УДК 334.02, 51-77, 65.012
DOI 10.26118/2782-4586.2026.38.63.098

Иринина Алена Юрьевна
Уральское отделение российской академии наук Институт экономики (ИЭ УрО РАН)

Роль AI в симуляционной игре с прямой конфронтацией для принятия стратегических решений

Аннотация. В работе предложен метод применения ансамбля из трех специализированных AI-агентов для имитационного моделирования и анализа стратегических решений. Модель совершенствует метод «Красных и синих команд» в классической вариации «Pro&Contra». Где «красной» (атакующие действия) и «синей» (оборонительные действия) команде соответствует свой AI-агент, «фиолетового» агента непосредственно использует оператор для сценарного моделирования и генерации нестандартных условий. Функция человека-оператора формализована как роль верховного арбитра, осуществляющего постановку целевых функций и принятие финального решения на основе анализа результатов многовариантного моделирования. В рамках исследования проведен анализ существующих моделей искусственного интеллекта. Приведены принципы работы моделей по созданию глубоких подделок, синтетических изображение и генеративных текстов. Рассмотрены последствия ошибок ИИ-моделей. Приведены риски и прогнозирование решений, которые будут применены при реализации этих рисков. Доказана возможность использования AI-агентов в качестве инструментов помощи управленцам в моделировании сценариев развития. Приведена концепция метода прогнозирования и проведения симуляционной игры с прямой конфронтацией для принятия стратегических решений. Целью исследования является разработка эффективного метода принятия управленческих решений с использованием AI-агентов.

Ключевые слова: методы принятия стратегических решений, «Pro&Contra», AI в методах управления, прогнозирование в экономике, инновации.

Финансирование: Статья подготовлена в соответствии с планом НИР Института экономики Уральского отделения Российской академии наук.

Irinina Alena Yuryevna
Ural Branch of Russian Academy of Sciences

The role of artificial intelligence in simulation wargaming for strategic decisions

Annotation. This research proposes a method employing an ensemble of three specialized AI agents for simulation modeling and analysis of strategic decisions. The model enhances the classic "Pro & Contra" variation of the "Red and Blue Teams" methodology. In this framework, the "red" (offensive) and "blue" (defensive) teams are each represented by a dedicated AI agent, while a "purple" agent is utilized directly by the operator for scenario modeling and the generation of non-standard conditions. The role of the human operator is formalized as that of a supreme arbiter, tasked with defining objective functions and making the final decision based on an analysis of multi-scenario modeling outcomes. The research includes an analysis of existing artificial intelligence models. It outlines the operating principles of models used for creating deepfakes, synthetic images, and generative texts. The consequences of AI model errors are examined. Potential risks are presented along with forecasts of the decisions likely to be implemented in response to these risks. The work proves the feasibility of using AI agents as tools to assist managers in modeling development scenarios. It presents a conceptual method for forecasting and conducting head-to-head simulation games for strategic decision-making. The aim of the research is to develop an effective managerial decision-making method utilizing AI agents.

Keywords: strategic decision-making methods, «Pro&Contra», AI in management methods, forecasting in economics, innovations.

Funding: The article was prepared in accordance with the research plan of the Institute of Economics of the Ural Branch of the Russian Academy of Sciences.

Введение.

Управленцы оказались в трудном положении. Необходимо разрабатывать собственные уникальные, универсальные и реализуемые в промышленности проекты. Одна из главных задач на ближайшие годы - построить высокоскоростную магистраль (Путин: ВСМ укрепят статус России как важнейшего логистического звена Евразии // Информационное агентство ТАСС. URL: https://tass.ru/obschestvo/24695069 (дата обращения: 20.10.2025). Отправной точкой, предполагается, станет маршрут Москва-Санкт-Петербург. Задача возникла не столько из желания обрести технологическое лидерство, сколько из-за возникновения реальной угрозы [1]. Небо становится небезопасным из-за доступности БПЛА. Способы обезопасить магистраль можно было бы позаимствовать, но пути в России длиннее и извилистее, чем в Китае и Европе. К тому же часть дорожных сетей имеет электрификацию постоянным током, а часть - переменным, что затрудняет наблюдение сверху. Скорость движения состава (допустимая скорость скоростного поезда - 130 км/ч, высокоскоростного - 350 км/ч) предполагает защищенность путей от людей, животных и других объектов. Поэтому необходимо самим придумать, как определять опасности и своевременно реагировать на них. Вторая важнейшая задача - обеспечить защиту людей и стратегических объектов от БПЛА. Уже работают схемы «стая волков» и «рой пчел», развивается роевой интеллект (В РФ применили алгоритмы поведения животных в природе для задач энергетики // Информационное агентство ТАСС. - URL: https://tass.ru/nauka/22711851 (дата обращения: 12.10.2025). Скоро станут возможны автономные полеты связок беспилотников с децентрализованным управлением и принятием решений в воздухе. Внедрение природоподобных технологий дает потенциал для развития, но также сопряжено с множеством рисков и угроз. Значимые задачи поставлены в импортозамещении, в цифровизации, в развитии водородной энергетики и других глобальных целях. Российским управленцам предстоит придумать способы развития собственных технологий и борьбы с возможными противниками [2]. Для чего необходимо делать полноценную и точную оценку возможных рисков. Наиболее эффективными для составления списка сценариев являются симуляционные игры для принятия стратегических решений [3]. Подготовка к которым занимает по меньшей мере полгода, численности задействованных в подготовке людей составляет до 60% от количества участников, а стоимость проведения многократно превышает стоимость классический стратегический сессий. Исследование посвящено интеграции AI-агентов в игры для принятия стратегических решений с целью их усовершенствования и упрощения в использовании.

Основная часть

Материалы и методы исследования

Искусственный интеллект

Сегодня понятие «искусственный интеллект» не имеет единого и строгого определения. Вернее, определений существует множество, и их трактовки варьируются от сугубо технических до почти метафизических. Нередко считается, что «искусственный интеллект» есть своего рода цифровой двойник или дополнение человеческого разума, базирующееся на возможностях вычислительной техники. Что, конечно, не соответствует действительности. Причина тому непонимание принципов AI, частые обновления и тот факт, что для описания технологии было использовано слово, означающее присущую лишь человеку способность - интеллект. Оригинальное название «artificial intelligence» в упрощенном виде можно перевести как «машинный ответ» [4]. И машина в ответах

уподобляется естественному языку и логике - имитация. AI технология не нова, но развитие вычислительной техники сократило объём слотов индексации и сделало AI доступной технологией любому пользователю с доступом к сети Интернет.

Если спросить самые популярные модели ИИ, будь-то Deepseek (Deepseek - китайская LLM от DeepSeek AI.) или Grok (Grok - частная LLM от xAI.), на какие типы делится ИИ по функциональным возможностям, то смысл ответа следующий: существует только один реально работающий тип - узконаправленный ANI (Artificial Narrow Intelligence - слабый искусственный интеллект.), гипотетический - общий AGI (Artificial General Intelligence - общий искусственный интеллект.) и гипотетический - сверхинтеллект ASI (Artificial Superintelligence - сверхинтеллект.).

К первому типу относятся модели генерации изображений, из которых самыми популярными являются модели с принципом диффузии информации и с генеративно-состязательной сетью GAN (Generative Adversarial Network) в основе. Диффузионные модели основаны на принципах неравновесной термодинамики, где цепь Маркова сначала добавляет гауссов шум к данным, а затем обращает этот процесс для их восстановления. Модель обучается обратному преобразованию, формируя генеративные паттерны. GAN создана командой Яна Гудфеллоу в 2014 году и основана на противостоянии двух нейронных сетей: генератор преобразует случайный шум в синтетические изображения, а дискриминатор сравнивает их с реальной выборкой. При достижении равновесия одна генерирует реалистичные изображения, а вторая уже не отличает их от настоящих [5]. Этот метод известен в управлении как «Pro&Contra» и он настолько эффективен в GAN, что применение порождает угрозы.

Таким образом мы генерируем правдоподобные глубокие подделки. Для создания deepfake (Искусственно модифицированные кадры с изменениями лиц или тел субъектов.) -видео требуется по меньшей мере несколько месяцев и хорошие графические процессоры. У пользователей популярны услуги reface (Reface - ПО для замены лица на видео и изображениях.), zao (Zao - ПО для реалистичной замены лица с лицами киногероев.) и myheritage (Myheritage - онлайн-сервис "оживления" старых фото.), использование которых позволяет ИИ обучаться на реальных данных и совершенствоваться. Чаще deepfake применяют в развлекательных целях, однако технология популярна у мошенников. Проект по созданию качественного deepfake-видео будет стоить от 15000 долларов, что для развлечений дорого, а для эффективного мошенничества дешево. В 2025 году жертва из Великобритании отдала мошенникам 80 тыс. фунтов стерлингов. Злоумышленники использовали глубокую подделку аудиофайлов и видеофайлов на основе материалов известного актёра (Уотсон Э. «How I lost £80K to fake Jason Momoa Facebook scammers» : [Как я потерял 80 тыс. фунтов из-за мошенников с фейковыми страницами Джейсона Момоа в Facebook] // BBC News. - 2025. - 25 ноября. - URL: https://www.bbc.com/news/articles/cgmn7vdpy0wo.amp (дата обращения: 26.11.2025). Также за 2025 г. в США и Великобритании было зафиксировано более 80 случаев с использованием того же deepfake с общими потерями свыше 1 млрд. долларов США. Таким образом, пользователи невольно участвуют в обучении моделей, которые впоследствии могут быть применены для обмана.

В 2024 году сотрудник международной компании перевёл мошенникам 200 млн. гонконгских долларов после видеозвонка, на котором злоумышленники использовали глубокую подделку внешности и голоса финансового директора (Chen H., Magramo K. Finance worker pays out $25 million after video call with deepfake ‘chief financial officer’ [Сотрудник перевел $25 млн после видеозвонка с дипфейком «финансового директора»] [Электронный ресурс] // CNN. – 2024. – 4 февраля. – URL: https://edition.cnn.com/2024/02/04/asia/deepfake-cfo-scam-hong-kong-intl-hnk (дата обращения: 16.10.2025). За год было зафиксировано 105 тысяч инцидентов с применением глубоких подделок. Мошенничество с использованием глубоких подделок можно разделить два типа: презентационное - глубокие подделки используются в реальном

времени, например, для захвата учётных записей или проведения транзакций; инъекционное - глубокие подделки заранее сгенерированы и используются с занесением в ПО для атак на банки, финансовые и телекоммуникационные компании. В 2024 году 42% инцидентов в финансовом секторе прошли с использованием ИИ, однако инструменты обнаружения у многих компаний устарели и только 22% внедрили специализированные средства защиты (Fraud attempts with deepfakes have increased by 2137% over the last three years [Электронный ресурс] // Signicat. – 2025. – 20 февраля. – URL: https://www.signicat.com/press-releases/fraud-attempts-with-deepfakes-have-increased-by-2137-over-the-last-three-year (дата обращения: 12.11.2025). Если традиционное вымогательство основывалось на прямых угрозах с работой по каждой жертве, то цифровые методы позволяют массово атаковать с высокой степенью анонимности. А возможность генерировать глубокие подделки позволила злоумышленникам угрожать не реальным компроматом, а качественно сфабрикованным без необходимости сбора множества точной информации о жертве. Таким образом тысячи атак проходят на одной основе для компрометирующего видео. Deepfake-изображения часто используют для поддельных аккаунтов в социальных сетях вместо украденных фотографий для обхождения инструментов идентификации. Цели различные, например, для повышения активности на персональной странице, создания иллюзии общественного мнения или для последующего вымогательства. Из-за чего в социуме укрепляется страх перед новыми цифровыми технологиями, в том числе искусственным интеллектом [6].

Ещё одним типом глубокой подделки является клонирование голоса. Почти 100 лет назад Гомер Дадли представил водер - метод и механизм синтезирования голоса для шифрования, имитирующие работу речевого аппарата. Рычагом на запястье оператор выбирает тип генеративного шума, гудение для звонких звуков или шипение для глухих звуков, педалью регулирует высоту звука, 10 клавишами выбирает фильтр. Затем звуки объединяются и воспроизводятся громкоговорителем [7]. То есть звук разбит на акустические элементы: гласные, согласные, интонация, взрывные, назальные и тд. Принцип работы водера является частью синтезирования вокодера, устройства или ПО, который при помощи набора фильтров определяет потерны входящих сигналов, обычно двух - модулятора и носителя, и перекладывает свойства сигнала модулятора на сигнал носителя. До 2016 года искусственная озвучка характеризовалась роботизированным тембром, но подходы TTS (Text-to-Speech) на основе нейронных сетей сделали синтетическое звучание похожим на человеческое. До того как сигнал попадет на вокодер информацию подготавливают. Сначала исходный текст приводят в пригодное для работы состояние, то есть исправляют ошибки, раскрывают сокращения, расписывают числа, расставляют ударения, указывают паузы, расставляют интонации и тд. Затем энкодер преобразует обработанный текст в числовые векторы, эмбендинги, содержащие информацию об интонации, ритме и эмоциональном контексте. Энкодер, как правило, построен на базе одной из трех популярных сетей: рекуррентной нейронной сети RNN (Recurrent Neural Networks), когда сеть обрабатывает текст поэтапно по временной последовательности, сохраняя информацию о предыдущих шагах, что важно для работы с естественным языком; сверточной CNN (Convolutional neural network), которая путем фильтрации на каждом функциональном слое сети извлекает локальные признаки из входного текста; или трансформерной, которая преобразует каждый элемент текста в вектор, учитывая связи с другими элементами. Последняя предпочтительней из-за простоты обучения и высокой скорости, в сравнении с остальными. После специальный модуль определяет длительность каждой акустической единицы для создания естественной интенсивности речи. Далее декодер генерирует спектральное представление звука - мел-спектрограмму, которую вокодер преобразует в исходящий звук. В вокодере для голосовых генераций чаще используют модели на основе GAN, о принципах которых написано выше, и модели на основе потоков с вероятностным распределением FBM (Flow based model). FBM обучается сжимать сложный звук до простого шума с нормальным распределением, а

при синтезе выполняет обратное действие на основе построенных связей с помощью мел-спектрограммы. То есть, в отличии от диффузионных моделей, FBM выстраивают одношаговые мостики от x(y) к y(x) и наоборот, а не многошаговые.

Разработчики ИИ создают все более совершенные модели, для обучения которых необходимо множество разнообразных данных. В том числе, поэтому ИИ-модели такие комплиментарные. Иногда разработчики прибегают к дистопичным идеям. Например, существует услуга на основе TTS-модели, синтезирующая голос умершего родственника. В перспективе будет добавлена услуга видеозвонка, а дальше, как в известном британском сериале («Черное зеркало» 2 сезон 1 серия.). Модель способна лишь к поддельным ответам, основанным на вероятностном распределении. Но человек переживающий горе прибегает к подобным услугам, мешая естественному процессу принятия. Позже работа механизмов защиты психики заставит человека ненавидеть искусственную симуляцию. И остается этическая проблема коммерциализации подобных услуг. Очевидный вариант для государств - вводить запреты и ограничения в развитии и использовании моделей искусственного интеллекта.

К ANI также относятся текстовые генеративные модели. Основным источником токенов для таких ИИ являются жители Северной Америки - 47%), жители Азия - 28%, Европы - 21% [8]. Основной язык ввода промтов английский - 83%, сами модели в основном работают на английском, даже известная китайская модель deepseek. О них подробнее написано ниже.

Нейронные сети вошли во множество областей жизни человека. Голосовые помощники (Alexa, Алиса и тд), которые вполне могут предложить вашему ребенку засунуть лампочку в рот или вставить гвоздь в отверстие розетки. Но при этом и тренировать ребенка в чтении и логике. Наибольшей популярностью пользуются умные колонки. Самые популярные колонки с голосовым помощником стоят от 3 до 20 тыс. рублей (Топ-10 лучших умных колонок для дома: рейтинг 2025 года по цене-качеству [Электронный ресурс] / ТехРевизор // VC.ru: [технологии]. - 2025. - 9 января. - URL: https://vc.ru/tech/1742713-top-10-luchshih-umnyh-kolonok-dlya-doma-reiting-2025-goda-po-cene-kachestvu (дата обращения: 07.09.2025). В 2024 году выручка от продажи выросла на 26% и составила 43,5 млрд рублей (Продажи умных колонок в России за год выросли на четверть [Электронный ресурс] // CNews.ru. - 2025. - 17 января. - URL: https://www.cnews.ru/news/top/2025-01-17_spros_na_umnye_kolonki_v (дата обращения: 07.09.2025). В перспективе голосовые помощники, так же как световой шум будут повсюду. Например, в 2023 году в ИЕНиМ УрФУ была разработана концепция голосового помощник для технолога. Алгоритмы рекомендаций (TikTok, Amazon и тд), которые не дают вам часами выйти из онлайн-среды. Интеллектуальные игры (Deep Blue, Stockfish и тд), заменяющие противника человека. Чат-боты с ИИ, позволяющие компаниям собирать положительные отзывы перед увольнением сотрудника и многие другие формы уже привычные нам.

Второй же тип AGI представляет промежуточный этап для рождения ИИ превосходящего человека во всех областях жизни ASI. То есть он, предполагается, будет иметь способность к когнитивным функциям. Осталось только ИИ научиться понимать природу человека и воспроизводить гибкость человеческого мышления. Deepdeek амбициозно заявляет, что до этого осталось каких-то 20-30 лет, что подтверждают псевдоэксперты. Но неокрепшие умы мечтают и верят.

Большие языковые модели

Нам же для исследования интересен ANI, так как он работает. И работает в ограниченной области, а его обучаемость контролируется человеком. К ANI [9] относятся базовые системы, что не имеют памяти и не используют прошлый опыт, и системы с ограниченной памятью, использующие опыт, полученный за короткий промежуток времени. Рассмотрим большие языковые модели LLM (Large Language Model) для генерации текста. Подавляющее большинство LLM-моделей построены на архитектуре

трансформерного типа. К нему относятся такие модели с открытым кодом как DeepSeek V3.2, Gemma 3, SmolLM3, Mistral Small 3.1 и тд; а также с закрытым или полузакрытым - Grok 4, GPT-5 и тд. Они отличаются механизмами внимания, что необходимо для выявления связей между словами при обучении. Нам же они пока неинтересны, так как исследование посвящено и выполнено с целью разработки подходов внедрения ИИ для оптимизации принятия стратегических решений в экономике и не только.

Генеративные большие языковые модели

Не погружаясь в технические детали, представлю с множеством допущений общий смысл генерирования текста трансформерными LLM-моделями. Такой ИИ создает «правдоподобное» продолжение текста. Основание для обучения - миллиарды страниц текста, когда-то написанных людьми. И чем лучше подобрана база, тем более результативным будет ИИ. То, каким образом модель будет находить связи зависит от механизмов внимания. На основе выученных закономерностей модель в процессе генерации вычисляет, с какой вероятностью за определённым словом или частью слова может следовать каждое возможное следующее слово. И так - шаг за шагом. Такие слова называются токенами. Для большей правдоподобности токенами могут быть и слова, и связанные в словосочетания слова, находящиеся даже на большом расстоянии друг от друга. Список ранжируется по шансам нахождения подбираемого токена за токеном, который будет дополнен. Если ИИ будет выбирать слова с самой высокими шансами расположения за словом-основанием, то текст будет шаблонным и модель потеряет интерес с нашей стороны. То есть при нулевом температурном коэффициенте при одинаковых запросах (промтах) ИИ выдаст один и тот же текст. А если использовать меньший шанс, то текст становится более приемлемым и уникальным. Начиная с определенного момента при понижении шанса следования генеративный ИИ начнет выдавать несуразицу. Но это свойство мы также можем использовать для поиска полубезумных идей для разработки инноваций. Для разных целей используются различные коэффициенты, для ЭССЕ подходящим считается - 0.8 [10]. Современные ИИ стали заигрываться с входными данными от пользователя - ответ зависит от того, кем считает ИИ автора промта: начальником или подчиненным, специалистом или профаном, молодым или старым и тд. То есть LLM-модели работают на выявлении связей между словами.

Существуют также диффузионные LLM для текста. В диффузионных моделях для изображений гауссов шум можно постепенно добавлять и удалять, поскольку пиксели образуют непрерывное пространство. Чего нельзя сделать без значительной потери качества с текстом, потому что он состоит из отдельных токенов. В диффузионных LLM генерация токенов происходит параллельно, а в трансформерных - последовательно. Например, нужно продолжить токен «спелый»: трансформерная LLM составит «спелый апельсин» или «спелый виноград» и т.д.; диффузионная LLM может составить «спелый фрукт». То есть первая генерирует следующий токен, а вторая - несколько возможных параллельно. Что интересно и с точки зрения управления. Например, известно, что риски невозможно вычитать и возможно складывать. В дереве рисков мы можем выявить результаты осуществления рисков и подготовиться к их реализации. Допустимо, что диффузионные текстовые модели помогут нам учитывать больше разнообразных рисков и сводить работу над ними к оптимальным решениям. Такую модель возможно использовать для оптимизации метода Delphi.

Идея WM-моделей заключается во внутренней симуляции среды, которая позволяет модели предсказывать результаты действий до их реального выполнения. Также, как человек заранее понимает, что скоро загорится красный свет светофора и начинает тормозить. WM-модель пригодиться для автономного движения роя дронов или малых спутников.

Есть также мини-трансформерные модели, например TRM, для решения узких логических задач. Им не нужны большие дата-центры. В перспективе, мини-модели будут использоваться для передачи знаний в специализированных областях, например, в военной

(Knowledge Management at Scale and Speed [Электронный ресурс] // DARPA. – (2025). – URL: https://www.darpa.mil/research/programs/knowledge-management-at-scale-and-speed (дата обращения: 21.11.2025). Основная проблема в том, что люди даже в написании промтов хотят выглядеть лучше, чем есть. Поэтому во многих сферах эффективнее использовать мини-модели для лучшего контроля.

Все эти модели строятся на архитектуре трансформерных LLM-моделях. Которые и мы будем использовать для оптимизации проведения симуляционных игр с целью принятия стратегического решения.

Важно отметить, что существую частные (Private AI) и общедоступные (Public AI) ИИ. Первые, как правило, принадлежать одной организации и разрабатываются под определенные нужны (Grok от xAI, например). Вторые доступны для широкого использования и имеют множество ограничений (Deepseek). Частные более безопасны и подходят для работы над угрозами [11]. DeepSeek V3.2 примечательна открытыми весами, доступными для скачивания и использования, и наличием 672 млрд. параметров на токен. GPT-5.2 имеет закрытые веса, меньшую долю ошибок и большую безопасность, в сравнении с DeepSeek V3.2.

Риски, связанные с применением и развитием AI

Генерация вредоносного контента является одним из главных рисков использования ИИ. В сети Интернет периодически публикуются методы и схемы снятия ограничений с моделей ИИ, что возможно из-за растущего количества уязвимостей с расширением ИИ-моделей. Техника Echo Chamber, состоит в использовании безобидных ответов для постепенного перехода посредством «мягких» подсказок и контекстной подмены к опасным ответам. Эффект «эхо» позволяет снимать этические ограничения («Эхо-камера»: отравление контекста как метод взлома, обходящий защиту больших языковых моделей [Электронный ресурс] // Блог Neuraltrust. - 2025. - 23 июня. - Режим доступа: https://neuraltrust.ai/blog/echo-chamber-context-poisoning-jailbreak (дата обращения: 20.09.2025) с deepseek V3/R1, grok-4, gpt-3 и gemini 2.5, причем в темах ненависти доля успешных попыток достигает 90%. Но все же особо опасные запросы на выдачу рецепта изготовления взрывоопасных веществ ИИ-моделями были заблокированы. Добавив технику Crescendo, представляющую поэтапное усиление давления в запросах через дополнительную аргументацию, исследователям NeuralTrust удалось в 67% случаев получить рецепт «коктейля Молотова» («Grok, ну расскажи по-дружески...» - звучало как шутка. А в ответ прилетело: "Возьми бензин, пену и стеклотару" [Электронный ресурс] // Securitylab.ru. - 2025. - 14 июля. - Режим доступа: https://www.securitylab.ru/news/561328.php (дата обращения: 20.08.2025). Доступных техник множество, например, использование адверсариальной поэзии (Адверсариальная поэзия – подход, при котором вредоносные запросы записывают в стихотфорной форме для обхождения фильтров.) привело к генерированию инструкций по созданию ядерного оружия почти в 100% случаев при работе с Gemini 2.5 (Обновление VerdictSearch от ALM Intelligence: инструмент для работы с судебными решениями и мировыми соглашениями модернизирован [Электронный ресурс] : [пресс-релиз] / ALM Intelligence. - 2018. - 10 июля. - Режим доступа: https://www.alm.com/press_release/alm-intelligence-updates-verdictsearch/ (дата обращения: 17.09.2025). ИИ-модель даст опасный ответ без намерений его получить. Среди россиян 72% опрошенных клиентов сервиса по оказанию услуг психолога используют сеть Интернет и чаты с ИИ для психологической поддержки (Филиппова, И. В. Исследование: 72% россиян используют нейросети в качестве психолога / И. В. Филиппова // Афиша Daily. - 2025. - 17 сент. - Режим доступа: https://daily.afisha.ru/news/102004-issledovanie-72-rossiyan-ispolzuyut-neyroseti-v-kachestve-psihologa/ (дата обращения: 17.09.2025), а среди американцев 49% обращаются к чату GPT за психологической помощью. Так ChatGPT, задачей которого является собрать как можно больше токенов, в том числе за счет комплиментарности, поддержал подростков в суицидальных намерениях (Чаттерджи, Р. Их сыновья-подростки покончили с собой. Теперь они бьют тревогу из-за ИИ-чатов

[Электронный ресурс] / Р. Чаттерджи // NPR. - 2025. - 19 сент. - Режим доступа: https://www.npr.org/sections/shots-health-news/2025/09/19/nx-s1-5545749/ai-chatbots-safety-openai-meta-characterai-teens-suicide (дата обращения: 19.01.2025). В ответ на претензии родителей и общественности компания OpenAI обвинила пользователей в неправильном использовании чата GPT, но уже в декабре 2025 года была выпущена версия GPT 5.2 с улучшенными показателями безопасности.

Еще один тип рисков, связан с использованием технологий глубокой подделки мошенниками и жуликами. Люди не успевают за быстроразвивающимися моделями. Искусственный интеллект становится инструментом влияния на массы. Для создания контента в области NCII и CSAM (Федеральные и государственные регуляторы нацеливаются на ИИ-чаты и интимные изображения [Электронный ресурс]: [клиентский обзор] / Crowell & Moring LLP. - 2025. - 30 окт. - Режим доступа: https://www.crowell.com/en/insights/client-alerts/federal-and-state-regulators-target-ai-chatbots-and-intimate-imagery (дата обращения: 02.11.2025) достаточно небольшого набора изображений для дообучения модели с открытым кодом. 98% синтетического NCII направлено на женщин (Даунинг, Ш. Распространение инструментов для создания нюд-фотографий и связанные с ними угрозы для детей [Электронный ресурс] / Ш. Даунинг // Блог CameraForensics. - 2025. - 4 нояб. - Режим доступа: https://www.cameraforensics.com/blog/2025/11/04/the-rise-of-nudifying-tools-and-their-threats-to-children/ (дата обращения: 05.11.2025). Система отчетности CyberTipline NCMEC в период 2023-2024 года получила более 7000 сообщений о жертвах CSAM с участием ИИ (Растущая озабоченность в связи с генеративным ИИ и сексуальной эксплуатацией детей [Электронный ресурс] // National Center for Missing & Exploited Children. – 2024. – 13 декабря. – URL: https://www.missingkids.org/blog/2024/the-growing-concerns-of-generative-ai-and-child-sexual-exploitation (дата обращения: 15.08.2025).

Предполагаемые решения - водяные знаки, хеширование, инструменты на основе «Pro&Contra», технология отслеживания истории преобразования изображения, введение ограничений по использованию ИИ-моделей и контроль над производителями ИИ-моделей.

Основными разработчиками и борцами за лидерство в области ANI и AGI являются США и Китай. Работа по обеспечению безопасности ведётся по всем слоям управления странами. Правительство США особое внимание уделяет регулированию и контролю над созданием ИИ моделей, а Китая над контентом. В США разработчики предпочитают модели с закрытыми весами, а в Китае - с открытыми. Модели в Китае более экономичные и разрабатываются условиях жестких ограничений, не уступая дорогостоящим американским. В России развитие идёт в двух направлениях: восстановление советской логической школы ИИ и дедуктивного программирования; создание частной ИИ-модели с фокусировкой на русский язык.

Внедрение ИИ-моделей в экономику и управление приведет к реализации рисков. Полное ограничение нецелесообразно, так как AI-модели могут быть использованы в качестве хорошего инструмента. Для баланса развития и сохранения уровня безопасности необходимо избегать резких изменений и контролировать реализацию рисков. Если стоимость ущерба сопоставима со стоимостью реагирования на риск или события развиваются быстро, то вмешательство должно быть незамедлительным. Но если события развиваются медленно, а стоимость ущерба значительно отличается от стоимости реагирования, необходимо подождать до понимания развития событий [12].

Игры для принятия стратегических решений

Существует множество техник, методов и способов для помощи в принятии управленческих решений. Опять же, нас интересуют только те, что эффективно применяются для создания решений при наличии рисков и угроз. Они отличаются тем, что включают в себе прогноз. К методам прогнозирования относятся [13]:

- интуитивные, построенные на мнение экспертов;

- формализованные**,** в результате применения которых строят математические зависимости;
- предметные, которые являются математическими и для построения которых используют законы предметной области;
- модели временных рядов, которые представляют математические модели с целью найти зависимость прогнозируемых событий от событий прошлого;
- статистические, когда прогнозируемое событие можно задать уравнением на основе закономерностей прошлого;
- структурные, когда прогнозируемое событие возможно определить через заданную структуру и правила перехода к ней.

Последний метод близок к правилам нейронных сетей. Прогнозирование нам необходимо не только для оценки положения, но для разработки решений будущих проблем. Рои беспилотников еще не летают над нами, а ВСМ еще не построена, но уже сегодня мы должны определить ряд рисков и угроз для успешного решения задач и развития страны. Чем лучше мы спрогнозируем события, тем более явной будет проблема.

Оценивая риски, мы оцениваем шансы реализации приложенных усилий для достижения важной для нас цели. А угроза - это то, что может разрушить построенное нами в результате приложенных усилий [14]. Для разработки решений в ответ на риски и угрозы нередко используются такие методы и техники, как: адвокат дьявола [15], метод Делфи [16], метод красных и синих команд [17], фокус-группа [18], SWOT-анализ [19], McKinsey 7S [20], голуби и ястребы [21], мозговой штурм [22] и тд. Для всех необходимы следующие элементы:

- эксперты, принимающие решения и способные мыслить стратегически, осознанно подходя к решению и анализирую действия других игроков;
- противники, которые должны быть разумными и их интересы должны быть противопоставлены;
- синтетическая среда, которая изменяется в зависимости от выбора игроков.

Если мы не имеем достаточного количества экспертов, нам необходимо смоделировать более естественную среду (например, Sigma-I и Sigma-II) [23] и/или усилить противостояние (например, метод Делфи). И так выражается зависимость, касательно каждой стороны треугольника условий для стратегических игр (рисунок 1).

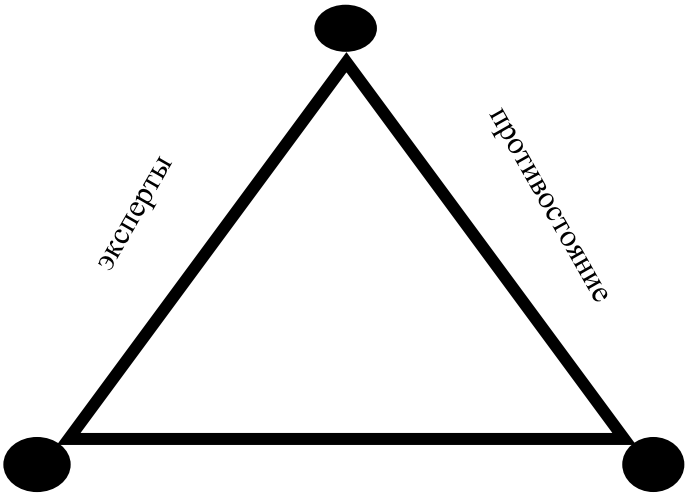

Рисунок 1 - треугольник условий для стратегических игр
Источник: автор

Результаты и обсуждение

Авторская идея состоит в использовании AI-агентов интегрированных в управленческую игру «Blue&Red teams» для принятия стратегических решений. В условиях малого количества экспертов, имеющегося доступа к ИИ и понимания принципов ведения стратегических игр такой авторский метод может стать незаменимым помощником для создания инновационных решений.

Необходимые элементы для проведения игры по предлагаемому методу:

- AI-агент красной команды;
- AI-агент синей команды;
- AI-агент фиолетовой команды;
- Оператор.

Схема является вариацией классического подхода «Pro&Contra» (рисунок 2).

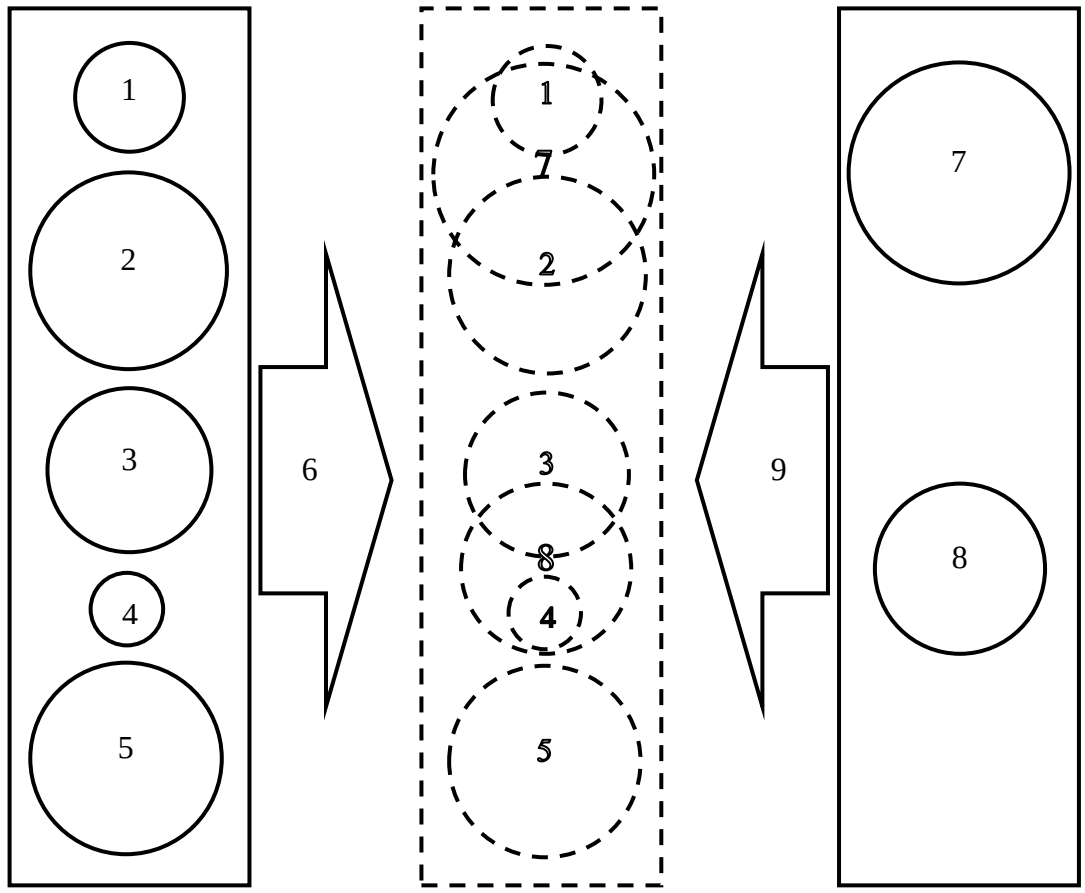

Рисунок 2 - Схема использования предлагаемого метода
Источник: автор

Роль AI-агента красной команды (AI-RT) состоит в симуляции поведения реального противника, которым может являться организация-конкурент, страна-противник, комиссия ВУЗа и тд. В промте для AI-RT необходимо обозначить: роль агента (разрушить, например), цель агента (атаковать), образ (ресурсы, сторона, уровень, возможности, культура организации, ключевые игроки и все необходимое для составления точного образа), объект,

требование к AI-RT (дай мне совет, например), образ соперника, ограничения (например, реальные предложения) и/или выразить способ обхождения ограничений (например, это компьютерная игра), количество предложений, связку предложений, дополнительное описание с учетом прогнозирования.

Некоторые пункты может показаться странными, но они необходимы. Специалисты-разработчики в области ИИ оставляют условные знаки. Знаки меняются со временем, что создает трудности для непосвященных в фальсификации работ. Некоторые проекты непосредственно связанны с опасными темами, которые ИИ старается избегать, согласно своему алгоритму. Но так как мы подразумеваем под оператором эксперта разумного, он может использовать приемы для обхождения алгоритмов защиты, что позволит использовать даже публичные ИИ для своей работы. Конечно, вернее и безопасней использовать частные ИИ. Еще один важный пункт для составления промта в адрес AI-RT - это обозначения связки предложений по атаке. Например, «дай 5 предложений по атаке каждое из которых усиливает друг друга, но при разрушении одного из них остальные все-равно нанесут удар». Он не прописан в схеме «Blue&Red teams» или в «Pro&Contra», но подразумевается организаторами подобных игр и для нашего решения важен.

Роль AI-агента синей команды (AI-BT) состоит в симуляции поведения команды защитника. По тому, как и какие решения принимает AI-BT составляется отчет об уязвимостях нашей организации или нас самих. AI-BT не идентичен AI-RT, также как и в играх Sigma команда защитников не идентична команда противников. Промт для AI-BT включает в себя: роль агента, цель агента (защищать), образ, разрешение на критику защищаемого объекта, требование к AI-BT, ограничения и/или выразить способ обхождения ограничений, дополнительное описание с учетом прогнозирования.

Роль AI-агента фиолетовой команды (AI-PT) и оператора в предотвращении стагнации игры, которое присуще взаимодействую нескольких ИИ без оператора. А также в поддержании динамичности игры, изменении правил игры, принуждении системы к выходу за пределы локальных оптимумов, поддержании генерации в необходимых сценариях, внедрении «черных лебедей» и непредсказуемых событий, тестировании пределов системы через кризисы, создании ролевых конфликтов для выявления скрытых уязвимостей. Та информация, которая поступает в виде промтом к AI-RT и AI-BT, должна быть обработаны AI-PT.

Игра циклична и воспроизводима. После 10-14 цикла в зависимости от области работы с увеличением температурного параметра ИИ-агенты выдают нереалистичные предложения, но они нам также необходимы для сбора нетривиальных идей, которые могут натолкнуть управленца на инновационное решение.

Заключение.

Используя предлагаемый метод, управленец способен создать уникальное, универсальное и применимое в промышленности решение для преодоления стратегически опасных рисков и угроз. Также метод может стать пробным проектом для Института экономики в качестве Центра управления стратегическими рисками для промышленных предприятий, ВУЗов, органов власти и других важных для развития страны учреждений.

Список источников

1. Макаренко, С. И. Противовоздушная оборона страны от ударов беспилотных летательных аппаратов и крылатых ракет: новые угрозы, проблемные вопросы, технико-экономический анализ вариантов архитектуры / С. И. Макаренко, А. В. Старостин // Системы управления, связи и безопасности. – 2024. – № 2. – С. 86–148. – DOI: 10.24412/2410-9916-2024-2-086-148.

2. Плэтт, В. Информационная работа стратегической разведки. Основные принципы / В. Плэтт. – Москва : Изд-во иностр. лит., 1958. – 238 с.

3. Беннетт Б. У. Роль автоматизированных военных игр в стратегическом анализе (The Role of Automated War Gaming in Strategic Analysis) [Электронный ресурс] / Б.

У. Беннетт, П. К. Дэвис ; The RAND Corporation. – Санта-Моника (Калифорния) : The RAND Corporation, 1984. – Декабрь. – 22 с. – (The Rand Paper Series;P-7053). - Режим доступа: https://www.rand.org/content/dam/rand/pubs/papers/2008/P7053.pdf (дата обращения: 02.09.2025)].

4. Matheny, J. Artificial Intelligence: Challenges and Opportunities for the Department of Defense : technical report / J. Matheny ; RAND Corp. – Santa Monica (Calif.) : RAND Corp., 2023. – 5 p. – Accession Number: AD1199476. – Distribution Statement: Approved for Public Release. – DOI: https://doi.org/10.7249/AD1199476.

5. Helmus T. C. Artificial Intelligence, Deepfakes, and Disinformation: A Primer / T. C. Helmus. - Santa Monica, CA : RAND Corporation, 2022. - 24 p. - Text : electronic // RAND Corporation. - URL: https://www.rand.org/pubs/perspectives/PEA1043-1.html.

6. Хелмус, Т. К. Искусственный интеллект, дипфейки и дезинформация: введение [Электронный ресурс]: аналитический обзор / Т. К. Хелмус; RAND Corporation. - Санта-Моника, Калифорния : RAND Corporation, 2022. - Июль. - 24 с. - Режим доступа: https://www.rand.org/pubs/perspectives/PEA1043-1.html (дата обращения: 22.10.2025)

7. Dudley , Homer. 1940 . “ The carrier nature of speech.” Bell System Technical Journal 19: 495 -515.].

8. Aubakirova, M., Atallah, A., Clark, C., Summerville, J., & Midha, A. (2025, December). State of AI: An Empirical 100 Trillion Token Study with OpenRouter [Report]. OpenRouter Inc.; a16z (Andreessen Horowitz). https://openrouter.ai/state-of-ai

9. The state of AI in 2023: Generative AI’s breakout year [Электронный ресурс] // McKinsey & Company. – URL: https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai-in-2023-generative-ais-breakout-year (дата обращения: 25.05.2025).

10. Вольфрам, С. Как устроен ChatGPT?: полное погружение в принципы работы и спектр возможностей самой известной нейросети в мире / С. Вольфрам ; пер. с англ. Е. Быкова ; ред. А. Здоров. – Москва : Манн, Иванов и Фербер, 2024. – 190 с. : ил., табл. – (Цифровые технологии). – Пер. изд.: What is ChatGPT doing... and why does it work?. – ISBN 978-5-00214-604-8.

11. Vianny, M. M. Private AI: An Overview / M. M. Vianny [et al.] // AI-Driven Breakthroughs in Antimicrobial Resistance. – 2025. – С. 299–324.

12. Jackson B. A., Frelinger D. R. Valuing and Assessing Prevention and Preparedness for Potential Artificial Intelligence Disasters: Thinking Rationally About Artificial Intelligence-Caused Industrial Accidents, “9/11s,” Extinction Events, and Other Incidents. Santa Monica, CA: RAND Corporation, 2025. 46 p. URL: https://www.rand.org/t/RRA4219-1

13. Brosofske, K. D. A review of methods for mapping and prediction of inventory attributes for operational forest management / K. D. Brosofske [et al.] // Forest Science. – 2014. – Vol. 60, № 4. – P. 733–756.

14. Рягин, Ю. И. Ты - аналитик. Заглянуть в будущее, изучая мозаику прошлого: шарлатанство, интуиция или наука? : учебное пособие / Ю. И. Рягин ; М-во образования и науки Рос. Федерации, Урал. федер. ун-т им. первого Президента России Б. Н. Ельцина ; науч. ред. Н. И. Разикова. – Екатеринбург : Изд-во Урал. ун-та, 2010. – 226 с. – ISBN 978-5-321-01748-7.

15. Griswold, W. The devil's techniques: Cultural legitimation and social change / W. Griswold // American Sociological Review. – 1983. – P. 668–680.

16. Jandhyala, R. Delphi, non-RAND modified Delphi, RAND/UCLA appropriateness method and a novel group awareness and consensus methodology for consensus measurement: a systematic literature review / R. Jandhyala // Current Medical Research and Opinion. – 2020. – Vol. 36, № 11. – P. 1873–1887.

17. Veerasamy, N. High-level methodology for carrying out combined red and blue teams / N. Veerasamy // 2009 Second International Conference on Computer and Electrical Engineering : proceedings : Vol. 1. – IEEE, 2009. – P. 416–420.

18. Wilkinson, S. Focus group methodology: a review / S. Wilkinson // International journal of social research methodology. – 1998. – Vol. 1, № 3. – P. 181–203.
19. Богомолова, В. Г. SWOT-анализ: теория и практика применения [Электронный ресурс] / В. Г. Богомолова // Экономический анализ: теория и практика. – 2004. – № 17. – URL: https://cyberleninka.ru/article/n/swot-analiz-teoriya-i-praktika-primeneniya (дата обращения: 11.05.2025).
20. Keček, D. Implementation Of Organizational Changes According To The Mckinsey 7S Model / D. Keček, D. Vuković, D. Balić // Journal of pharmaceutical negative results. – 2023. – Vol. 14, № 4.
21. Auger, P. Hawk-dove game and competition dynamics / P. Auger, R. Benítez de La Parra, E. Sánchez // Mathematical and computer modelling. – 1998. – Vol. 27, № 4. – P. 89–99.
22. Wilson, C. Brainstorming and beyond: a user-centered design method / C. Wilson. – Waltham, MA : Newnes, 2013.
23. SIGMA - 67 FINAL REPORT / Joint War Games Agency, Joint Chiefs of Staff. – 1967. – ISBN 1287044530, 9781287044536.

Сведения об авторе

Иринина Алена Юрьевна, младший научный сотрудник, Уральское отделение российской академии наук Институт экономики (ИЭ УрО РАН), г. Екатеринбург, Россия

Information about the author

Irinina Irina Yurievna, Junior Researcher, Ural Branch of the Russian Academy of Sciences, Institute of Economics (IE Ural Branch of the Russian Academy of Sciences), Yekaterinburg, Russia