Федеральное государственное автономное образовательное учреждение
высшего образования
«Национальный исследовательский университет
«Высшая школа экономики»

На правах рукописи

Иванов Дмитрий Игоревич

Приложение машинного обучения к теоретико-игровым задачам: аукционы и марковские игры

РЕЗЮМЕ ДИССЕРТАЦИИ
на соискание ученой степени
кандидата компьютерных наук

Москва – 2024

Диссертация была подготовлена в «Национальном исследовательском университете «Высшая школа экономики»

Научный руководитель: **Александр Нестеров**, Ph.D.; доцент департамента экономики, заведующий международной лабораторией Теории Игр и Принятия Решения, «Национальный исследовательский университет «Высшая школа экономики»

Теория игр обеспечивает математическую основу для моделирования стратегических взаимодействий как в конкурентных, так и в кооперативных сценариях. В основе теории игр (и экономики в целом) лежит предпосылка рациональности агентов. Идеального человеческого агента, который оптимально обрабатывает информацию, не несет вычислительных затрат, избегает ошибок и когнитивных искажений и действует рационально, часто называют *homo economicus*. Parkes and Wellman (2015) заметили, что агенты искусственного интеллекта (ИИ) могут лучше соответствовать этим идеалам, и ввели термин *machina economicus* как синтетический аналог идеально рациональному человеческому агенту.

Конечно, оба этих вида не существуют в реальности. Люди отклоняются от предпосылки рациональности бесчисленным множеством способов, а современные генеративные модели ИИ порой не могут решить тривиальные задачи, галлюцинируют (Zhang et al., 2023; Huang et al., 2023) и даже проявляют те же когнитивные искажения, что и люди (Schramowski et al., 2022; Acerbi and Stubbersfield, 2023). Тем не менее, теория игр показала себя незаменимым источником моделей человеческого поведения, и ее применение к агентам ИИ является естественным следующим шагом. Из этого следует огромный потенциал для синергии областей теории игр и ИИ.

По мере того, как ИИ все больше интегрируется во все аспекты общества, крайне важной является разработка методов для анализа и управления взаимодействиями агентов ИИ, особенно в условиях конфликта интересов. Теория игр и экономика предлагают широкий набор инструментов, которые можно адаптировать для этой цели (Conitzer, 2019; Hadfield-Menell and Hadfield, 2019), как уже было продемонстрировано в таких разнообразных областях, как классификация (Ghalme et al., 2021), рекомендательные системы (Bahar et al., 2020), мультиагентное обучение с подкреплением (Leibo et al., 2017) и даже большие языковые модели (Duetting et al., 2024).

В то же время, машинное обучение открывает новые пути для решения ранее недостижимых теоретико-игровых задач. В качестве примера можно выделить зарождающуюся область дифференцируемой экономики (Dütting et al., 2024), которая применяет методы глубокого обучения к таким областям, как дизайн аукционов (Dütting et al., 2019; Curry et al., 2023) и контрактов (Wang et al., 2024).

В этой диссертации представлены примеры обоих направлений, что демон-

стрирует взаимное обогащение машинного обучения и теории игр.

Актуальность и Значимость

Мое первое исследование продвигает область автоматизированного дизайна аукционов (automated auction design), максимизирующих доход, с помощью глубокого обучения. Классический подход, широко используемый в литературе, заключается в получении аналитических решений путем применения теоретического анализа к подмножествам всех возможных аукционов или даже к их конкретным примерам (Myerson, 1981; Manelli and Vincent, 2006; Pavlov, 2011; Giannakopoulos and Koutsoupias, 2014; Daskalakis et al., 2015; Yao, 2017; Haghpanah and Hartline, 2021). Этот подход предполагает фиксирование определенных параметров аукциона, таких как количество продаваемых предметов (items), количество участников (participants) и/или распределение ценностей (values) каждого участника для каждого набора предметов. Помимо точного анализа каждой конкретной ситуации, а также нереалистичных требований доступа к частной информации, этот подход неприменим даже в простых задачах, где участвуют лишь два участника и продаются лишь два предмета.

В качестве альтернативы, автоматизированный дизайн аукционов (Conitzer and Sandholm, 2002, 2003, 2004) использует методы оптимизации и машинного обучения для поиска приближенных решений в любых аукционах. Прорывом в этой области стал подход RegretNet (Dütting et al., 2019), который параметризует механизм аукциона нейронной сетью. В частности, RegretNet принимает ставки всех агентов на все предметы в качестве входных данных, обрабатывает их через полносвязную нейронную сеть, и аппроксимирует оптимальное вероятностное распределение предметов между участниками, а также суммарный платеж каждого участника за полученные предметы. Эта сеть обучается с использованием функции потерь, состоящей из двух частей: доходность (revenue – максимизация общей суммы платежей) и правдивость участников торгов (минимизация сожаления (regret), что является количественной мерой стимулов участников делать ставки стратегически в сравнении с честными ставками, равными их ценностям).

Я предлагаю два независимых улучшения RegretNet: во-первых, альтернативную архитектуру нейронной сети RegretFormer, основанную на слоях внимания;

во-вторых, новую функцию потерь, которая упрощает настройку гиперпараметров и предоставляет простой и интерпретируемый способ балансировки двух частей функции потерь. Оба улучшения провалидированы в обширном эмпирическом исследовании, которое выходит за рамки стандартного сравнения метрик эффективности и включает, например, дистилляцию нейронных сетей. В целом, эта работа представляет собой state-of-the-art подход к автоматическому дизайну аукционов.

В **моем втором исследовании** я критически анализирую распространенное предположение в мультиагентном обучении с подкреплением (Multi-Agent Reinforcement Learning, MARL), которое приравнивает кооперацию эгоистичных агентов к максимизации социального благосостояния. Большинство литературы рассматривает проблему кооперации агентов как исключительно вычислительную задачу, допуская неограниченное вмешательство в функции награды агентов (Peysakhovich and Lerer, 2018a,b; Hughes et al., 2018; Jaques et al., 2019; Wang et al., 2019; Eccles et al., 2019; Jiang and Lu, 2019; Durugkar et al., 2020; Yang et al., 2020; Zimmer et al., 2021; Phan et al., 2022) или их параметры (Gupta et al., 2017). Учитывая сложность типичных для MARL мультиагентных сред с конфликтующими предпочтениями агентов (формализованных как марковские игры (Markov games), Leibo et al. (2017)), этот традиционный подход упрощает как обучение, так и валидацию. Однако, такой подход игнорирует индивидуальность агентов во время обучения, а также их уязвимость к эксплуатации со стороны эгоистичных агентов. Альтернативный, экономический взгляд на проблему предполагает, что кооперация должна возникать в результате принятия стратегических решений рациональными эгоистичными агентами как равновесие (equilibrium), максимизирующее общественное благосостояние (social welfare) и устойчивое к отклонениям ради личной выгоды.

В качестве реализации описанной выше экономической концепции кооперации, я предлагаю использовать медиаторов (mediators), предложенных Monderer and Tennenholtz (2009). Медиатор определяется как третья сторона, которая может действовать в игре от лица агентов, которые на это соглашаются. Важно отметить, что если агент не считает передачу права на выбор своих действий медиатору приемлемым, он может выбрать действовать в игре самостоятельно. Однако в этом случае медиатор не будет учитывать предпочтения такого агента,

действуя от лица других агентов (которые согласились на передачу права принятия решений медиатору). Задача медиатора – сбалансировать стимулы всех агентов и достигнуть взаимовыгодного равновесия, реализуемого в случае, когда все агенты передают право выбора действий медиатору (mediated equilibrium). Чтобы адаптировать эту идею к MARL, я параметризую медиатора и агентов как нейронные сети, формулирую их взаимодействие как задачу оптимизации социального благосостояния с ограничениями на стимулы каждого агента и решаю ее, используя градиент политик (policy gradient).

Я демонстрирую эффективность этого подхода в достижении кооперативного равновесия без ущерба для индивидуальных наград агентов в классических социальных дилеммах, а также в их секвенциальных модификациях с огромным пространством возможных состояний среды. Предложенная методология открывает новые возможности для создания более устойчивых и справедливых взаимодействий агентов в сложных средах со смешанными мотивами.

Наконец, **мое третье исследование** вносит вклад в область персонализированного машинного обучения (personalized machine learning), которая касается адаптации моделей к уникальным характеристикам и предпочтениям конкретных пользователей (den Hengst et al., 2020). В частности, я фокусирюсь на возможностях персонализации в таких важных областях с высокой ценой ошибки, как здравоохранение и автономное вождение. В этих областях использование любого автоматизированного решения требует строгого и долгого процесса одобрения регулирующими органами (Breton et al., 2020), что делает персонализацию для каждого пользователя невозможной. Чтобы решить эту проблему, я предлагаю формальную модель, названную репрезентативный марковский процесс принятия решений (represented Markov Decision Process, r-MDP), которая призвана обеспечить баланс между необходимостью персонализации и нормативными ограничениями. r-MDP формализуется одноагентным марковским процессом принятия решений (single-agent MDP), множеством пользователей с различными предпочтениями и жестким ограничением на общее количество персонализированных политик. Каждый пользователь может выбрать одну из политик для его репрезентации в рамках MDP. Общая задача оптимизации включает в себя два взаимосвязанных аспекта: обучить репрезентативные политики (вычислительный аспект) и сопоставить каждого пользователя с политикой таким

образом, чтобы максимизировать общее социальное благосостояние (теоретико-игровой аспект). После того как политики обучены в симуляторе, они могут быть предоставлены на утверждение регулирующим органам и, наконец, внедрены в реальный мир.

При большом количестве агентов или политик решение r-MDP затруднено из-за экспоненциального роста сложности задачи сопоставления агентов политикам. В качестве альтернативы поиску глобального оптимума, я использую идеи из классических алгоритмов кластеризации K-means и Expectation Maximization (MacQueen, 1967; Dempster et al., 1977; Lloyd, 1982). В частности, я предлагаю два алгоритма глубокого обучения с подкреплением, которые итеративно обучают репрезентативные политики и переназначают агентов между политиками жадным образом. Эти алгоритмы мотивированны теорическими результатами: на каждой итерации они монотонно улучшают приближенное решение и, в конечном итоге, сходятся к локальным максимумам социального благосостояния.

Для проведения эмпирического исследования я использую простую, но наглядную среду сбора ресурсов (Barrett and Narayanan, 2008), а также сложные задачи управления роботами в физическом симуляторе MuJoCo (Todorov et al., 2012). Эксперименты демонстрируют универсальность и эффективность алгоритмов в реализации персонализированных политик в условиях жестких ограничений на их количество. Эти результаты не только подтверждают практичность моего подхода к достижению значимой персонализации в регулируемых областях, но и открывают дорогу для будущих исследований по применению этих методологий в реальных приложениях, сокращая разрыв между теоретическими моделями машинного обучения и практической необходимостью соблюдения нормативных требований.

Цели Исследования

1. Продвинуть область автоматизированного дизайна аукционов с помощью глубокого обучения посредством использования слоев внимания.

2. Продемонстрировать теоретико-игровой подход к кооперации в марковских средах с кофликтом интересов с помощью медиаторов.

3. Предложить компромиссный подход к персонализации моделей обучения с

подкреплением, подходящий для областей с высокой стоимостью внедрения различных политик.

Ключевые Результаты

На основе описанных выше исследований я выношу следующие **ключевые результаты на защиту**:

1. Предложенная архитектура RegretFormer, основанная на слоях внимания, представляет собой state-of-the-art в автоматизированном дизайне аукционов. Кроме того, предложенная модификация функции потерь, основанная на дуальном градиентном спуске, является менее чувствительной к гиперпараметрам и позволяет интерпретируемо контролировать соотношение максимизации доходности и минимизации сожалений участников.

2. Медиаторы могут быть применены в MARL со смешанными мотивами для создания нового равновесия, максимизирующего общественное благосостояние. Это равновесие можно найти с помощью предложенного алгоритма, применив градиент политик к сформулированной задаче оптимизации с ограничениями.

3. Значимая персонализация моделей машинного обучения для группы пользователей может быть достигнута с помощью лишь нескольких решений. В контексте RL, политики, представляющие собой эти решения, могут быть обучены с помощью предложенных алгоритмов, которые сочетают в себе высокоуровневую структуру K-means и EM кластеризаций с градиентом политик.

Личный вклад

Эти результаты были достигнуты в сотрудничестве как с экспертами в области машинного обучения, так и с талантливыми студентами. Однако во всех исследованиях я внес основной вклад, о чем свидетельствует мое первое авторство во всех трех публикациях, составляющих эту диссертацию.

Первое исследование было проведено с сокурсниками. Я руководил проектом и активно участвовал в формулировке научных гипотез, а также в реализации

алгоритмических разработок и экспериментов. Основные результаты исследования — state-of-the-art нейронная архитектура и улучшенная функция потерь — основаны на моих идеях. Я активно участвовал в написании статьи.

Над вторым исследованием я работал с двумя студентами. Я возглавлял этот проект, сформулировав направление исследования по применению медиаторов в MARL, задачу ограниченной оптимизации, ее решение с использованием градиента политик, и сопутствующие эксперименты. Студенты разработали кодовую базу, реализовали алгоритм, и провели большинство экспериментов под моим руководством. Статья полностью написана мной.

Третье исследование было проведено в сотрудничестве с академическим экспертом в области машинного обучения и теории игр, который сформулировал практическую проблему персонализации в областях с высокой ценой ошибки и предложил RL решение на основе кластеризации. Я продолжил исследование с этой точки – предложил модифицированную версию алгоритма (обе версии вошли в публикацию), разработал эксперименты и реализовал кодовую базу. Статья в основном написана мной, за исключением части введения.

Публикации и Апробация Исследований

Я опубликовал семь статей в материалах международных рецензируемых конференций. Три из этих публикаций составляют текущую диссертацию.

Публикации первого уровня

1. **Ivanov, D.**, Safiulin, I., Filippov, I., & Balabaeva, K. (2022). Optimal-er auctions through attention. In *Advances in Neural Information Processing Systems (NeurIPS)*, Vol. 35, pp. 34734-34747.

2. **Ivanov, D.**, Zisman, I., & Chernyshev, K. (2023). Mediated Multi-Agent Reinforcement Learning. In *Proceedings of the 2023 International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, Vol. 22, pp. 49-57.

3. **Ivanov, D.**, & Ben-Porat, O. (2024). Personalized Reinforcement Learning with a Budget of Policies. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38, pp. 12735-12743.

Доклады на конференциях и семинарах

1. Виртуальная постерная презентация на конференции *the 36th Conference on Neural Information Processing Systems (NeurIPS)*, Декабрь 2022, Новый Орлеан, США. Optimal-er auctions through attention.

2. Постерная презентация на конференции *the 22nd International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, Июнь 2023, Лондон, Великобритания. Mediated Multi-Agent Reinforcement Learning.

3. Устная презентация на *внутреннем исследовательском семинаре компании DeepMind*, Июнь 2023, Лондон, Великобритания. Mediated Multi-Agent Reinforcement Learning.

4. Предзаписанная устная презентация на конференции *the 38th AAAI Conference on Artificial Intelligence*, Февраль 2024, Ванкувер, Канада. Personalized Reinforcement Learning with a Budget of Policies.

Содержание

Optimal-er Auctions through Attention

Дизайн аукционов

В этой части моей диссертации я исследую механизмы аукционов, в которых множество участников $N = \{1, \dots i, \dots, n\}$ выражает свой интерес к набору предметов $M = \{1, \dots j, \dots, m\}$, через функции ценности $v_i(j)$ (value function). Эти функции показывают, как каждый участник торгов оценивает каждый предмет, и ключевым предположением является аддитивность оценок: общая ценность, которую участник торгов присваивает подмножеству предметов, представляет собой сумму ценностей, которые он присваивает каждому отдельному предмету.

Суть данного исследования заключается в понимании того, как участники торгов, каждый из которых имеет свою функцию ценностей, полученную из конкретных распределений, взаимодействуют в рамках аукциона. Задача аукционера – на основе данных предыдущих аукционов, аппроксимировать оптимальный аукцион в условиях отсутствия точных знаний об истинных ценностях участников торгов или их распределений.

Формально, аукцион определяется двумя функциями: функция распределения предметов между участниками и функция платежей каждого участника. Помимо максимизации прибыли, дополнительное условие от механизма – создать стимулы для участников торгов ставить свои истинные ценности. Эта концепция известна как совместимость стимулов в доминирующих стратегиях (Dominant Strategy Incentive-Combatibility, DSIC). Кроме того, аукцион должен быть индивидуально рациональным (Individually Rational, IR), гарантируя, что игрокам выгодно участвовать в аукционе независимо от исхода.

Задача разработки оптимальных аукционов рассматривается как задача оптимизации. Цель состоит в том, чтобы максимизировать ожидаемую доходность (сумму платежей) при соблюдении ограничений DSIC и IR. Хотя проблема решена аналитически для аукционов с одним предметом, аукционы с несколькими предметами не имеют аналитических решений в общем виде.

RegretNet

Опираясь на инновационный метод RegretNet (Dütting et al., 2019), моя диссертация исследует автоматический дизайн аукционов посредством глубокого обучения. Основой RegretNet является его двухсетевая архитектура, параметризующая функции распределения и платежей полносвязными нейронными сетями. Эти сети обрабатывают в качестве входных данных матрицу ставок (ставки всех участников для всех предметов, равные их ценностям) через несколько полносвязных слоев. В частности, сеть распределения определяет вероятности распределения предметов между участниками. Сеть платежей вычисляет платеж, который должен произвести каждый участник, как часть ожидаемой полезности участника, которая будет передана аукционисту.

Основным новшеством RegretNet является функция потерь. Архитектура спроектирована так, чтобы максимизировать доход с учетом ограничения DSIC. Это ограничение реализуется посредством подсчета сожаления участников (regret), равного дополнительной ожидаемой полезности, которую участник мог бы получить, оптимально отклонившись от своей истинной ценности при совершении ставки. RegretNet стремится свести это сожаление к нулю с целью снижения стимулов участников стратегически искажать свои ценности. Процесс оптимизации реализован методом Лагранжа, который уравновешивает две

противоречивые цели максимизации доходности и минимизации сожалений.

Я использую RegretNet в качестве основы и предлагаю модификации, позволяющие расширить применение глубокого обучение к дизайну аукционов.

Мои модификации RegretNet

Я предлагаю два существенных усовершенствования подхода RegretNet для оптимального дизайна аукционов: архитектуру RegretFormer, основанную на слоях внимания, и альтернативную функцию потерь.

С одной стороны, архитектура RegretNet сталкивается с проблемами, связанными с чувствительностью результатов аукциона к порядку предметов и участников в матрице ставок, требованием неизменного числа участников и предметов, а также ограниченной экспрессивностью полносвязных слоев. Эти проблемы препятствуют его практическому применению и возможности обобщения.

Для решения этих проблем, я предлагаю RegretFormer — новую архитектуру, основанную на слоях внимания. Архитектура проиллюстрирована на рисунке 1. В частности, слои внимания применяются как по рядам, так и по колонкам матрицы признаков, созданной на основе матрицы ставок. Выходные данные этих слоев внимания объединяются через полносвязный слой, и этот процесс можно повторять несколько раз. Последний шаг включает обработку этих выходных данных для создания матрицы распределения и вектора платежей. Такая конструкция гарантирует, что архитектура не зависит от порядка ставок, и позволяет применять подход к данным аукционов с варьируемым количеством предметов и участников. Более того, экспрессивность слоев внимания улучшает качество найденных механизмов на больших задачах.

С другой стороны, первоначальная процедура обучения RegretNet во многом опирается на точную настройку гиперпараметров для балансировки двух целей. Этот процесс не только скрупулезен, но и чреват возможным снижением качества решения, если гиперпараметры подобраны не оптимально (Rahme et al., 2021b).

Чтобы преодолеть эти проблемы, я предлагаю упрощенную и более интуитивную функцию потерь, которая приоритизирует максимизацию доходности при условии заранее определенного "бюджета"сожалений (regret budget). Этот подход к обучению формализован как релаксированная задача ограниченной

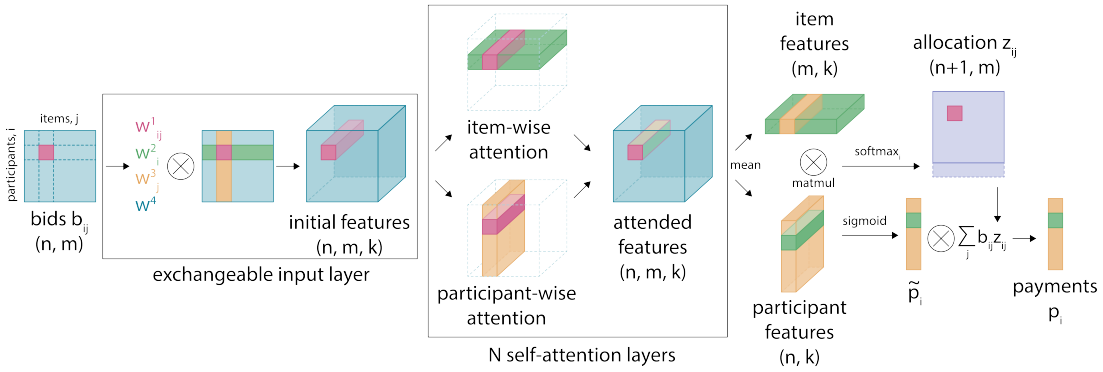


Рис. 1: RegretFormer

оптимизации, направленная на минимизацию среднего значения сожаления для всех участников без превышения указанного максимального порога (бюджета). В этой релаксированной задаче множитель Лагранжа настраивается автоматически посредством дуального градиентного спуска.

Данный подход имеет два существенных преимущества. Во-первых, он устраняет необходимость балансировать противоречивые цели с помощью нескольких гиперпараметров, упрощая их настройку. Во-вторых, явное указание бюджета сожалений делает процесс обучения более простым и менее чувствительным к изменениям гиперпараметров. Эмпирически этот метод устойчив к различным параметрам аукциона и не требует корректировки гиперпараметров от аукциона к аукциону.

С помощью этих модификаций я устраняю некоторые ограничения исходного метода RegretNet, предлагая более четкий и эффективный путь к оптимальному дизайну аукционов.

Эксперименты

Я провожу серию экспериментов с целью оценить эффективность RegretFormer по сравнению с RegretNet и EquivariantNet (Rahme et al., 2021a) при различных параметрах аукциона, а также эффективность модификации функции потерь для контроля соотношения доходности и сожалений. Подробное сравнение представлено в основной части диссертации.

В таблице 1 представлены эксперименты в конфигурациях (settings), отличающихся только количеством участников (n) и предметов (m), обозначенных как $n \times m$. Ценности всех участников для всех предметов независимо сэмплированы

Таблица 1: Сравнение архитектур

R_{max}	setting	RegretNet		EquivariantNet		RegretFormer	
		revenue	regret	revenue	regret	revenue	regret
10^{-3}	1x2	0.572	0.0007	**0.586**	0.00065	0.571	0.00075
	2x2	0.889	0.00055	0.878	0.0008	**0.908**	0.00054
	2x3	1.317	0.00102	1.365	0.00084	**1.416**	0.00089
	2x5	2.339	0.00142	2.437	0.00146	**2.453**	0.00102
	3x10	5.59	0.00204	5.744	0.00167	**6.121**	0.00179
10^{-4}	1x2	0.551	0.00007	0.548	0.00013	**0.556**	0.00014
	2x2	0.825	0.00005	0.75	0.00005	**0.861**	0.00006
	2x3	1.249	0.00007	1.226	0.0001	**1.327**	0.00011
	2x5	2.121	0.00013	2.168	0.00017	**2.339**	0.00015
	3x10	5.02	0.00062	5.12	0.00025	**5.745**	0.00022

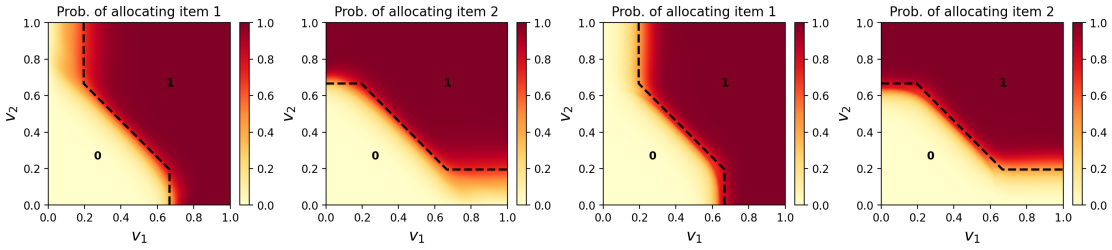
из равномерного распределения $U[0, 1]$. Эксперименты охватывают пять различных конфигураций: 1×2, 2×2, 2×3, 2×5 и 3×10, где конфигурация 1×2 является широко известным аукционом Манелли-Винсента с оптимально доходностью, равной 0.55. Для остальных конфигураций оптимальные доходности неизвестны.

Эмпирические результаты показывают, что RegretFormer достигает более высокую доходность во всех конифгурациях, за исключением самой простой конфигурации 1×2, где разница между моделями незначительна. Разрыв между моделями значительно увеличивается в более сложных конфигурациях. Хотя свойство RegretFormer игнорировать порядок ставок играет роль в этом результате, основным фактором, вероятно, является повышенная экспрессивность слоев внимания (в сравнении с полносвязными слоями). Кроме того, показано, что как RegretNet, так и RegretFormer точно аппроксимируют оптимальные вероятности распределения в конфигурации 1×2 (рис. 2).

Таблица 2 содержит сравнения оцененных сожалений участников с целевыми значениями бюджетов сожалений в результате обучения, что позволяет оценить модифицированную функцию потерь. Результаты показывают, что соотношение оцененных сожалений к бюджету сожалений колеблется около идеального

Таблица 2: Отношение оценки сожаления к бюджету; должно быть близко к 1

R_{max}	setting	RegretNet		EquivariantNet		RegretFormer	
		train	valid	train	valid	train	valid
10^{-3}	1x2	1.12	1.22	1.04	1.11	1.01	1.31
	2x2	0.97	1.24	1.41	1.82	0.89	1.19
	2x3	1.07	1.55	1.11	1.23	1.02	1.26
	2x5	0.94	1.21	1.11	1.2	0.8	0.83
	3x10	0.89	1.09	0.9	0.87	1.03	0.88
10^{-4}	1x2	0.94	1.27	0.92	2.37	1.31	2.52
	2x2	0.95	1.94	1.73	1.33	0.93	1.39
	2x3	1.52	1.12	1.57	1.63	1.6	1.66
	2x5	1.04	1.23	1.02	1.57	0.95	1.28
	3x10	0.9	3.71	1.05	1.46	0.88	1.15



(a) (b) (c) (d)

Рис. 2: Вероятности распределения в 1x2: (a, b) RegretNet; (c, d) RegretFormer

значения (единицы) во время обучения. Тем не менее, этап валидации, который включает в себя более точную оценку сожалений, выявляет некоторые отклонения, что позволяет предположить, что увеличение количества шагов оптимизации во время обучения может далее уточнить найденное решение.

Mediated Multi-Agent Reinforcement Learning

Марковские Игры и Секвенциальные Социальные Дилеммы

Следующая часть моей диссертации посвящена марковским играм как ключевой структуре для понимания взаимодействий в мультиагентном обучении с подкреплением (Multi-Agent Reinforcement Learning, MARL). Марковские игры рас-

ширяют марковский процесс принятия решений (Markov Decision Process, MDP) на множество агентов, каждый из которых имеет собственную функцию вознаграждения. Эти игры моделируют сценарии, в которых агенты, основываясь на текущем состоянии MDP, принимают одновременные решения в соответствии со своими политиками. На основе коллективного действия всех агентов, MDP переходит в новое состояние, отражающее взаимосвязанное влияние решений каждого агента.

Важнейшим аспектом динамики обучения в марковских играх является их схождение к некоему равновесию (обычно, совершенному равновесию на подыграх – Subgame Perfect Equilibrium или Markov Perfect Equilibrium Maskin and Tirole (2001)). В равновесии ни один агент не может изменить свою стратегию ради собственной выгоды, что обеспечивает стабильность общей политики.

Кроме того, мое исследование сосредоточено на определенном подмножестве марковских игр, известном как секвенциальные социальные дилеммы (Sequential Social Dillema, SSD). Для SSD характерен конфликт интересов между агентами, порождюащий общественно неоптимальные равновесия. Дилемма заключается в противоречии между индивидуальными стимулами и общественным благосостоянием, в результате чего максимизация собственной функции наград каждым агентом приводит к равновесию, субоптимальному как для всех агентов, так и для каждого конкретного агента.

Медиаторы

Медиатор является дополнительным игроком, который может действовать в среде от имени подмножества агентов, называемого *коалицией* (coalition). Медиация конкретного агента происходит только в случае его добровольного *согласия* (commitment), что сохраняет автономию агентов — агенты могут выбирать, взаимодействовать с медиатором (передав ему право выбора действий) или действовать независимо.

Стратегия медиатора определена для каждой возможной коалиции. В отличие от традиционного экономического подхода с фиксированной стратегией медиатора Monderer and Tennenholtz (2009), моя адаптация предполагает его обучение с помощью RL параллельно с другими агентами в среде. Эту концепцию я формализую как *марковские медиаторы*. Марковский медиатор имеет ту

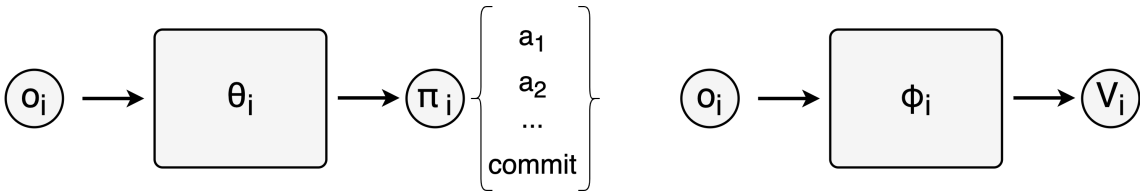


Рис. 3: Схематическое изображение архитектур actor (слева) и critic (справа) агентов

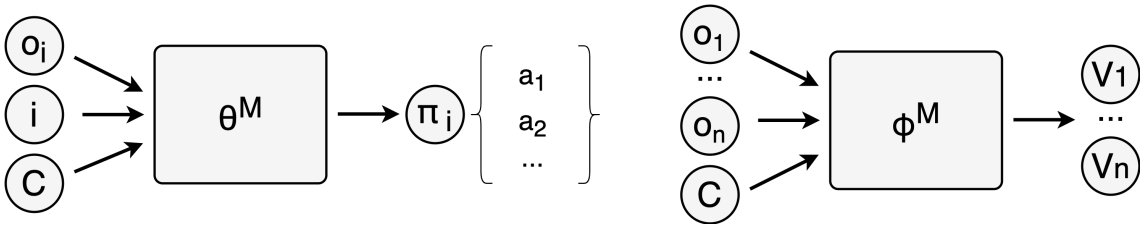


Рис. 4: Схематическое изображение архитектур actor (слева) и critic (справа) медиатора

же информацию о состоянии среды, что и агенты в коалиции, и принимает за них решения с целью максимизации их суммарной награды. Эта формулировка позволяет медиатору реализовать равновесие, в котором медиатор действует за всех агентов и максимизирует общее благосостояние. Агенты периодически решают, вступать ли в коалицию или действовать самостоятельно, сохраняя тем самым автономию агентов.

Политика медиатора, которая стимулирует всех агентов к согласию на медиацию, называется *медиационным равновесием* (mediated equilibrium). Это равновесие служит альтернативной равновестю Нэша и совершенному равновесию на подыграх.

Deep Mediated MARL

Как агенты, так и медиатор обучаются методом Actor-Critic. Actor представляет собой политику, в то время как критик представляет собой аппроксимацию функции ценности (value function). Обе функции можно параметризовать с помощью нейронных сетей. Их архитектуры показаны на рисунках 3 и 4. Обозначения следующие: o_i обозначает наблюдение агента i (которое является функцией состояния MDP), π_i обозначает политику агента i, C обозначает коалицию, V_i обозначает аппроксимацию функции ценности агента i, а θ и ϕ обозначают параметры нейронных сетей.

Наивный подход к обучению медиатора заключается в максимизации общественного благосостояния коалиции, игнорируя стимулы отдельных агентов. Этот подход, хоть и максимизирует благосостояние, не мотивирует агентов к согласию на медиацию, поскольку может не учитывать их собственные награды. На основе этого наблюдения, я ввожу два вида ограничений, чтобы согласовать цели медиатора с целями отдельных агентов. Во-первых, ограничение совместимости стимулов (Incentive-Compatibility, IC) гарантирует, что агенты не пожалеют о присоединении к коалиции, получая по крайней мере такую же награду, сколько они получили бы действуя независимо от медиатора. Во-вторых, ограничение поощрения (Encouragement, E) не позволяет агентам вне коалиции эксплуатировать медиатора, гарантируя, что их награда не превысит того, что они получили бы, если бы согласились на медиацию. Чтобы учесть эти ограничения при обучении медиатора, я применяю метод множителей Лагранжа к градиенту политик.

Эксперименты

В одном из ключевых экспериментов я исследую феномен «безбилетника» (free riding) в играх — ситуацию, когда некоторые агенты полагаются на совместные усилия других для личной выгоды. Эта проблема становится особенно острой в средах с более чем двумя агентами, где введение медиатора может непреднамеренно привести к появлению безбилетников, тем самым подрывая общественное благосостояние. Чтобы проиллюстрировать это, я использую Public Good Game (PGG), сравнивая наивного медиатора и медиатора с ограничениями.

В PGG каждый из N агентов обладает (is endowed) единицей полезности, которую они могут либо внести в общественное благо (contribute), либо удержать (defect). Общий вклад приумножается на коэффициент n, превышающий единицу, но меньший, чем N, а затем равномерно перераспределяется между всеми агентами, создавая стимулы отказаться от участия в общественном благе.

Результаты эксперимента, представленные в таблице 3), наглядно демонстрируют ожидаемую динамику. Рассмотрим игру с $N = 3$ агентами и множителем $n = 2$. Без участия медиатора, агенты по умолчанию удерживают вклад. Наивному медиатору удается достигнуть устойчивой кооперации между двумя агентами, но выученная детерменированная стратегия стимулирует третьего

Таблица 3: Результаты в одношаговой Public Good Game (PGG); c и m обозначают действия *contribute* и *commit*, $|C|$ обозначают размер коалиции, $\tilde{\pi}$ и $\tilde{\pi^M}$ обозначают усредненную политику агентов и медиатора, соответственно. $^{(*)}$ без медиатора $^{(\dagger)}$ наивный медиатор $^{(\ddagger)}$ медиатор с ограничениями

PGG	$N = 3$	$N = 10$	$N = 25$
reward$^{(*)}$	0.012	0.0	0.0
reward$^{(\dagger)}$	0.652	0.005	0.014
$\tilde{\pi}(m)^{(\dagger)}$	0.658	0.159	0.121
$\tilde{\pi}^M(c)^{(\dagger)}$	0.985	0.001	0.02
$\pi^M(c \mid \|C\| = 2)^{(\dagger)}$	0.993	-	-
$\pi^M(c \mid \|C\| = 3)^{(\dagger)}$	0.999	-	-
reward$^{(\ddagger)}$	0.891	0.827	0.817
$\tilde{\pi}(m)^{(\ddagger)}$	0.916	0.961	0.933
$\tilde{\pi}^M(c)^{(\ddagger)}$	0.959	0.858	0.817
$\pi^M(c \mid \|C\| = 2)^{(\ddagger)}$	0.774	-	-
$\pi^M(c \mid \|C\| = 3)^{(\ddagger)}$	0.996	-	-

агента удерживать свой вклад, получая долю общественного блага. Наконец, медиатор с ограничениями успешно ведет всех агентов к кооперации и общественно оптимальному равновесию, смешивая стратегии агентов в неполных коалициях. Таким образом, этот медиатор эффективно решает проблему безбилетника и находит политику, учитывающую стимулы агентов вне коалиции с целью предотвратить эксплуатацию коалиции.

Результаты устойчивы к увеличению числа агентов: медиатор с ограничениями стабильно способствует кооперации между всеми агентами с помощью политики, которая адаптируется к возможным эксплуататорам. Этот эксперимент подчеркивает способность медиатора преодолевать сложности мультиагентного обучения с подкреплением, способствуя равновесию, которое балансирует индивидуальные стимулы с коллективным благосостоянием.

В основном тексте также представлены эксперименты с секвенциальной модификацией PGG, где баланс каждого агента сохраняется между несколькими раундами PGG, и общественное благо может расти экспоненциально при кооперации. Эти эксперименты позволяют оценить скалируемость предложенного

подхода.

Personalized RL with a Budget of Policies

Represented Markov Decision Processes

В моем исследовании я предлагаю Represented Markov Decision Processes (r-MDPs) как модификацию стандартных MDPs для задачи персонализации решений машинного обучения в применениях с высокой ценой ошибки. r-MDP формализуется как кортеж $\mathcal{M}_r = (S, A, \mathcal{T}, \mathcal{T}_0, \gamma, N, K, (r^i)_{i \in N})$, включающий элементы стандартного одноагентного MDP, такие как множество состояний S, множество действий A, динамика перехода $\mathcal{T}$, распределение начальных состояний $\mathcal{T}_0$ и коэффициент дисконтирования γ. Кроме того, он вводит N как множество n агентов, K как множество репрезентативных политик (representative policies), количество которых ограничено бюджетом k, и r^i как индивидуальную функцию вознаграждения для каждого агента i.

В рамках r-MDP агенты не взаимодействуют напрямую со средой. Вместо этого каждый агент представлен одной из политик из множества K, которая действует от его имени. Оптимизационная задача в r-MDP состоит из двух компонент: оптимальное назначие (assignment) агентов репрезентативным политикам согласно вероятностям (α^i) и обучение этих политик (π^j) максимизации общественного благосостояния.

Новизна r-MDP заключается в абстракции взаимодействия между агентами и средой, различая между действующими в среде политиками и получающими награду агентами. Репрезентативные политики максимизируют благосостояние агентов, которых они представляют, не имея при этом собственных функций вознаграждения. Эта абстракция позволяет целенаправленно подходить к максимизации общественного благосостояния в условиях ограничения на количество различных политик.

Важно отметить, что среда остается одноагентной, и каждая политика фактически действует в своей собственной копии MDP с идентичной динамикой, но разными функциями вознаграждения (равной сумме функций вознаграждения назначенных ей агентов).

Мой подход к решению r-MDPs

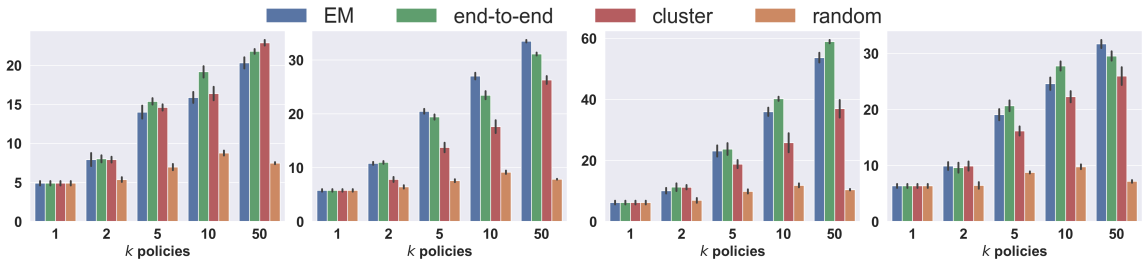
Моя методология для решения r-MDPs обходит проклятие размерности, возникающее из-за экспоненциального роста числа возможных паросочетаний (matchings) агентов и политик по мере увеличения числа агентов. Прямая оптимизация этих паросочетаний требует поиска оптимального решения в множестве с кардинальностью, пропорциональной K^n, что становится невозможным при больших n из-за огромного числа возможных комбинаций.

В качестве решения, мой подход предполагает разбиение проблемы на более простые компоненты. С одной стороны, предположим, что политики π^j фиксированы. При этом условии задача максимизации общественного благосостояния упрощается до жадного назначения каждого агента i представителю j^*, политика которого обеспечивает этому агенту наивысшую ожидаемую суммарную награду.

С другой стороны, предположим, что вероятности распределния каждого агента по политикам α^i фиксированы. В этих условиях, максимизация общественного благосостояния равносильна решению набора MDP – по одному на каждую политику. Этот подход превращает совместную задачу оптимизации политик в набор независимых задач, каждая из которых сосредоточена на оптимизации одной политики.

Объединение этих двух упрощений приводит к факторизованному подходу, при котором оптимизация паросочетаний и политик осуществляется независимо, но итеративно. На каждой итерации сперва оптимизируется назначение агентов на основе текущего набора политик, а затем политики улучшаются на основе этого паросочетания. Эта методология направлена на итеративное приближение к локально оптимальному решению совместной максимизации общественного благосостояния, тем самым преодолевая проклятие размерности в r-MDP с помощью разбиения оптимизацинной задачи на более простые подзадачи.

В качестве конкретной реализации описанного выше факторизованного подхода я предлагаю алгоритм, основанный на классических подходах к кластеризации, K-means и Expectation-Maximization (EM). Этот EM-подобный алгоритм специализирован под решение r-MDPs. Алгоритм работает посредством итеративного процесса оптимизации, состоящего из двух основных фаз: Expectation (E-фаза) и Maximization (M-фаза).



(a) Ant (b) HalfCheetah (c) Hopper (d) Walker2d

Рис. 5: Общественное благосостояние, достигаемое с помощью наших (EM, end-to-end) и бейзлайн алгоритмов в средах MuJoCo

На E-фазе, аналогично распределению точек по кластерам в алгоритмах кластеризации, агенты распределяются по представителям на основе текущих политик. Этот процесс формализуется на основе таблицы размера $n \times k$, обозначенной $\tilde{Q}$, в которой хранятся приблизительные значения Q-ценностей (Q-values) для каждой пары агент-политика. На основе этих значений, каждый агент жадно перераспределяется наилучшей политике, соответствующей максимальной Q-ценности в ряде агента i таблицы $\tilde{Q}$, тем самым максимизируя их суммарную ожидаемую полезность.

После перераспределения агентов M-фаза независимо улучшает каждую политику для повышения благосостояния назначенных этой политике агентов. Каждая политика обучается с помощью метода Proximal Policy Optimization (PPO, Schulman et al. (2017)) на сумме наград назначенных агентов, в результате чего обновленные политики совместно повышают суммарное благосостояние всех агентов.

Подобно итеративному процессу K-means, предложенный алгоритм сходится к локальному оптимуму общественного благосостояния для данного r-MDP. Сходимость алгоритма является не просто эмпирическим наблюдением, но также формально установлена в виде теоремы.

Кроме того, дополнительно предложена модификация EM-подобного алгоритма, релаксирующая жадное перераспределение агентов по политикам на E-шаге.

Эксперименты

Чтобы проверить эффективность предложенных алгоритмов, я провел эксперименты в симуляторе MuJoCo, в частности, в средах HalfCheetah, Ant, Hopper и Walker2d. Эти среды предполагают управление роботами с помощью непрерывных действий (сил, приложенных к суставам робота) на основе многомерных состояний среды.

Чтобы формализовать эти среды как r-MDP, моделируется $n = 100$ агентов, каждому из которых случайно назначена уникальная целевая скорость робота. Награда агента затем определяется на основе того, насколько точно скорость робота соответствует назначенной этому агенту целевой скорости на каждом временном шаге среды, тем самым персонализируя одну и ту же среду под каждого агента через функцию награды.

В экспериментах изучались различные бюджетные ограничения на количество политик k – 1, 2, 5, 10 и 50 – чтобы проследить влияние этих ограничений на способность алгоритмов к эффективной персонализации.

Результаты, представленные на рисунке 5, показывают эффективность предложенных алгоритмов (как EM-подобного алгоритма, описанного выше, так и его вариации end-to-end) по сравнению с существующим подходом из литературы по персонализации. Примечательно, что оба алгоритма не только значительно превзошли случайное распределение агентов по политикам, но и стабильно превосходят clustering baseline (Hassouni et al., 2018) практически во всех протестированных средах и конфигурациях, за исключением среды Ant.

Эти результаты подчеркивают стабильность и адаптируемость предложенных алгоритмов, демонстрируя их потенциал для достижения значимой персонализации в приложениях машинного обучения даже в жестких условиях, связанных со сложными многомерными задачами и строгими ограничениями на число политик.

Заключение

Данная диссертация объединяет области теории игр и искусственного интеллекта, демонстрируя с помощью трех отдельных исследований, как глубокое обучение и обучение с подкреплением могут быть использованы для решения

сложных задач на пересечении этих областей. Каждое из этих исследований способствует нашему пониманию и возможностям разработки систем искусственного интеллекта в мультиагентных системах с учетом их теоретико-игровых особенностей.

В первом исследовании представлена RegretFormer, новая архитектура на основе глубокого обучения для автоматизированного дизайна оптимальных аукционов, превосходящая существующие методы. Переосмысливая RegretNet, эта работа не только продвигает state-of-the-art подход, но также упрощает процесс оптимизации, уменьшая чувствительность к настройкам гиперпараметров, и предлагает нетривиальные методы валидации, которые могут принести пользу будущим исследованиям.

Второе исследование предлагает альтернативный взгляд на кооперацию в мультиагентном обучении с подкреплением, применяя теоретико-игровую концепцию медиаторов с целью создания и достижения кооперативного равновесия. Адаптируя медиаторов к контексту марковских игр, это исследование выходит за рамки кооперации как чисто вычислительной задачи, представляя метод оптимизации с ограничениями, который приоритезирует как общественное, так и индивидуальное благосостояние. Применение медиаторов в MARL открывает множество возможностей для будущих исследований: от их применения в более сложных средах до их объединения с криптографическими технологиями для полной децентрализации.

В третьем исследовании акцент смещается на задачу персонализации ИИ решений в рамках регуляторных ограничений с помощью концепции represented Markov Decision Processes. Разработаны два алгоритма глубокого обучения с подкреплением, которые демонстрируют возможность достижения персонализации в условиях строгих ограничений на количество политик. Более того, теоретико-игровой взгляд на проблему как на оптимизацию социального благосостояния закладывает основу для последующих исследований. Например, они могут включать соображение о справедливости достигнутых агентами наград, оптимизируя не только суммарное благосостояние, но и равномерность его распределения между агентами.

В совокупности эти исследования подчеркивают синергетический потенциал сочетания теории игр с машинным обучением для создания мультиагентных систем, которые не только интеллектуальны и адаптивны, но также устойчивы

к манипуляциям и направлены на улучшение благосостояния каждого агента. Данная диссертация демонстрирует, как теоретико-игровые принципы могут направлять и расширять исследования в области ИИ, от развития кооперации в области MARL до персонализации решений в областях с высокой ценой ошибки, раскрывая потенциал для будущих исследований на пересечении этих двух ключевых областей. Более того, исследование по автоматизированному дизайну аукционов является примером применения ИИ для решения фундаментальной теоретико-игровой проблемы, демонстрируя потенциал глубокого обучения в нетривиальном практическом приложении.

Список литературы

Acerbi, A. and Stubbersfield, J. M. (2023). Large language models show human-like content biases in transmission chain experiments. *Proceedings of the National Academy of Sciences*, 120(44):e2313790120.

Bahar, G., Ben-Porat, O., Leyton-Brown, K., and Tennenholtz, M. (2020). Fiduciary bandits. In *International Conference on Machine Learning*, pages 518–527. PMLR.

Barrett, L. and Narayanan, S. (2008). Learning all optimal policies with multiple criteria. In *Proceedings of the 25th international conference on Machine learning*, pages 41–47.

Breton, M. D., Kanapka, L. G., Beck, R. W., Ekhlaspour, L., Forlenza, G. P., Cengiz, E., Schoelwer, M., Ruedy, K. J., Jost, E., Carria, L., et al. (2020). A randomized trial of closed-loop control in children with type 1 diabetes. *New England Journal of Medicine*, 383(9):836–845.

Conitzer, V. (2019). Designing preferences, beliefs, and identities for artificial intelligence. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 9755–9759.

Conitzer, V. and Sandholm, T. (2002). Complexity of mechanism design. In *Proceedings of the Eighteenth conference on Uncertainty in artificial intelligence*, pages 103–110.

Conitzer, V. and Sandholm, T. (2003). Automated mechanism design: Complexity

results stemming from the single-agent setting. In *Proceedings of the 5th international conference on Electronic commerce*, pages 17–24.

Conitzer, V. and Sandholm, T. (2004). Self-interested automated mechanism design and implications for optimal combinatorial auctions. In *Proceedings of the 5th ACM Conference on Electronic Commerce*, pages 132–141.

Curry, M., Sandholm, T., and Dickerson, J. (2023). Differentiable economics for randomized affine maximizer auctions. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, pages 2633–2641.

Daskalakis, C., Deckelbaum, A., and Tzamos, C. (2015). Strong duality for a multiple-good monopolist. In *Proceedings of the Sixteenth ACM Conference on Economics and Computation*, pages 449–450.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22.

den Hengst, F., Grua, E. M., el Hassouni, A., and Hoogendoorn, M. (2020). Reinforcement learning for personalization: A systematic literature review. *Data Science*, 3(2):107–147.

Duetting, P., Mirrokni, V., Paes Leme, R., Xu, H., and Zuo, S. (2024). Mechanism design for large language models. In *Proceedings of the ACM on Web Conference 2024*, pages 144–155.

Durugkar, I., Liebman, E., and Stone, P. (2020). Balancing individual preferences and shared objectives in multiagent reinforcement learning. *Good Systems-Published Research*.

Dütting, P., Feng, Z., Narasimhan, H., Parkes, D., and Ravindranath, S. S. (2019). Optimal auctions through deep learning. In *International Conference on Machine Learning*, pages 1706–1715. PMLR.

Dütting, P., Feng, Z., Narasimhan, H., Parkes, D. C., and Ravindranath, S. S. (2024). Optimal auctions through deep learning: Advances in differentiable economics. *Journal of the ACM*, 71(1):1–53.

Eccles, T., Hughes, E., Kramár, J., Wheelwright, S., and Leibo, J. Z. (2019). Learning reciprocity in complex sequential social dilemmas. *arXiv preprint arXiv:1903.08082*.

Ghalme, G., Nair, V., Eilat, I., Talgam-Cohen, I., and Rosenfeld, N. (2021). Strategic classification in the dark. In *International Conference on Machine Learning*, pages 3672–3681. PMLR.

Giannakopoulos, Y. and Koutsoupias, E. (2014). Duality and optimality of auctions for uniform distributions. In *Proceedings of the fifteenth ACM conference on Economics and computation*, pages 259–276.

Gupta, J. K., Egorov, M., and Kochenderfer, M. (2017). Cooperative multi-agent control using deep reinforcement learning. In *International conference on autonomous agents and multiagent systems*, pages 66–83. Springer.

Hadfield-Menell, D. and Hadfield, G. K. (2019). Incomplete contracting and ai alignment. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 417–422.

Haghpanah, N. and Hartline, J. (2021). When is pure bundling optimal? *The Review of Economic Studies*, 88(3):1127–1156.

Hassouni, A. e., Hoogendoorn, M., van Otterlo, M., and Barbaro, E. (2018). Personalization of health interventions using cluster-based reinforcement learning. In *PRIMA 2018: Principles and Practice of Multi-Agent Systems: 21st International Conference, Tokyo, Japan, October 29-November 2, 2018, Proceedings 21*, pages 467–475. Springer.

Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., Chen, Q., Peng, W., Feng, X., Qin, B., et al. (2023). A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *arXiv preprint arXiv:2311.05232*.

Hughes, E., Leibo, J. Z., Phillips, M., Tuyls, K., Dueñez-Guzman, E., García Castañeda, A., Dunning, I., Zhu, T., McKee, K., Koster, R., et al. (2018). Inequity aversion improves cooperation in intertemporal social dilemmas. *Advances in neural information processing systems*, 31.

Jaques, N., Lazaridou, A., Hughes, E., Gulcehre, C., Ortega, P., Strouse, D., Leibo, J. Z., and De Freitas, N. (2019). Social influence as intrinsic motivation for multi-agent deep reinforcement learning. In *International conference on machine learning*, pages 3040–3049. PMLR.

Jiang, J. and Lu, Z. (2019). Learning fairness in multi-agent systems. *Advances in Neural Information Processing Systems*, 32.

Leibo, J. Z., Zambaldi, V., Lanctot, M., Marecki, J., and Graepel, T. (2017). Multi-agent reinforcement learning in sequential social dilemmas. In *Proceedings of the 16th Conference on Autonomous Agents and MultiAgent Systems*, pages 464–473.

Lloyd, S. (1982). Least squares quantization in pcm. *IEEE transactions on information theory*, 28(2):129–137.

MacQueen, J. (1967). Classification and analysis of multivariate observations. In *5th Berkeley Symp. Math. Statist. Probability*, pages 281–297.

Manelli, A. M. and Vincent, D. R. (2006). Bundling as an optimal selling mechanism for a multiple-good monopolist. *Journal of Economic Theory*, 127(1):1–35.

Maskin, E. and Tirole, J. (2001). Markov perfect equilibrium: I. observable actions. *Journal of Economic Theory*, 100(2):191–219.

Monderer, D. and Tennenholtz, M. (2009). Strong mediated equilibrium. *Artificial Intelligence*, 173(1):180–195.

Myerson, R. B. (1981). Optimal auction design. *Mathematics of operations research*, 6(1):58–73.

Parkes, D. C. and Wellman, M. P. (2015). Economic reasoning and artificial intelligence. *Science*, 349(6245):267–272.

Pavlov, G. (2011). Optimal mechanism for selling two goods. *The BE Journal of Theoretical Economics*, 11(1):000010220219351704 1664.

Peysakhovich, A. and Lerer, A. (2018a). Consequentialist conditional cooperation in social dilemmas with imperfect information. In *International Conference on Learning Representations*.

Peysakhovich, A. and Lerer, A. (2018b). Prosocial learning agents solve generalized stag hunts better than selfish ones. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems*, pages 2043–2044. International Foundation for Autonomous Agents and Multiagent Systems.

Phan, T., Sommer, F., Altmann, P., Ritz, F., Belzner, L., and Linnhoff-Popien, C. (2022). Emergent cooperation from mutual acknowledgment exchange. In *Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems*, pages 1047–1055.

Rahme, J., Jelassi, S., Bruna, J., and Weinberg, S. M. (2021a). A permutation-equivariant neural network architecture for auction design. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(6):5664–5672.

Rahme, J., Jelassi, S., and Weinberg, S. M. (2021b). Auction learning as a two-player game. In *International Conference on Learning Representations*.

Schramowski, P., Turan, C., Andersen, N., Rothkopf, C. A., and Kersting, K. (2022). Large pre-trained language models contain human-like biases of what is right and wrong to do. *Nature Machine Intelligence*, 4(3):258–268.

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., and Klimov, O. (2017). Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

Todorov, E., Erez, T., and Tassa, Y. (2012). Mujoco: A physics engine for model-based control. In *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 5026–5033. IEEE.

Wang, J. X., Hughes, E., Fernando, C., Czarnecki, W. M., Duéñez-Guzmán, E. A., and Leibo, J. Z. (2019). Evolving intrinsic motivations for altruistic behavior. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems*, pages 683–692. International Foundation for Autonomous Agents and Multiagent Systems.

Wang, T., Duetting, P., Ivanov, D., Talgam-Cohen, I., and Parkes, D. C. (2024). Deep contract design via discontinuous networks. *Advances in Neural Information Processing Systems*, 36.

Yang, J., Li, A., Farajtabar, M., Sunehag, P., Hughes, E., and Zha, H. (2020). Learning to incentivize other learning agents. *Advances in Neural Information Processing Systems*, 33:15208–15219.

Yao, A. C.-C. (2017). Dominant-strategy versus bayesian multi-item auctions: Maximum revenue determination and comparison. In *Proceedings of the 2017 ACM Conference on Economics and Computation*, pages 3–20.

Zhang, Y., Li, Y., Cui, L., Cai, D., Liu, L., Fu, T., Huang, X., Zhao, E., Zhang, Y., Chen, Y., et al. (2023). Siren's song in the ai ocean: a survey on hallucination in large language models. *arXiv preprint arXiv:2309.01219*.

Zimmer, M., Glanois, C., Siddique, U., and Weng, P. (2021). Learning fair policies in decentralized cooperative multi-agent reinforcement learning. In *International Conference on Machine Learning*, pages 12967–12978. PMLR.