl): агент, среда, состояние, наблюдение, действие, награда, эпизод, стратегия (policy), функция ценности;

— формальную постановку задачи обучения с подкреплением через марковский процесс принятия решений (markov decision process, mdp);

— уравнения беллмана и принцип оптимальности беллмана;

— различия между методами на основе текущей политики (on-policy) и другой политики (off-policy), на основе функции ценности (value-based) и политики (policy-based);

- основные классические алгоритмы обучения с подкреплением: динамическое программирование, монте-карло методы, td-обучение (td-learning), sarsa, q-обучение (q-learning);

- причины неустойчивости алгоритмов глубокого обучения с подкреплением (deep rl) и методы стабилизации обучения;

- архитектуру и основные идеи deep q-network (dqn): буфер воспроизведения (replay buffer), целевая сеть (target network), double dqn и dueling dqn;

- подходы на основе градиента политики: reinforce, функция преимущества (advantage function), актор-критик (actor-critic), gae;

- принципы методов с доверительной областью (trust-region): trpo и ppo;

- подходы к непрерывному управлению (continuous control) и алгоритмы типа soft actor-critic (sac);

- основы мультиагентного обучения с подкреплением;

- подходы rlhf и dpo для больших языковых моделей человеческим предпочтениям;

- связь обучения с подкреплением и задач оптимального управления: линейно-квадратичный регулятор, линейно-гауссовский регулятор, фильтр калмана;

- области применения обучения с подкреплением: игры, робототехника, рекомендательные системы, управление ресурсами, финансы, большие языковые модели.

уметь:

- формализовывать задачи в виде марковского процесса принятия решений;

- объяснять связь обучения с подкреплением с оптимальным управлением и классическими методами теории управления;

- реализовывать и запускать базовые алгоритмы;

- обучать и анализировать агентов в симуляционных средах;

- сравнивать различные подходы и выбирать подходящий алгоритм под задачу;

- настраивать базовые гиперпараметры моделей;

- анализировать сходимость и устойчивость обучения агентов;

- использовать нейросетевые модели для приближения функций ценности и политик;

- применять актор-критик и ppo-подобные подходы для решения задач управления;

- интерпретировать результаты обучения агента и диагностировать типичные проблемы обучения;

- работать с современными библиотеками и инструментами для глубокого обучения с подкреплением.

владеть:

- ориентация в основных направлениях современного обучения с подкреплением и примерах их практического применения;

- реализация и тестирование классических и глубоких алгоритмов обучения с подкреплением;

- анализ качества обучения агентов и интерпретация поведения стратегий;

- использование нейросетевых моделей для решения задач обучения с подкреплением.

2. Перечень планируемых результатов обучения

Компетенции, формируемые в результате освоения дисциплины (модуля) при проведении учебных занятий в форме контактной работы обучающихся с педагогическими работниками Университета и в форме самостоятельной работы обучающихся:

Код и содержание компетенции	Индикатор компетенции и перечень планируемых результатов обучения по дисциплине (модулю)
ППК-Р4. Способен использовать базы данных при создании программных модулей и компонентов	ППК-Р4.1. Пишет программный код с использованием языков определения и манипулирования данными в базах данных (базовый) ППК-Р4.1. 3-1. Знает архитектуры современных систем управления баз данных, включая SQL и noSQL ППК-Р4.2. У-3. Умеет использовать методы и средства выбранного языка программирования для работы с базами данных
	ППК-Р4.2. Проектирует базы данных для программных модулей и компонентов (средний) ППК-Р4.2. 3-1. Знает нормативно-технические документы (стандарты и регламенты), определяющие требования к оформлению моделей баз данных выбранной архитектуры ППК-Р4.2. У-1. Умеет выбирать тип базы данных в зависимости от решаемой задачи
	ППК-Р4.3. Оптимизирует производительность работы с базами данных (средний) ППК-Р4.3. У-1. Умеет применять методы и средства мониторинга производительности запросов к базе данных ППК-Р4.3. У-2. Умеет вырабатывать варианты оптимизации производительности запросов в базе данных
ППК-Р5. Способен разрабатывать требования и проектировать программное обеспечение	ППК-Р5.1. Анализирует возможности реализации требований к компьютерному программному обеспечению (средний) ППК-Р5.1. 3-3. Знает методологии разработки компьютерного программного обеспечения и технологии программирования ППК-Р5.1. У-2. Умеет выявлять взаимосвязи и документировать требования к компьютерному программному обеспечению
	ППК-Р5.2. Разрабатывает технические спецификации на программные компоненты и их взаимодействие (средний) ППК-Р5.2. 3-2. Знает методы и средства проектирования программных интерфейсов ППК-Р5.2. У-2. Умеет вырабатывать варианты реализации компьютерного программного обеспечения
	ППК-Р5.3. Проектирует компьютерное программное обеспечение (продвинутый) ППК-Р5.3. 3-2. Знает типовые решения, библиотеки программных модулей, шаблоны, классы объектов, используемые при разработке компьютерного программного обеспечения ППК-Р5.3. У-3. Умеет проектировать программные интерфейсы
	ППК-Р5.4. Выполняет логическое проектирование системы (продвинутый) ППК-Р5.4. 3-1. Знает устройство и функционирование ИТ-систем/продуктов ППК-Р5.4. У-1. Умеет декомпозировать ИТ-системы и ИТ-продукты на подсистемы и элементы поставки

ППК-Р6. Способен участвовать в промышленной разработке программного обеспечения	ППК-Р6.1. Работает в соответствии с промышленными методологиями разработки (средний) ППК-Р6.1. 3-1. Знает принципы Agile и их применение в промышленных проектах ППК-Р6.1. У-2. Умеет работать в команде с использованием инструментов управления проектами
	ППК-Р6.2. Использует инструменты промышленной разработки (средний) ППК-Р6.2. 3-1. Знает принципы Continuous Integration and Continuous Delivery (CI/CD) ППК-Р6.2. У-2. Умеет настраивать мониторинг в продуктивной среде
	ППК-Р6.3. Разрабатывает масштабируемый и поддерживаемый код (продвинутый) ППК-Р6.3. 3-3. Знает паттерны проектирования и антипаттерны ППК-Р6.3. У-3. Умеет проводить рефакторинг для повышения качества кода
	ППК-Р6.4. Участвует в развертывании и поддержке программного обеспечения (продвинутый) ППК-Р6.4. 3-1. Знает стратегии развертывания промышленного программного обеспечения ППК-Р6.4. У-2. Умеет обнаруживать и устранять инциденты с работой продуктивного программного обеспечения
ППК-Р7. Способен применять искусственный интеллект (ИИ) для генерации и отладки программного кода	ППК-Р7.1. Применяет ИИ-инструменты для генерации программного кода (базовый) ППК-Р7.1. 3-1. Знает принципы работы современных генеративных ИИ-моделей для генерации кода ППК-Р7.1. У-1. Умеет формулировать корректные текстовые запросы (промты) для генерации кода
	ППК-Р7.2. Использует ИИ для анализа и отладки кода (средний) ППК-Р7.2. 3-1. Знает методы ИИ-анализа кода ППК-Р7.2. У-2. Умеет интерпретировать рекомендации ИИ по исправлению кода
	ППК-Р7.3. Оптимизирует код с помощью ИИ (продвинутый) ППК-Р7.3. 3-1. Знает методы ИИ-оптимизации ППК-Р7.3. У-2. Умеет проверять корректность оптимизаций, предложенных ИИ
	ППК-Р7.4. Оценивает этические и профессиональные аспекты применения ИИ в разработке (продвинутый) ППК-Р7.4. 3-1. Знает этические нормы использования ИИ (конфиденциальность, плагиат кода и т.п.) ППК-Р7.4. У-2. Умеет документировать использование ИИ в разработке
ППК-ДА4. Способен выявлять бизнес-проблемы или бизнес-возможности	ППК-ДА4.1. Собирает информацию о бизнес-проблемах или бизнес-возможностях (средний) ППК-ДА4.1. 3-1. Знает языки и инструменты визуального моделирования ППК-ДА4.1. У-3. Умеет оформлять результаты бизнес-анализа в соответствии с выбранными подходами
	ППК-ДА4.2. Выявление истинных бизнес-проблем или бизнес-возможностей (продвинутый) ППК-ДА4.2. 3-3. Знает перспективные и существующие цифровые технологии и цифровые возможности для бизнеса в контексте предметной области и специфики деятельности организации ППК-ДА4.2. У-1. Умеет представлять информацию о выявленных бизнес-проблемах или бизнес-возможностях

	различными способами и в различных форматах для обсуждения с заинтересованными сторонами
ППК-ДА5. Способен применять методы статистического анализа и теорию эксперимента для планирования, проведения и интерпретации результатов экспериментов	ППК-ДА5.1. Разрабатывает дизайн эксперимента (базовый) ППК-ДА5.1. 3-1. Знает основные принципы планирования экспериментов (рандомизация, контрольные группы, мощность теста) ППК-ДА5.1. У-2. Умеет выбирать метрики для оценки эффекта воздействия
	ППК-ДА5.2. Проводит статистический анализ данных эксперимента (средний) ППК-ДА5.2. 3-1. Знает методы проверки гипотез (t-тест, χ^2 , ANOVA) ППК-ДА5.2. У-2. Умеет интерпретировать p-value и уровень значимости
	ППК-ДА5.3. Интерпретирует результаты экспериментов (продвинутый) ППК-ДА5.3. 3-2. Знает ограничения и риски некорректной интерпретации (ложные положительные/отрицательные результаты) ППК-ДА5.3. У-2. Умеет формулировать рекомендации на основе статистических выводов
МО-2. Способен проектировать и внедрять системы автоматизации жизненного цикла машинного обучения (MLOps)	МО-2.1. Проектирует и настраивает CI/CD-пайплайны для машинного обучения (базовый) МО-2.1. 3-1. Знает принципы CI/CD применительно к моделям машинного обучения (тестирование, сборка, развертывание) МО-2.1. У-1. Умеет настраивать CI/CD-пайплайны для МО с использованием инструментов
	МО-2.2. Организует версионирование моделей, данных и экспериментов (средний) МО-2.2. 3-1. Понимает системы версионирования моделей и данных МО-2.2. У-2. Умеет использовать инструменты MLOps для воспроизводимости и масштабируемости экспериментов
	МО-2.3. Внедряет системы мониторинга моделей в продуктивную эксплуатацию (средний) МО-2.3. 3-1. Знает метрики мониторинга моделей МО МО-2.3. У-1. Умеет развертывать системы мониторинга моделей МО
	МО-2.4. Автоматизирует процессы переобучения и развертывания моделей (продвинутый) МО-2.4. 3-1. Знает стратегии автоматизации переобучения МО-2.4. У-1. Умеет автоматизировать переобучение моделей и А/В-тестирование развертываний
РГ-1. Способен руководить процессами разработки компьютерного программного обеспечения	РГ-1.1. Руководит разработкой программного кода (средний) РГ-1.1. 3-2. Знает основные принципы и методы управления персоналом РГ-1.1. У-3. Оценивает качество и эффективность программного кода
	РГ-1.2. Руководит проверкой работоспособности компьютерного программного обеспечения (средний) РГ-1.2. 3-2. Знает методы и средства проверки работоспособности компьютерного программного обеспечения РГ-1.2. У-1. Умеет распределять задачи на проверку работоспособности компьютерного программного обеспечения между исполнителями

	РГ-1.3. Руководит интеграцией программных модулей и компонентов (продвинутый) РГ-1.3. 3-3. Знает методы проектирования и разработки программных интерфейсов взаимодействия внутренних модулей компьютерного программного обеспечения РГ-1.3. У-2. Умеет оценивать результаты выполнения назначенных заданий на разработку процедур интеграции, сборку, подключение к внешней среде, проверку работоспособности выпусков программного продукта
РГ-2. Способен организовать процессы разработки компьютерного программного обеспечения	РГ-2.1. Управляет проектированием компьютерного программного обеспечения (средний) РГ-2.1. 3-1. Знает принципы построения архитектуры компьютерного программного обеспечения и виды архитектуры программного обеспечения РГ-2.1. У-2. Умеет распределять задания на проектирование компьютерного программного обеспечения, структуры базы данных, программных интерфейсов
	РГ-2.2. Управляет процессом разработки компьютерного программного обеспечения (продвинутый) РГ-2.2. 3-2. Знает методы оценки качества плана разработки программного продукта (ресурсы, сроки, риски) РГ-2.2. У-3. Умеет наблюдать за исполнением планов разработки программного продукта
	РГ-2.3. Управляет запросами на изменения, дефектами и проблемами (продвинутый) РГ-2.3. 3-3. Знает методы планирования и документирования вносимых изменений в компьютерное программное обеспечение РГ-2.3. У-2. Умеет принимать управленческие решения о реализации запросов на изменения (решений о необходимости и сроках внесения изменений в программное обеспечение и документацию)
	РГ-2.4. Управляет конфигурациями и выпусками программного продукта (продвинутый) РГ-2.3. 3-2. Знает состав и методы использования коллективной среды разработки компьютерного программного обеспечения и системы контроля версий РГ-2.5. У-3. Умеет определять перечень функциональных требований, реализуемых в новой версии программного продукта
РГ-4. Способен организовать промышленную разработку программного обеспечения	РГ-4.1. Внедряет промышленные методологии разработки (средний) РГ-4.1. 3-1. Знает принципы Agile и их применение в промышленных проектах РГ-4.1. У-3. Умеет организовать работу в команде с использованием инструментов управления проектами
	РГ-4.2. Внедряет инструменты промышленной разработки (средний) РГ-4.2. 3-1. Знает принципы Continuous Integration and Continuous Delivery (CI/CD) РГ-4.2. У-3. Умеет организовать мониторинг в продуктивной среде
	РГ-4.3. Организует разработку масштабируемого и поддерживаемого кода (продвинутый) РГ-4.3. 3-3. Знает паттерны проектирования и антипаттерны РГ-4.3. У-3. Умеет управлять техническим долгом, в том числе организовать рефакторинг для повышения качества кода
	РГ-4.4. Организует развертывание и поддержку программного

	обеспечения (продвинутый) РГ-4.4. 3-1. Знает стратегии развертывания промышленного программного обеспечения РГ-4.4. У-2. Умеет управлять обнаружением и устранением инцидентов с работой продуктивного программного обеспечения
РГ-5. Способен организовать процессы применения искусственного интеллекта (ИИ) для генерации и отладки программного кода	РГ-5.1. Организует применение ИИ-инструментов для генерации программного кода (средний) РГ-5.1. 3-1. Знает принципы работы современных генеративных ИИ-моделей для генерации кода РГ-5.1. У-2. Умеет организовывать процесс применения ИИ для генерации программного кода
	РГ-5.2. Организует применение ИИ для анализа и отладки кода (продвинутый) РГ-5.2. 3-1. Знает методы ИИ-анализа кода РГ-5.2. У-3. Умеет организовать применение ИИ для проведения code review
	РГ-5.3. Организует оптимизацию кода с помощью ИИ (продвинутый) РГ-5.3. 3-1. Знает методы ИИ-оптимизации РГ-5.3. У-2. Умеет управлять проверкой корректности оптимизаций, предложенных ИИ
	РГ-5.4. Оценивает этические и профессиональные аспекты применения ИИ в разработке (продвинутый) РГ-5.4. 3-1. Знает этические нормы использования ИИ (конфиденциальность, плагиат кода и т.п.) РГ-5.4. У-2. Умеет организовать процесс документирования использования ИИ в разработке
УК-4. (УНК-01.) Способен осуществлять деловую коммуникацию в устной и письменной формах на государственном языке Российской Федерации и иностранном(ых) языке(ах) <i>(соответствует УНК-01 – Способен применять коммуникационные компетенции в профессиональной деятельности)</i>	УНК-01.01. Осуществляет коммуникации в профессиональной деятельности (базовый) УНК-01.01.01. У-4 Умеет адаптироваться к изменениям и неопределенностям в работе УНК-01.01.01. У-5 Способен конструктивно воспринимать критику
	УНК-01.02. Демонстрирует владение профессиональной культурой (базовый) УНК-01.02.01. У-2. Способен поддерживать уважительные и конструктивные отношения с коллегами, клиентами и партнерами УНК-01.02.01. У-4. Способен демонстрировать аккуратность, пунктуальность и ответственность в выполнении своих обязанностей
УК-3. (УНК-02.) Способен осуществлять социальное взаимодействие и реализовывать свою роль в команде <i>(соответствует УНК-02 – Способен применять методы коллективной работы в профессиональной деятельности)</i>	УНК-02.01. Принимает участие в групповом взаимодействии (базовый) УНК-02.01.01. У-2. Умеет ясно выражать свои мысли в группе УНК-02.01.01. У-3. Способен слушать и учитывать мнения других участников команды
	УНК-02.02. Проявляет лидерство и осуществляет наставничество (базовый) УНК-02.02.01. У-2. Способен организовывать работу группы, распределять обязанности и контролировать выполнение задач УНК-02.02.01. У-4. Способен принимать ответственные решения и демонстрировать пример профессионализма

3. Тематический план

№п/ п	Наименование раздела дисциплины (модуля)	Трудоемкость, академические часы				ТКУ (текущий контроль успеваемости)
		Очная форма				
		Аудиторная работа		Контроль	Самостоятельная работа	
		Лекции	Практические занятия			
1	Введение в обучение с подкреплением	4	4		16	Домашнее задание
2	Классическая теория обучения с подкреплением	4	4		16	Домашнее задание
3	Глубокое обучение с подкреплением	4	4		16	Домашнее задание
4	Алгоритмы градиента политики (Policy Gradient) и Актор-Критик (Actor-Critic)	8	8	2	34	Домашнее задание Контрольная работа
5	Алгоритмы непрерывного управления (Continuous Control) и современные применения обучения с подкреплением	10	10	2	40	Домашнее задание Контрольная работа Кейс
	Экзамен			4		
Итого:		30	30	8	122	
Объем дисциплины (модуля) (в ак. ч.)		190				
Объем дисциплины (модуля) (в зач. ед.)		5				

4. Содержание дисциплины (модуля)

№п/п	Наименование раздела дисциплины (модуля)	Содержание дисциплины (модуля) по темам
1	Введение в обучение с подкреплением	Введение в обучение с подкреплением Формальная постановка обучения с подкреплением
2	Классическая теория обучения с подкреплением	Динамическое программирование Табличное обучение с подкреплением
3	Глубокое обучение с подкреплением	Введение в глубокое обучение с подкреплением. Алгоритм DQN Модификации алгоритма DQN
4	Алгоритмы градиента политики (Policy Gradient) и Актор-Критик (Actor-Critic)	Алгоритмы на основе градиента политики (Policy Gradient) Актор-Критик (Actor-Critic) алгоритмы Алгоритмы на основе доверительной области (trust-region) Алгоритм Proximal Policy Optimization (PPO)
5	Алгоритмы непрерывного управления (Continuous Control) и современные	Алгоритмы непрерывного управления (Continuous Control) Мультиагентное обучение с подкреплением Алгоритмы для дообучения больших языковых моделей Обучение с подкреплением и оптимальное управление

	применения обучения с подкреплением	
--	--	--

5. Учебно-методическое обеспечение

Университет располагает полным набором лицензионного и свободно распространяемого программного обеспечения, включая продукты отечественного производства.

Каждый студент в течение всего периода обучения получает индивидуальный неограниченный доступ к электронно-библиотечной системе и электронной информационно-образовательной среде университета. Эти системы предоставляют возможность доступа к ресурсам из любой точки, где есть подключение к сети Интернет, как на территории университета, так и за его пределами.

Студентам обеспечен удаленный доступ к современным профессиональным базам данных и информационным справочным системам.

Основная литература:

1. Карау, Х. Изучаем Spark. Молниеносный анализ данных : практическое руководство / Х. Карау, Э. Конвински, П. Венделл, З. Матей ; пер. с англ. - 2-е изд. - Москва : ДМК Пресс, 2023. - 305 с. - ISBN 978-5-89818-320-2. - Текст : электронный. - URL: <https://znanium.ru/catalog/product/2102607>.

2. Харенслак, Б. Apache Airflow и конвейеры обработки данных : практическое руководство / Б. Харенслак, Д. дэ Руйтер ; пер. с англ. Д. А. Беликова. - Москва : ДМК Пресс, 2022. - 502 с. - ISBN 978-5-97060-970-5. - Текст : электронный. - URL: <https://znanium.ru/catalog/product/2155905>.

Дополнительная литература:

1. Лесковец, Ю. Анализ больших наборов данных : практическое руководство / Д. Дж. Ульман, Ю. Лесковец, А. Раджараман ; пер. с англ. А. А. Слинкина. - 2-е изд. - Москва : ДМК Пресс, 2023. - 500 с. - ISBN 978-5-89818-304-2. - Текст : электронный. - URL: <https://znanium.ru/catalog/product/2102592>.

6. Материально-техническое обеспечение

Университет располагает материально-технической базой, соответствующей действующим противопожарным правилам и нормам и обеспечивающей проведение всех видов дисциплинарной и междисциплинарной подготовки, практической и научно-исследовательской работ обучающихся, предусмотренных учебным планом.

Помещения, которые представляют собой учебные аудитории для проведения занятий лекционного типа, занятий семинарского (практического) типа, групповых и индивидуальных консультаций, текущего контроля и промежуточной аттестации, а также помещения для самостоятельной работы и помещения для хранения и профилактического обслуживания учебного оборудования. Помещения укомплектованы специализированной мебелью и техническими средствами обучения, служащими для представления учебной информации большой аудитории.

Изучение дисциплины (модуля) обеспечивается в учебных аудиториях, оснащенных:

- столами и стульями;
- компьютерной техникой;
- механическими калькуляторами;
- специализированным оборудованием, включая демонстрационное оборудование.

Помещения для самостоятельной работы обучающихся, в том числе приспособленные для использования инвалидами и лицами с ограниченными возможностями здоровья, оснащены компьютерной техникой с возможностью подключения к сети «Интернет» и обеспечением доступа в электронную информационно-образовательную среду Университета.

Обучающимся предоставляется доступ (в том числе удаленный) к ресурсам информационно-телекоммуникационной сети «Интернет», электронным ресурсам (в том числе электронным библиотечным системам, современным профессиональным базам данных и информационным справочным системам):

№	Наименование портала (издания, курса, документа)	Ссылка
1.	Научная электронная библиотека elibrary.ru библиотека	https://elibrary.ru/defaultx.asp
2.	База данных для IT-специалистов	https://habr.com
3.	База данных ScienceDirect	https://www.sciencedirect.com
4.	Официальный сайт Министерства науки и высшего образования Российской Федерации	https://minobrnauki.gov.ru/
5.	Федеральный портал «Российское образование»	https://www.edu.ru/
6.	Информационная система "Единое окно доступа к образовательным ресурсам"	http://window.edu.ru/
7.	Единая коллекция цифровых образовательных ресурсов	http://school-collection.edu.ru/
8.	Федеральный центр информационно - образовательных ресурсов	http://fcior.edu.ru/

Перечень информационных технологий, используемых при осуществлении образовательного процесса по дисциплине (модулю), в том числе комплект лицензионного программного обеспечения, современные профессиональные базы данных и информационные справочные системы:

Наименование ПО	Производство	Лицензионное / свободно распространяемое
Операционные системы:		
Microsoft Imagine (Windows Client, Server)	зарубежное	лицензионное
Браузеры:		
Яндекс.Браузер	отечественное	свободно распространяемое
Google Chrome	зарубежное	свободно распространяемое
Офисные приложения:		
Microsoft Imagine (Visio, OneNote)	зарубежное	лицензионное
TeXstudio	зарубежное	свободно распространяемое
Adobe Acrobat Reader	зарубежное	свободно распространяемое
Программное обеспечение для планирования и учета времени:		
Toggle app	зарубежное	свободно распространяемое
Системы управления проектами:		
Microsoft Imagine (Project)	зарубежное	лицензионное
Системы управления базами данных:		
Microsoft Imagine (SQL Server)	зарубежное	лицензионное
Системы резервного копирования (backup):		
Acronis Backup Advanced for HyperV	зарубежное	лицензионное
Справочно-правовые системы:		
КонсультантПлюс: справочно-правовая система	отечественное	лицензионное
Средства антивирусной защиты:		
Kaspersky Endpoint Security для бизнеса Стандартный Russian Edition	отечественное	лицензионное
Среды разработки:		
Visual Studio Code	зарубежное	свободно распространяемое
Bash (Unix shell)	зарубежное	свободно распространяемое

Anaconda	зарубежное	свободно распространяемое
Robotic Operating System	зарубежное	свободно распространяемое
CopelliaSim	зарубежное	свободно распространяемое
Google Colaboratory	зарубежное	свободно распространяемое
Пакеты программных средств и библиотек:		
AutoPsy	зарубежное	свободно распространяемое
Interactive Disassembler (IDA)	зарубежное	свободно распространяемое
Системы управления библиографической информацией:		
Zotero	зарубежное	свободно распространяемое
Сервисы и службы:		
Bind	зарубежное	свободно распространяемое
Docker	зарубежное	свободно распространяемое

7. Методические и оценочные материалы

Методические указания для обучающихся по освоению дисциплины (модуля)

В процессе изучения дисциплины (модуля) «Обучение с подкреплением» в рамках текущего контроля успеваемости используются такие виды учебной работы, как лекции, семинары, домашние задания, тесты, контрольная работа, а также различные виды самостоятельной работы обучающихся по заданию преподавателя, направленные на развитие навыков профессиональной лексики, закрепление практических профессиональных компетенций, поощрение инициатив.

Лекция – систематическое, последовательное, монологическое изложение преподавателем учебного материала, как правило, теоретического характера.

В процессе лекций рекомендуется вести конспект лекций: кратко и схематично фиксировать основные идеи, выводы и обобщения лекции; выделять важные мысли, ключевые слова и термины. Необходимо отметить вопросы или материалы, которые вызывают затруднения, и попытаться найти ответы в рекомендованной литературе. Если разобраться в материале не удастся, следует сформулировать вопрос и задать его преподавателю на консультации или во время семинарского (практического) занятия.

Семинар – это форма учебной деятельности, проводимая в учебном заведении под руководством преподавателя, где студенты активно участвуют в обсуждениях, практических заданиях и других формах взаимодействия.

Для успешной подготовки к семинару рекомендуется заранее ознакомиться с темой занятия и основными материалами, чтобы иметь возможность активно участвовать в обсуждении. Также полезно подготовить вопросы и идеи для обсуждения, что поможет глубже понять материал и продемонстрировать заинтересованность.

Аудиторная работа – активная работа студента на семинаре, его ответы на вопросы преподавателя и участие в дискуссии.

Для успешного участия в семинаре студентам рекомендуется заранее ознакомиться с темой обсуждения, прочитать необходимые материалы и подготовить вопросы. Важно активно слушать и вовлекаться в дискуссию, высказывая свои мнения и аргументируя их. При ответах на вопросы преподавателя стоит быть уверенным, четким и логичным, опираясь на изученный материал. Также полезно поддерживать диалог с однокурсниками, чтобы обогатить обсуждение и расширить свои знания.

Домашнее задание – набор задач по темам недели.

При работе над домашними заданиями важно внимательно ознакомиться с требованиями и сроками выполнения. Рекомендуется разбивать задания на этапы, чтобы избежать перегрузки и лучше усвоить материал. Использовать различные источники

информации, включая учебники и онлайн-ресурсы, для более глубокого понимания темы.

Кейс – практическая работа студентов над реальными или смоделированными задачами, что позволяет студенту применять теоретические знания на практике.

Студент самостоятельно разрабатывает стратегию решения поставленной задачи, что способствует развитию навыков критического мышления и самостоятельного принятия решений. Такой подход помогает подготовить будущих специалистов к реальным вызовам в их профессиональной деятельности.

Контрольная работа – письменная работа с набором задач, которые нужно решить за ограниченное время.

Цель контрольной работы – получить специальные знания по одной или нескольким темам дисциплины и продемонстрировать навыки их практического применения.

Бонусные баллы — это оценки, которые студенты могут получить за выполнение дополнительных заданий.

Формат бонусных баллов позволяет студентам улучшить общую оценку по дисциплине (модулю) и стимулирует углубленное изучение материала.

Самостоятельная работа – работа студентов, направленная на углубленное изучение отдельных тем и вопросов учебной дисциплины (модуля).

В процессе самостоятельной работы студенты взаимодействуют с рекомендованными материалами при минимальном участии преподавателя. Задачи студента включают работу с конспектами лекций (обработка текста), повторное изучение учебных материалов планов и тезисов ответов, изучение дополнительных тем, выполнение учебно-исследовательских заданий и другое.

Система оценивания результатов обучения по дисциплине (модулю)

Критерии получения уровня и оценивания сформированности компетенций по дисциплине (модулю) «Обучение с подкреплением»

Оценивание уровня учебных достижений, обучающихся по дисциплине (модулю), осуществляется в виде текущего контроля успеваемости и промежуточной аттестации.

Промежуточная аттестация по дисциплине (модулю) осуществляется в форме *экзамена*, при этом проводится оценка компетенций, сформированных по дисциплине.

Для оценивания текущего контроля успеваемости и промежуточной аттестации используется десятибалльная шкала оценивания, которая соотносится с традиционной пятибалльной шкалой следующим образом, если иное не указано в силлабусе дисциплины:

Десятибалльная оценка	Пятибалльная оценка	Общая характеристика результата обучения по дисциплине (модулю)
10	Отлично	Студент полностью владеет знаниями, изложенными в рабочей программе, и глубоко осмысляет дисциплину. Он самостоятельно и логически последовательно отвечает на все вопросы, акцентируя внимание на наиболее важном. Умеет анализировать, сравнивать, классифицировать, обобщать, конкретизировать и систематизировать изученный материал, выделяя ключевые моменты и устанавливая причинно-следственные связи. Четко формулирует ответы, уверенно интерпретирует результаты анализов и других исследований, а также решает сложные
9	Отлично	
8	Отлично	

Десятибалльная оценка	Пятибалльная оценка	Общая характеристика результата обучения по дисциплине (модулю)
		задачи. Студент хорошо знаком с методами исследования, необходимыми для практической деятельности, и умеет связывать теоретические аспекты дисциплины (модуля) с практическими задачами.
7	Хорошо	Студент обладает знаниями предмета почти в полном объеме рабочей программы и самостоятельно, логически последовательно и всесторонне отвечает на все вопросы, акцентируя внимание на наиболее значимых моментах. Он умеет анализировать, сравнивать, классифицировать, обобщать, конкретизировать и систематизировать изученный материал, выделяя его ключевые аспекты и устанавливая причинно-следственные связи. Формулирует свои ответы, уверенно интерпретирует результаты анализов и других исследований, а также решает сложные ситуационные задачи. Студент хорошо знаком с методами исследования, необходимыми для практической деятельности, и умеет связывать теоретические аспекты предмета с практическими задачами.
6	Хорошо	
5	Удовлетворительно	Студент обладает базовыми знаниями по дисциплине (модулю), но испытывает трудности при самостоятельных ответах и использует неточные формулировки. В ходе ответов он допускает ошибки, касающиеся сути вопросов. Студент способен решать только самые простые задачи и владеет лишь минимальным набором методов исследования.
4	Удовлетворительно	
3	Не сдан	Студент не овладел обязательным минимумом знаний по предмету и не может ответить на вопросы, даже если преподаватель задает дополнительные наводящие вопросы.
2	Не сдан	
1	Не сдан	

Дисциплина (модуль) «Обучение с подкреплением» оценивается следующим образом:

Активность	Вес	Количество	Описание
Домашние задания	50%	14	Набор задач по темам недели
Кейс	20%	1	Практическая работа студентов над реальными или смоделированными задачами
Контрольная работа	15%	2	Письменная работа с набором задач, которые нужно решить за ограниченное время
Экзамен	15%	1	Письменная или устная работа над заданием, направленным на проверку полученных знаний и навыков по дисциплине (модулю)

В рамках изучения дисциплины (модуля) возможно получение бонусных баллов.

Формула расчёта итоговой оценки по дисциплине (модулю) «Обучение с подкреплением»: $0,5 \times \text{среднее за домашние задания} + 0,2 \times \text{кейс} + 0,15 \times \text{контрольная работа} + 0,15 \times \text{экзамен}$.

Текущий контроль успеваемости обучающихся по дисциплине (модулю)

Примерные домашние задания

Домашнее задание 1.

1. Объясните, в чём заключается суть обучения с подкреплением и как взаимодействуют агент и среда в процессе обучения;
2. Опишите основные компоненты задачи обучения с подкреплением: состояние, действие, награда, стратегия и функция ценности;
3. Сформулируйте задачу обучения с подкреплением в виде марковского процесса принятия решений и объясните, зачем требуется марковское свойство;
4. Объясните, чем отличаются эпизодическое и непрерывное обучение, и приведите примеры задач для каждого типа;
5. Приведите примеры реальных приложений обучения с подкреплением в робототехнике, играх и управлении ресурсами.

Домашнее задание 2.

1. Объясните, как методы динамического программирования используют уравнения Беллмана для поиска оптимальной стратегии;
2. Опишите, как работает итерация по стратегиям и итерация по ценности в табличных методах обучения с подкреплением;
3. Реализуйте простой алгоритм Q-обучения для среды с конечным числом состояний и действий (например, сетка 4x4);
4. Объясните, в чём разница между методами Монте-Карло и временным разностным обучением (TD-learning) при оценке функции ценности;
5. Сравните алгоритмы SARSA и Q-обучение, указав различия в подходах текущей политики (on-policy) и другой политики (off-policy).

Домашнее задание 3.

1. Объясните, почему для сред с высокой размерностью состояний недостаточно табличных методов и требуется аппроксимация функции ценности;
2. Опишите архитектуру алгоритма DQN и роль буфера воспроизведения в стабилизации обучения;
3. Объясните, как целевая сеть (target network) помогает уменьшить корреляцию между обучающими выборками и стабилизирует сходимость;
4. Реализуйте или опишите схему модификации DQN с двойным обучением (Double DQN) и объясните её преимущества;
5. Опишите принцип работы Dueling DQN и как разделение оценки состояния и преимущества действия улучшает обучение.

Примерное описание задания и критерии оценивания к кейсу

Цель задания: применить теоретические знания и практические навыки по обучению с подкреплением для анализа, проектирования и обоснования решения реальной или смоделированной задачи в одной из предметных областей: робототехника, игры, управление ресурсами, рекомендательные системы или выравнивание языковых моделей.

Задание для студентов:

- выбрать или получить кейс, описывающий задачу принятия решений в динамической среде;
- формализовать задачу в терминах обучения с подкреплением: определить состояние, действие, награду, стратегию и тип среды (эпизодическая / непрерывная, дискретная / непрерывная);

- обосновать выбор алгоритма (например, q-обучение, dqn, ppo, sac) с учётом особенностей задачи и доступных данных;
- описать архитектуру решения: структуру нейросети (если используется), способ аппроксимации функции ценности или стратегии, метод стабилизации обучения;
- провести анализ возможных проблем: нестабильность, сходимость, исследование/эксплуатация, смещение политики;
- подготовить презентацию и защитить решение перед комиссией или группой, ответив на вопросы по обоснованности выбора и альтернативным подходам.

Примеры кейсов:

- обучение агента проходить лабиринт с изменяющейся структурой;
- оптимизация управления температурой в умном доме;
- дообучение языковой модели по предпочтениям пользователя (RLHF);
- управление автономным транспортным средством на перекрёстке;
- балансировка нагрузки в облачной инфраструктуре.

Критерии оценивания кейса:

- корректность формализации задачи в терминах обучения с подкреплением: чёткое определение состояния, действия, награды и среды;
- обоснованность выбора алгоритма с учётом типа задачи (дискретное/непрерывное управление, размерность пространства, доступ к среде);
- глубина анализа архитектуры решения: использование современных подходов (буфер воспроизведения, целевые сети, gae, энтропия и др.);
- способность выявить и прокомментировать потенциальные проблемы обучения: сходимость, нестабильность, компромисс между исследованием и эксплуатацией;
- качество презентации и защита: логичность изложения, ответы на вопросы, сравнение с альтернативными методами.

Примерные задания для контрольной работы

1. Объясните, в чём заключается отличие обучения с подкреплением от обучения с учителем и без учителя, и приведите пример задачи, где требуется именно обучение с подкреплением;
2. Опишите, как формализуется задача обучения с подкреплением в виде марковского процесса принятия решений: укажите состояния, действия, награду и стратегию;
3. Объясните, как уравнение Беллмана используется для оценки функции ценности и почему оно лежит в основе большинства алгоритмов;
4. Сравните методы Монте-Карло и временного разностного обучения (TD) по способу обновления Q-функции и сходимости;
5. Опишите, как работает алгоритм Q-обучения и в чём его основное отличие от SARSA;
6. Объясните, зачем в DQN используется буфер воспроизведения и целевая сеть, и как эти механизмы стабилизируют обучение;
7. Опишите, как архитектура актор-критик сочетает оценку ценности и непосредственное моделирование стратегии;
8. Объясните, как функция преимущества (advantage function) и метод GAE уменьшают дисперсию градиента в алгоритмах на основе политики;
9. Опишите, как алгоритм PPO использует обрезку вероятностей для ограничения шага обновления стратегии;
10. Объясните, почему алгоритмы, такие как DQN, не подходят для задач с непрерывным пространством действий, и как Soft Actor-Critic решает эту проблему.

Задания для промежуточной аттестации по дисциплине (модулю)

№ п/п	Задание	Ответ	Компетенция	Индикатор
1.	Назовите основную цель агента в процессе обучения с подкреплением.	Максимизация суммарной дисконтированной награды	ППК-Р5	ППК-Р5.1. 3-3
2.	Укажите, как называется среда, в которой будущее состояние зависит только от текущего состояния и действия.	Марковская среда / MDP	ППК-Р5	ППК-Р5.4. 3-1
3.	Объясните, в чём заключается принцип «исследование против эксплуатации» в обучении с подкреплением.	Баланс между изучением новых действий и использованием известных оптимальных	ППК-Р5	ППК-Р5.1. У-2
4.	Назовите тип задачи, в которой агент взаимодействует со средой без заранее заданной модели переходов и наград.	Model-free learning / Обучение без модели	ППК-Р5	ППК-Р5.3. 3-2
5.	Как называется функция, оценивающая полезность состояния при следовании определённой стратегии?	Value function / Функция ценности	ППК-Р5	ППК-Р5.4. 3-1
6.	Назовите функцию, которая оценивает, насколько выгодно выполнение конкретного действия в заданном состоянии.	Q-функция / Функция действия-состояния	ППК-Р5	ППК-Р5.3. У-3
7.	Укажите, как называется процесс улучшения стратегии на основе оценки функции ценности.	Policy iteration / Итерация по стратегии	ППК-Р5	ППК-Р5.4. У-1
8.	Назовите метод, использующий уравнение Беллмана для пошагового улучшения оценки ценности состояний.	Value iteration / Итерация по ценности	ППК-Р5	ППК-Р5.3. 3-2
9.	Назовите алгоритм табличного обучения, который обновляет Q-оценки на основе разности временных разностей.	Q-learning	ППК-Р5	ППК-Р5.3. У-3
10.	Укажите, как называется алгоритм, обновляющий Q-функцию с учётом текущей стратегии выбора действия.	SARSA	ППК-Р5	ППК-Р5.4. 3-1
11.	Объясните, в чём разница между on-policy и off-policy методами.	On-policy оценивает текущую стратегию, off-policy — может использовать другую	ППК-Р5	ППК-Р5.3. 3-2
12.	Назовите подход, при котором агент обучается на основе траекторий, собранных с помощью другой стратегии.	Off-policy обучение	ППК-Р5	ППК-Р5.4. У-1
13.	Назовите алгоритм глубокого обучения с подкреплением, сочетающий нейросеть и Q-обучение.	DQN / Deep Q-Network	ППК-Р5	ППК-Р5.3. 3-2

14.	Как называется механизм в DQN, позволяющий хранить и повторно использовать переходы среды?	Replay buffer / Буфер воспроизведения	ППК-P4	ППК-P4.2. 3-1
15.	Назовите компонент DQN, используемый для стабилизации целевых значений.	Target network / Целевая сеть	ППК-P5	ППК-P5.3. У-3
16.	Назовите модификацию DQN, которая устраняет переоценку действий за счёт разделения выбора и оценки.	Double DQN	ППК-P5	ППК-P5.3. 3-2
17.	Назовите метод, в котором стратегия обновляется напрямую с использованием градиента ожидаемой награды.	Policy Gradient / Метод градиента политики	ППК-P5	ППК-P5.4. У-1
18.	Укажите, как называется архитектура, сочетающая оценку ценности (критик) и моделирование стратегии (актор).	Actor-Critic / Актор- критик	ППК-P5	ППК-P5.3. 3-2
19.	Назовите алгоритм, использующий обрезку вероятностей для стабильного обновления стратегии.	PPO / Proximal Policy Optimization	ППК-P5	ППК-P5.3. У-3
20.	Назовите алгоритм непрерывного управления, максимизирующий награду и энтропию стратегии.	SAC / Soft Actor- Critic	ППК-P5	ППК-P5.4. 3-1
21.	Какой тип базы данных подходит для хранения опыта агента (transitions)?	NoSQL / Time-series DB / Redis	ППК-P4	ППК-P4.2. У-1
22.	Какой метод используется для оптимизации запросов к буферу воспроизведения?	Индексирование / Batch sampling	ППК-P4	ППК-P4.3. У-2
23.	Какой принцип Agile применяется в разработке RL-проектов?	Scrum / Kanban / Iterative	РГ-4	РГ-4.1. 3-1
24.	Что означает аббревиатура CI/CD в контексте MLOps для RL?	Continuous Integration / Continuous Delivery	МО-2	МО-2.1. 3-1
25.	Какой инструмент используется для версионирования моделей и экспериментов в RL?	MLflow / DVC / Weights & Biases	МО-2	МО-2.2. У-2
26.	Какая метрика важна для мониторинга качества RL-модели в продакшене?	Average Reward / Success Rate / Convergence	МО-2	МО-2.3. 3-1
27.	Какой процесс позволяет автоматизировать переобучение RL-агента?	Auto-retraining Pipeline / Adaptive RL	МО-2	МО-2.4. У-1
28.	Какой навык необходим для распределения задач в команде RL-разработчиков?	Лидерство / Делегирование	РГ-1	РГ-1.1. 3-2
29.	Как называется документ, описывающий архитектуру RL-системы?	Техническая спецификация / Архитектурное описание	ППК-P5	ППК-P5.2. 3-2

30.	Какой принцип чистого кода важен при разработке RL-пайплайнов?	DRY / SOLID / KISS	ППК-Р6	ППК-Р6.3. 3-3
31.	Что такое технический долг в контексте разработки RL-систем?	Накопленные проблемы кода / Долг рефакторинга	РГ-4	РГ-4.3. У-3
32.	Какой инструмент ИИ помогает генерировать код для RL-алгоритмов?	GitHub Copilot / ChatGPT / Cursor	ППК-Р7	ППК-Р7.1. У-1
33.	Что необходимо учитывать при использовании ИИ-кода с точки зрения этики?	Лицензирование / Безопасность / Качество кода	ППК-Р7	ППК-Р7.4. 3-1
34.	Какой метод используется для проверки гипотез в эксперименте с RL-агентом?	A/B тестирование / T-test / ANOVA	ППК-ДА5	ППК-ДА5.2. У-2
35.	Как называется процесс сбора информации о бизнес-проблеме для RL-проекта?	Интервью / Анализ требований	ППК-ДА4	ППК-ДА4.1. У-3
36.	Какой навык важен для презентации результатов RL-модели заказчику?	Коммуникация / Визуализация	УНК-01 (УК-4)	УНК-01.01.01. У-4
37.	Какое действие проявляет поддержку коллег в команде data science?	Наставничество / Code Review / Обратная связь	УНК-02 (УК-3)	УНК-02.01.01. У-3
38.	Какой инструмент используется для управления задачами в RL-проекте?	Jira / Trello / Asana	РГ-4	РГ-4.1. У-3
39.	Что такое стратегия развертывания промышленного RL-решения?	Canary / Blue-Green / Rolling	ППК-Р6	ППК-Р6.4. 3-1
40.	Какой метод используется для оптимизации инференса RL-агента?	Model Pruning / Quantization / ONNX	ППК-Р6	ППК-Р6.3. У-3
41.	Как называется план процесса разработки RL-продукта?	Roadmap / План проекта	РГ-2	РГ-2.2. У-3
42.	Какой принцип важен при проектировании архитектуры мультиагентной RL-системы?	Modularity / Scalability / Fault Tolerance	ППК-Р5	ППК-Р5.3. 3-2
43.	Как называется процесс адаптации к изменениям в требованиях к RL-агенту?	Fine-tuning / Transfer Learning	УНК-01 (УК-4)	УНК-01.01.01. У-4
44.	Какой инструмент используется для поиска уязвимостей в коде RL-систем?	SAST / Security Scan	ППК-Р7	ППК-Р7.2. У-2
45.	Как называется метод интерпретации рекомендаций ИИ по оптимизации кода?	Code Review / Validation	ППК-Р7	ППК-Р7.3. У-2
46.	Что такое долгосрочное хранение данных эпизодов в RL-системах?	Experience Replay / Data Lake	ППК-Р4	ППК-Р4.2. У-1
47.	Какой навык необходим для работы в команде над RL-проектом?	Коллаборация / Коммуникация	УНК-02 (УК-3)	УНК-02.01.01. У-2

48.	Как называется процесс оценки результатов проверки работоспособности RL-продукта?	Testing / Validation / QA	РГ-1	РГ-1.2. У-1
49.	Какой метод используется для выявления бизнес-проблем в RL-проекте?	Анализ предметной области / Business Analysis	ППК-ДА4	ППК-ДА4.2. У-1
50.	Как называется документация, оформленная в соответствии с профессиональными стандартами?	Техническая документация / API Docs	УНК-01 (УК-4)	УНК-01.02.01. У-4