Человек в информационном обществе

ОБЩЕСТВО И ИСКУССТВЕННЫЙ ИНТЕЛЛЕКТ: ПУТЬ К ЧЕЛОВЕКОЦЕНТРИРОВАННОМУ ПОДХОДУ

Статья рекомендована к публикации членом редакционного совета Т.В. Ершовой 25.06.20.

Абрамова Ольга Александровна

Соискатель степени phd психологических наук
Национальный исследовательский университет «Высшая школа экономики», Департамент психологии Факультета социальных наук, преподаватель
Москва, Российская Федерация
oabramova@hse.ru

Аннотация

Оптимальный сценарий безопасного и ответственного внедрения искусственного интеллекта (далее ИИ) предполагает человекоцентрированный подход – использование технологий для помощи человеку, а не для его замены. Такая стратегия позволит снизить сопротивление инновациям, страх нового в обществе и ускорит положительный эффект от автоматизации мыслительных процессов. Для создания более развитого, универсального ИИ обществу потребуется решить ряд задач: интеграция социально-психологических конструктов в технологии ИИ, внедрение этических норм в структуру ИИ, ответственность разработчиков, безопасность и полезность ИИ. Теоретическая статья рассматривает перспективы принятия ИИ обществом с описанием основных социальных рисков, сравнивает интеллект машины и человека с целью лучшего понимания роли человека при создании и распространении ИИ.

Ключевые слова

искусственный интеллект; человекоцентрированный подход; технологическое общество; риски искусственного интеллекта; полезный ИИ

Judge: So why don't you take over the world?
Entity: I wouldn't know what to do with it…Anyway,
I'm pretty busy already.[1]
Диалог с машиной в тесте Тьюринга [32]

Введение

Искусственный интеллект (далее ИИ) в последние годы привлек особое внимание социальных наук. Когда стало ясно, что техническое воплощение идеи развитого ИИ становится возможным, – последствия этих изменений и их безопасность для человека ознаменовали взлет междисциплинарных исследований современного ИИ как неотъемлемой части общества.

Рост вычислительной мощности компьютеров и скорость решения высокоинтеллектуальных задач заставили мировое научное сообщество снова пересмотреть исторические, социальные и философские концепты для понимания и предсказания влияния технологий на будущее. Создание этики ИИ, направленной на поддержание гуманистических ценностей, поиск модели инклюзивного общества с равным доступом всех людей к благам, обещаемым новыми технологиями ИИ, запустили новую волну поиска оптимальных решений и баланса в отношениях между человеком и технологиями. Важной вехой стало принятие Принципов работы с ИИ на Асиломарской конференции в 2017 году в США с закреплением этического подхода в создания ИИ

[1]Судья: Почему ты не захватишь мир?
Машина: Я не знаю, что с ним делать...В любом случае,
Я уже весьма занят(а).

– действовать исключительно для пользы и в помощь человеку [40]. Новые ценностные ориентиры апеллируют к безупречному ИИ. Реальность же обременена алгоритмическими, когнитивными и психологическими ограничениями, которые с большой вероятностью отразятся на применении и восприятии новой технологии в обществе.

Для лучшего понимания роли человека в процессах создания и распространения ИИ в обществе, данный теоретический обзор представляет дискуссионные вопросы, возникшие на международных научных площадках и в зарубежных публикациях за последние 5 лет, но все еще мало освещенные российскими авторами. Акцент делается (1) на основных социальных рисках, связанных с автономными интеллектуальными системами, включая этический аспект; (2) взаимодействии человека и ИИ при распределении потенциальных ролей в принятии решений; а также (3) перспективе принятия ИИ обществом с описанием возможной траектории развития.

1 Риски и вызовы Искусственного интеллекта - человеку и обществу

Социальными рисками, связанными с развитием ИИ называют безработицу, всеобщую праздность, потерю целей и смысла жизни, аномию общества и, как результат, рост конфликтов, социального расслоения и риск вымирания. Если будет создан Сверхинтеллект (Artificial Superintelligence (ASI)) – интеллект, в значительной степени превышающий человеческий, то сам человек может оказаться менее полезным для столь рационального и предсказуемого общества. Сохраняя оптимистический рационализм, эксперты в области ИИ считают, что человекоподобный интеллект (human-like machine intelligence (HLMI)) будет иметь положительное влияние на общество, но не исключают и риск негативного сценария. Поэтому, 48% опрошенного экспертного сообщества настаивает на приоритете исследований в области безопасного ИИ над всеми другими направлениями [15]. Предотвращение экзистенционального риска (риска выживания человеческой расы в мире развитых технологий) стало основной темой Асиломарской конференции по безопасности ИИ, организованной Институтом Будущего Жизни (FLI) в 2017 году (Калифорния, США), где в качестве потенциального решения были предложены 23 Принципа работы с ИИ, основанные на общечеловеческих ценностях и контроле ИИ человеком. Одобренные более 4000 ученых и предпринимателей принципы декларируют финансовую поддержку создания ИИ, безопасного и полезного для человечества; поощрение культуры взаимодействия и открытости между исследователями; ответственность разработчиков за морально-этические последствия создаваемых интеллектуальных систем; обязательное внедрение в автономные системы алгоритмов, защищающих ценность человека и гуманизм; долгосрочное планирование использования ИИ с учетом и предупреждением возможных рисков нанесения вреда человеку и обществу [40].

Еще одним высоковероятным социально-экономическим последствием развитого ИИ называется потеря рабочих мест. Технологическая безработица, описанная Адамом Смитом и Йозефом Шумпетером, всегда оказывалась естественным процессом усовершенствования технологий производства. Каждые 100 лет перечень востребованных профессий полностью менялся, и эти структурные изменения не вели к невостребованности человеческих ресурсов, а наоборот, помогали человеку развиваться и переходить на более сложные интеллектуальные работы [2]. В 21 веке, с внедрением ИИ в рабочие процессы, с развитием робототехники профессии не будет уничтожены, они станут другими. Многие службы исчезнут, потому что данные могут анализироваться отдельными лицами, а не корпорациями, знания будут децентрализованы. Вызов заключается в последствиях этих изменений: при использовании разумных систем все больше людей потеряет свой опыт в определенных областях. Люди будут следовать простым «машинным техникам» из-за веры, что ИИ лучше решает проблемы и, вероятно, непогрешим. Эта нисходящая спираль может сделать человека менее креативным, менее оригинальным, и может экспоненциально увеличить человеко-машинное несоответствие. Уже сейчас есть технологии, которые делают людей умнее, когда они используются, и которые заставляют чувствовать себя плохо, когда этого не делается. Важно, чтобы ИИ попал в первую категорию и не стал новым «феноменом смартфонов», от которого человек полностью зависим.

Следующий вызов жизни с ИИ – прозрачность и предсказуемость поведения человека, пользующегося рекомендательными искусственными ассистентами. Это подводит к вопросу конфиденциальности жизни человека. Вся персональная информация сохраняется в цифровых облачных ресурсах, повторяемое поведение становится предсказуемым. История положительной оценки постов друзей пользователем в социальной сети помогает определить личностные черты человека из «большой пятерки» (Big 5): экстраверсию, доброжелательность, добросовестность, нейротизм, открытость опыту [20]. Последние алгоритмы позволяют определить «большую

пятерку» черт уже по фотографии [17]. Эти данные можно использовать как для положительного воздействия на человека (например, предлагая ему желаемое место отдыха, заботу о здоровье, поиск подходящего партнера для жизни), так и для отрицательного (манипулирование без учета пожеланий пользователя, дискриминация на работе, вовлечение в неоправданные кредиты и покупки) [7][26]. Способом защиты утечки информации и одним из решений проблемы конфиденциальности называют федеративное обучение ИИ (federated learning) – метод обучения ИИ, позволяющий обрабатывать информацию в защищенной среде, без перемещения данных в общее хранилище. Так, персональные данные на смартфонах могут использоваться для глубокого обучения без их копирования, непосредственно на каждом устройстве пользователя. Отличие федеративного обучения – децентрализованное хранение данных на периферийных устройствах или серверах, без обмена ими – позволяет обрабатывать личную информацию финансового и медицинского характера, защищенную GDPR, а также данные, собираемые автономными беспилотными системами, с наименьшими рисками взлома и с сохранением приватности.

Но насколько доступ к информации возможно защитить от несанкционированного вмешательства? В настоящее время не только узкозадачность, но и технические ошибки могут привести к нежелательным последствиям. Например, не так давно создан метод, позволяющий ввести в заблуждение компьютерное зрение. Интеллектуальное программное обеспечение для распознавания изображений может быть обмануто изменением изображений таким образом, чтобы ИИ классифицировал данные как принадлежащие к другому классу [21]. Интересно, что этот метод не обманул бы человеческий глаз.

Причина, по которой сейчас социальные факторы влияния ИИ изучаются более активно, объясняется демократизацией информации и подходов к ее интеллектуальной обработке и интерпретации. С одной стороны, для конечных пользователей теперь доступно большое количество онлайн-сервисов и инструментов, с другой стороны, реальная власть и доступ к данным, общему видению картины концентрируется в руках нескольких крупных IT-компаний, владеющих цифровым следом каждого пользователя и вычислительными ресурсами для использования ИИ на более высоком уровне.

В переходные периоды между «ограниченным», узкоспециализированным ИИ к «общему», универсальному ИИ, а потом и к сверхинтеллекту остро стоят вопросы: кто будет отвечать за ошибки и неисправности? Кому и на каких условиях будут принадлежать супервозможности ИИ? Решение предстает в принятии правил обязательной «полезности» автономных систем и социальной ответственности разработчиков ИИ с оставлением человеку полномочий управления исключениями. На пути к развитому технологическому и гуманистическому обществу важен баланс власти, сопутствующих рисков и полезности ИИ вкупе с его доступностью каждому человеку.

2 Человек и Искусственный интеллект

2.1 Отличие моделей принятии решений ИИ и человека

В последние несколько лет можно наблюдать культурный сдвиг: изначально человек был ответственен за креативные, высокоинтеллектуальные задачи, в то время как машина была устройством безопасности от нежелательных сценариев, однако, сегодня роли меняются, и машины выходят на первый план. Если говорить о познании как о вычислительном процессе с символическим представлением, тогда такое вычисление может быть описано когнитивной наукой, а затем реализовано в искусственной вычислительной системе. Такая система становится центральным агентом контроля, который рационально выбирает действие, обычно заданное функцией полезности. Хотя данный подход привлекает наибольшее число исследователей ИИ, сравнение моделей естественной и искусственной когнитивных систем приводит к пониманию кардинального различия этих систем.

Интеллект, предоставляемый машинами, по-прежнему ограничен: он не имеет представления о том, что такое объект, у него нет более ранней памяти о неудавшихся попытках, он не сознателен, не имеет здравого смысла Люди лучше обобщают информацию, эффективнее действует в неконтролируемой учебной среде, решают интуитивные задачи, практически невозможные для компьютера. Действия, необходимые для выживания в критической ситуации, которые выполняются человеком без усилий, для машины крайне сложны и чаще нереализуемы. Это иллюстрирует парадокс Моравека: высокоуровневое рассуждение требует небольших вычислений, и тогда это легко для машины, в то время как простые низкоуровневые сенсомоторные навыки требуют гигантских вычислительных усилий [29]. Сегодня возможности ИИ расширяются

в сторону решения абстрактных задач, приближая ИИ к умению работать с неопределенностью. Победа алгоритмов, не использовавших человеческий опыт (например, AlphaGo Zero) над алгоритмом, обученным на человеческом опыте (AlphaGo) [34] дополнилась созданием архитектуры динамической памяти (дифференцируемые нейронные компьютеры (DNC)), способной переформировывать сложные структуры данных, увеличивая гибкость и точность ответов рекомендательной системы [16]. Еще одним вектором развития ИИ стала генерирующая сеть запросов (GQN) для идентификации объектов в пространстве: их количества, цвета и расположения в помещении с дальнейшим прогнозированием новых визуальных сцен на основе полученных данных [13].

В чем же основные отличия ИИ от возможностей человека?

Первое – это процесс обучения: даже ребенок может обучиться на нескольких примерах опосредованно через наблюдение или через непосредственный опыт [4], машина же требует в сотни тысяч раз больше итераций для достижения того же результата. Второе – человек при обучении усваивает концепции, включая сложные, нерациональные, многоуровневые, с возможностью применения их в других областях жизни. Например, человек понимает смысл игры после нескольких минут включения в нее, способен к ежеминутному переходу от игры к другим задачам и обратно, учитывая контекст и изменения, ставя новые задачи и производя общую классификацию объектов. Машина же использует паттерны для решения только конкретной задачи, например, выиграть игру, но и такая цель требует сотни тысяч партий. Третье отличие – интуитивное мышление, присущее человеку, – составляющая, включающая анализ возможных последствий принятых решений, сохранение отношений с другими людьми, долгосрочные цели, интеграция решения в разные контексты. Как отмечает Лейк с соавторами, человек способен предсказывать и объяснять, а машина – только предсказывать [23]. Четвертое отличие лежит в сфере эмоций. Эмоции отвечают за распознавание отношения к нам, выбор представителей своей социальной группы и безопасность. Для человека важно, настоящие эмоции перед ним или симуляция. Так дружба, любовь, чувство сплоченности и единения, физическая и эмоциональная боль требуют искренности и глубокого совместного переживания, которое не могут дать машины, несмотря на усовершенствование алгоритмов распознавания эмоций за последние годы [19]. Пятое отличие человека от машины – потребность в поиске смысла, подкрепляемая самоосознанием своей уникальности, вера. Поиск причины сопровождает нас до самой смерти. Экзистенциональные вопросы: кто я, как устроен мир, для чего я живу, – влияют на наши решения [1]. Человек идет дальше привычных повторяемых паттернов поведения, машина остается на этом уровне.

Другой актуальный аспект рассмотрения принятия решений человеком и машиной – это преобладание иррационального при совершении выбора человеком [25]. Как отмечает Саймон, иррациональность людей в принятии решений проистекает из высокой ресурсозатратности мыслительного процесса оптимизации и из ограниченности естественных вычислительных способностей человека [35]. Канеман и Тверски в теории перспектив подтверждают иррациональный характер принятия решений человеком и осуществление выбора в зависимости от эмоционального состояния и поиска наибольшей удовлетворенности в настоящий момент. Личные предубеждения и ограниченные знания о контексте решения позволяют человеку приходить к ложным выводам. Решения принимаются эвристически, основываясь на опыте и данных из прошлого: человек учится выбирать, получая положительные или отрицательные отзывы о результатах своей деятельности. Однако при изменении контекста и возрастании неопределенности старые привычки перестают отвечать новым условиям, и тогда усвоение новой информации человеком и адаптация механизма оценки альтернатив замедляется [18]. В этом ключе ИИ имеет ряд преимуществ перед человеком: анализ новой информации и адаптация алгоритма в большинстве задач происходит быстро, потому что устраняются любые поведенческие отклонения, присущие человеку, а при выработке оптимального рационального решения учитываются все факторы и возможные сценарии. Рекомендательные системы на основе ИИ способны снизить количество ошибок, вызываемых «человеческим фактором» в таких областях как медицина, производство, транспорт, городская инфраструктура. Новые архитектуры нейросетей разрабатываются на основе знаний об обработке информации человеком. В этом направлении многообещающая нейросеть – Transformer уже позволяет заменить традиционные рекуррентные и сверточные сети распараллеливанием – методом обработки естественного языка (NLP), основанном на функции внимания (Attention), позволяющем заменить последовательное кодирование и декодирование отдельных слов – одновременным рассмотрением всех связей между словами с расчетом веса внимания по набору значений [38]. Такие сети радикально сокращают скорость обучения, позволяют анализировать огромные массивы данных и уже используются не

только в машинных переводах (например, GPT-3), но и в компьютерном зрении [39]. Движение ИИ в сторону работы с абстрактными задачами с высокой степенью неопределенности привлекает исследователей к моделям глубокого обучения без учителя (unsupervised learning). Подобные алгоритмы на шаг приближают машины к восприятию мира человеком – через наблюдение, поиск связей и закономерностей до начала обучения с помощью кластеризации и обобщения, без использования размеченный человеком базы данных. Узкозадачность, присущая моделям обучения ИИ с учителем (supervised learning), и ограничения, связанные с малым количеством размеченных данных, могут быть нивелированы самообучением машин и открыть путь к решению более сложных задач искусственными интеллектуальными агентами.

Сейчас машины уже способны ответить на вопросы «что» и «как», безопасный ИИ должен начать спрашивать «зачем», «с какой пользой». Любая система машинного обучения хорошо обнаруживает шаблоны и помогает принимать решения, и поскольку многие алгоритмы все еще жестко закодированы, их легко понять и просчитать. Следующий уровень развития ИИ – общий алгоритм, способный создавать каузальные модели мира, как физические, так и психологические, – не так прозрачен и требует благородных намерений своего создателя [23]. Искусственные экспертные системы должны учитывать контекст и имитировать логику рассуждения человека – эксперта, ведь человек способен интегрировать среду в решение задачи, а машины – пока нет [22].

2.1 Распределение ответственности в принятии решений между человеком и ИИ. Полезный ИИ

Один из спорных, но активно обсуждаемых подходов к распределению ответственности между людьми и искусственными интеллектуальными агентами в социальных системах основан на Акторно-сетевой теории Бруно Латура, Мишеля Каллона и Джона Ло, которая объединяет машины и людей, ставя их на равные позиции в обществе. Общество описывается как сеть физических объектов, людей, технологий, социальных процессов, идей – акторов или актантов, имеющих одинаковый вес в сети, и во взаимодействии создающих живой, меняющийся социальный контекст – саму жизнь [24]. В такой системе регулирование механизмов ответственности сведено к минимуму, и каждый новый агент независимо от происхождения может стать активным актором, выполняющим работу лучше предыдущего и заменяющего его. Видение общества как самоорганизующейся сети в противовес структурированному и контролируемому подходу – возможно следующий этап развития общества. В текущих реалиях социальная интеграция ИИ предполагает активное участие человека в решениях искусственного интеллектуального агента, даже самого умного и автономного. ИИ существует в мире людей, – живой социальной системе, требующей разных уровней интерактивности, – социального интеллекта, которым машины не обладают, например, способностью к быстрому распознаванию незнакомых ситуаций [3]. Научные дебаты об ответственности за принятые машиной решения приходят к мысли оставить последнее слово за человеком. Человеческий контроль над ИИ включает: понимание особенностей алгоритма, заложенного в машину для решения конкретной задачи; выставление рамок безопасного применения машины и не введении пользователя в заблуждение путем создания иллюзии универсальности ИИ.

Речь идет о полезном ИИ (beneficial AI), требующем сознательности от создателя – его личной ответственности за последствия самообучения и использования. Полезность ИИ для человека, вшитая в изначальный дизайн интеллектуальной системы с постоянной корректировкой работы алгоритма в зависимости от изначальных целей разработчика ИИ, повлечет последующее безопасное его применение. Снижение ответственности создателей ИИ и краткосрочность целей его создания – то, что может помешать успешному диалогу «человек – машина» [5]. Бостром называет полезный ИИ «машиной с моральным статусом», где заложенный алгоритм прозрачен, а разработчики учли все негативные сценарии использования и автономной работы ИИ [6].

Стюарт Рассел из Беркли предлагает свои критерии безопасного и полезного ИИ:

- ИИ должен действовать по формальным ограниченным спецификациям,
- не создавать нежелательное поведение и функционировать в соответствии с ограничениями предыдущего пункта;
- ИИ должен быть защищен от преднамеренных манипуляций со стороны третьих сторон внешне и изнутри;
- люди должны иметь гарантированные способы восстановления контроля над ИИ при необходимости [30].

3 Принятие Искусственного интеллекта в обществе

Помимо технологической поляризации ресурсов и социально-психологических вызовов ИИ, встают вопросы адаптации человека к новому порядку: текущий контекст и восприятие ИИ обществом, социальные феномены и условия, ускоряющие или замедляющие распространение ИИ.

3.1 Коллективный интеллект и квантовые вычисления

Вполне возможно, что сверхинтеллект не будет единственным терминалом, способным принимать сложные решения, а скорее сетью терминалов. Общий ИИ вероятнее всего будет коллективным интеллектом [31]. По словам Розенберга, существующие методы формирования человеческого коллективного интеллекта не позволяют пользователям синхронно влиять друг на друга, не вызывая негативные искажения. ИИ, в свою очередь, может заполнить пробелы в синхронности и создать единый коллективный интеллект, похожий на другие виды коллективного разума, известные в природе. Например, сообщество пчел, которое при принятии решений использует большое количество единиц популяции, действующих параллельно для поиска доказательств и взвешивания альтернатив (например, при выборе нового места обитания). Каждая подгруппа пчел поддерживает свой выбор танцем, и консенсус достигается вовлечением остальных членов сообщества в лучший вариант из предложенных, что происходит при «достаточном кворуме возбуждения» и заканчивается единогласным решение всего сообщества [31].

С другой стороны, для создания следующих уровней ИИ - сверхинтеллекта необходимы квантовые вычисления. Квантовые компьютеры позволят выполнять вычисления, которые природа делает мгновенно. Разницу между традиционными и квантовыми вычислениями часто иллюстрируют «проблемой телефонной книги». Традиционный подход к поиску номера в телефонной книге происходит через ввод записи, чтобы найти правильное совпадение. Вместо этого, основной алгоритм квантового поиска полагается на так называемую «квантовую суперпозицию состояний», которая в основном анализирует каждый элемент сразу и определяет вероятностно правильный ответ.

Сегодня квантовые компьютеры присутствуют не более чем в нескольких организациях и университетах, работающих в области квантовых вычислений, в 2019 IBM представила первый коммерческий квантовый компьютер. Препятствиями распространения квантовых компьютеров сейчас являются: повышенная температура, необходимая для сверхпроводящих материалов компьютера; небольшое время когерентности, которое является временным окном выполнения вычислений; время выполнения отдельных операций; разница между правильным и неправильным ответами, которая может быть настолько мала, что ее трудно обнаружить.

3.2 Социальные роли ИИ и человека в обществе

Машина не учитывает эмоции и ценности, а понимание алгоритма, заложенного в машину для решения конкретной задачи, позволяет не вводить пользователя в заблуждение иллюзией универсальности ИИ, устанавливая рамки безопасного применения ИИ [14].

Тем не менее, согласно исследованиям, реакции человека в коммуникации с ИИ такие же, как в общении с живыми людьми: если опыт общения с чатботом превосходит ожидания, это вызывает эмоции радости, удивления, счастья, и, напротив, при неудачном общении люди испытывают разочарование и грусть [33]. С интеграцией искусственных интеллектуальных агентов в жизнь в обществе формируются ожидания, что роботы лучше людей позаботятся о пожилых людях и детях; будут эффективнее учить и лечить, а если человек не сможет найти подходящего спутника жизни или друга – роботы заполнят пустоту и удовлетворят базовые потребности в любви и принятии [37]. Перекладывание функций человека на искусственный алгоритм происходит из-за страха людей потерпеть неудачу с иллюзорным расчетом, что машины смогут решить проблемы, с которыми сам человек не справляется. Возрастающая роль, которую играют машины, и их способность влиять на людей могут в конечном счете сделать человека слабее.

Это еще один аргумент в пользу того, что социальные роли человека лучше оставить за ним, продолжая испытывать человечность на прочность, идя глубже в самопознание и духовность, следуя ценностям живого в мире вещей [27]. То есть оставаясь «моральными агентами» (moral agents) в противоположность «моральным (нравственным) пациентам» (moral patients). «Моральный агент» знает нравственные нормы и ценности и регулирует свои действия в соответствии с этим знанием и пониманием, он берет ответственность за реализацию высоких идеалов в своей жизни и действует. Отказываясь от владения этим правом и переходя в

нравственного реципиента (пассивного наблюдателя), человек может потерять часть своей идентичности [11].

3.3 Репутации ИИ в обществе

Положение ИИ в обществе, его принятие и положительная репутация во многом зависят от того, как преподносится ИИ в открытых информационных каналах игроками рынка. Каждое явление, феномен подается обществу в контексте. В социальной психологии такое обрамление называется концептуальное фреймирование или архитектура выбора. Создание истории вокруг феномена или новой технологии, акцент на полезных или вредных характеристиках, выбор референтных точек сравнения влияют на принятие нового явления обществом или его отвержение [10]. Ричард Талер и Касс Санстейн описывают архитектуру выбора человека в теории наджинга. Наджинг (nudging) – подталкивание к желаемому выбору через скрытое управление контекстом, недирективное управление мнением человека и его поведением, давая во многом иллюзорную свободу выбора [36]. Так, акцентирование на победах ИИ над человеком, многочисленные статьи о замене человека машинами наряду с фантастическими фильмами о выходе ИИ из-под контроля человека и следующей затем катастрофе может породить в обществе убежденность в опасности новых технологий и затормозить распространение ИИ в обществе. Природа человека такова, что анализ альтернатив начинается с вычисления потенциальных потерь, обещаемых новой технологией, и лишь потом преимуществ. Волны восхищения достижениями машинного интеллекта, сменяющиеся сопротивлением приходу новых технологий и страхом негативного влияния, присущи текущему состоянию общества и вызваны неполнотой информации о пути развития ИИ.

Заключение

Будущий тандем человека и ИИ должен определяться в междисциплинарных рамках: в сотрудничестве философии, психологии, социологии, культуры и технических наук [22][23][28][12]. «Основной вызов [нашего времени] – остаться человеком» в мире, где большинство процессов автоматизированы, а ИИ интегрирован в ежедневную жизнь с раннего детства [19]. Люди воспринимают себя как часть общества, связанного глубокими душевными узами с другими людьми. Человек играет по социальным правилам и избегает поведения, которое ведет к остракизму и отчуждению от групп, к которым он себя относит. ИИ может быть сконструирован так, чтобы давать видимость осознанных ответов, создавать имитацию эмоций, но создать самосознающее искусственное существо, которое автономно осуществляет выбор, мотивировано добиваться вознаграждения и признания, избегая страданий, – скорее невозможно. Список различий человека и машины – не закрытый. Здесь описана лишь часть социально-психологических вопросов взаимоотношения человека и ИИ, обсуждаемых в последние годы.

Основные аргументы объединены идеей принятия современной версии человекоцентрированного (human-centered) подхода Майкла Кули, основанного на необходимости видеть человека как главного в отношениях «человек - машина», а дизайн технологических решений – создавать вокруг потребностей пользователя, проверяя такие решения на социальную полезность и поддержку ценности человека [8; 9].

Не стать зависимыми, продолжать развиваться в культурно-ценностном аспекте, воспринимать ИИ как возможность личного качественного скачка и путь к свободе – задачи, стоящие перед человеком во время технологической трансформации общества и расцвета ИИ.

Благодарности

Статья подготовлена в результате проведения исследования в рамках Программы фундаментальных исследований Национального исследовательского университета «Высшая школа экономики» (НИУ ВШЭ).

Литература

1. Франкл В. Человек в поисках смысла. — Москва: Прогресс, 1990. с. 368.
2. Шумпетер Й.А. Капитализм, Социализм и Демократия пер. с англ. /предисл. и общ. ред. В.С. Автономова. Москва: Экономика, 1995. с. 540.

3. Abbass H. A. Social Integration of Artificial Intelligence: Functions, Automation Allocation Logic and Human-Autonomy Trust// Cognitive Computation. New York: Springer US, 2019, Vol.11, N.2, P. 159–171.
4. Bandura A. Social learning theory. Englewood Cliffs, New Jersey: Prentice Hall, 1977.
5. Baum, S.D. On the promotion of safe and socially beneficial artificial intelligence // AI & Society. New York: Springer US, 2017, Vol. 32, P. 543–551.
6. Bostrom, N. Superintelligence: Paths, dangers, strategies. Oxford: Oxford University Press, 2014.
7. Burr C., Cristianini, N., Ladyman, J. An Analysis of the Interaction Between Intelligent Software Agents and Human Users Minds and Machines. New York: Springer US, 2018, Vol. 28, N 4, P. 735–774.
8. Cooley M. Human-centered Systems. Designing Human-centered Technology // The Springer Series on Artificial Intelligence and Society. New York: Springer US, 1989, P. 133–143.
9. Cooley M. On Human-Machine Symbiosis. In: Gill S. (eds) Cognition, Communication and Interaction // Human-Computer Interaction Series. London: Springer-Verlag London, 2008, P. 457-485.
10. Cunneen M., Mullins M., Murphy F. Autonomous Vehicles and Embedded Artificial Intelligence: The Challenges of Framing Machine Driving Decisions // Applied Artificial Intelligence. Abingdon: Taylor & Francis, 2019, Vol. 33, N 8, P. 706–731.
11. Danaher J. The rise of the robots and the crisis of moral patiency // AI & Society. London: Springer-Verlag London, 2019, Vol. 34, N 1, P. 129–136.
12. Duan Y., Edwards J.S., Dwivedi Y.K. Artificial intelligence for decision making in the era of Big Data – evolution, challenges and research agenda // International Journal of Information Management. Amsterdam: Elsevier, 2019, Vol. 48, P. 63-71.
13. Eslami S.M., Rezende D.J., Besse F.,Viola F.,Morcos A.S., Garnelo M., Ruderman A., Rusu A.A., Danihelka I., Gregor K., Reichert D.P., Buesing L.,Weber T.,Vinyals O., Rosenbaum D., Rabinowitz N., King H., Hillier C., Botvinick M., Wierstra D., Kavukcuoglu K., Hassabis D. Neural scene representation and rendering//Science. Ney-York: Science Publishing group, 2018, Vol. 360, N. 6394, P. 1204-1210.
14. Gelepithis P. A.M. AI and Human Society // AI & Society. London: Springer-Verlag London Limited, 1999, Vol. 13, P. 312-321.
15. Grace K., Salvatier J., Dafoe A., Zhang B., Evans O. When Will AI Exceed Human Performance? Evidence from AI Experts //Journal of Artificial Intelligence Research. Cambridge: AAAI Press, 2018, Vol. 62, P. 729-754.
16. Graves A., Wayne G., Reynolds M., Harley T., Danihelka I., Grabska-Barwińska A., Colmenarejo S.G., Grefenstette E., Ramalho T., Agapiou J., Badia A.P., Hermann K.M., Zwols Y., Ostrovski G., Cain A., King H., Summerfield C., Blunsom P., Kavukcuoglu K., Hassabis D. Hybrid computing using a neural network with dynamic external memory// Nature, London: Springer Nature, 2016, Vol. 538, N. 7626, P. 471-476.
17. Kachur A., Osin E., Davydov D. et al. Assessing the Big Five personality traits using real-life static facial images // Scientific Reports. Nature, 2020, Vol. 10, 8487, URL: https://www.nature.com/articles/s41598-020-65358-6 (дата обращения: 02.06.2020).
18. Kahneman D., Tversky A. Prospect Theory: An Analysis of Decision under Risk // Econometrica. Cleveland: The Econometric Society, 1979, Vol. 48, N 2, P. 263—291.
19. Kile F. Artificial intelligence and society: a furtive transformation // AI & Society. London: Springer-Verlag London, 2013, Vol. 28, P. 107–115.
20. Kosinski M., Stillwell D., Graepel T. Private traits and attributes are predictable from digital records of human behavior // PNAS. Washington: National Academy of Sciences, 2013, Vol. 15, N 110, P. 5802–5805.
21. Kurakin A., Goodfellow I. J., Bengio S. Adversarial Examples in the Physical World // Technical report. Google Inc., 2016.
22. Lake B. M., Salakhutdinov R., Tenenbaum J. B. Human-level concept learning through probabilistic program induction // Science. Washington DC: American Association for the Advancement of Science, 2015, Vol. 350, N6266, P. 1332–1338.
23. Lake B. M., Ullman T. D., Tenenbaum J. B., Gershman S. J. Building Machines That Learn and Think Like People // Behavioral and Brain Sciences. Cambridge: Cambridge University Press, 2017, P. 1-9.
24. Latour B. Technology is society made durable. In: Law J (ed) A Sociology of Monsters: essays on power, technology and domination //Sociological Review Monograph. Abingdon: Routledge, 1991, Vol. 38, P. 103–132.

25. Lo A. W. The adaptive markets hypothesis: Market efficiency from an evolutionary perspective // Journal of Portfolio Management. London: Euromoney Institutional Investor, 2004, Vol. 30, P. 15–29.
26. Matz S. C., Kosinski M., Nave G., Stillwell D. Psychological Targeting as an Effective Approach to Digital Mass Communication // Proceedings of the National Academy of Science. Washington DC: United States National Academy of Sciences, 2017.
27. McCarthy E. The dynamics of culture, innovation and organizational change: a nano psychology future perspective of the psycho-social and cultural underpinnings of innovation and technology // AI & Society. London: Springer-Verlag London, 2013, 28, P. 471–482.
28. Miller T. Explanation in artificial intelligence: Insights from the social sciences // Artificial Intelligence. Amsterdam: Elsevier, 2018, Vol. 267, P. 1-38.
29. Muller V. C., Bostrom N. Future progress in artificial intelligence: A survey of expert opinion // In V. C. Müller (Ed.) // Fundamental issues of artificial intelligence. Cham: Springer International Publishing, 2016, P. 553–570.
30. Russell S., Dewey D., Tegmark M. Research priorities for robust and beneficial artificial intelligence //AI Magazine. Cambridge: AAAI Press, Vol. 36, N 4, 2015, P. 105–114.
31. Rosenberg L. B. Human Swarms, a real-time method for collective intelligence. In Proceedings of the European Conference on Artificial Life. Artificial Life. London: MIT Press, 2015, P. 658– 659.
32. Shah H., Warwick K. Machine humour: examples from Turing test experiments // AI & Society. London: Springer-Verlag London, 2017, Vol. 32, P. 553–561.
33. Shank D.B., Graves C., Gott A., Gamez P., Rodriguez S. Feeling our way to machine minds: People's emotions when perceiving mind in artificial intelligence // Computers in Human Behavior. Amsterdam: Elsevier, 2019, Vol. 98, P. 256-266.
34. Silver D., Schrittwieser J.,Simonyan K., Antonoglou I., Huang A., Guez A., Hubert T., Baker L., Lai M., Bolton A., Chen Y., Lillicrap T., Hui F., Sifre L., Van den Driessche G., Graepel T., Hass D. Mastering the game of Go without human knowledge // Nature. London: Nature Research, 2017, Vol. 550, P. 354–359.
35. Simon H. A. A behavioral model of rational choice // The Quarterly Journal of Economics. Oxford: Oxford University Press, 1955, Vol. 69, N. 1, P. 99–118.
36. Thaler R., Sunstein C. Nudge: improving decisions about health, wealth, and happiness. London: Yale University Press, 2008.
37. Turkle S. Alone Together: Why We Expect More from Technology and Less from Each Other. New York: Hachette Book Group, 2011.
38. Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A.N., Kaiser L. Attention is all you need// NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems, 2017, P. 6000–6010.
39. Anonymous authors. Paper under double-blind review as a conference paper at ICLR 2021. An image is worth 16x16 words: transformers for image recognition at scale. URL: https://openreview.net/pdf?id=YicbFdNTTy (дата обращения: 18.10.2020).
40. URL: https://futureoflife.org/ai-principles/

SOCIETY AND ARTIFICIAL INTELLIGENCE: PATH TO THE HUMAN-CENTERED APPROACH

Abramova, Olga Alexandrovna

PHD candidate
National Research University Higher School of Economics, Faculty of Social Sciences, School of Psychology, lecturer
Moscow, Russian Federation
oabramova@hse.ru

Abstract

The theoretical paper describes the main challenges of artificial intelligence (hereinafter AI) development, faced by society, analyzes the ethical standards of AI, the responsibility of the AI developers, the safety and control of AI decision-making, based on beneficial AI and human-centered approach.

Keywords

artificial intelligence, human-centered approach, risks of artificial intelligence, beneficial AI

References

1. Frankl, V. Chelovek v poiskakh smysla. [Man in a search of meaning]. Moscow: Progres, 1990. P.368.
2. Shumpeter I.A. Kapitalizm, Sotsializm i Demokratiya per. s angl., predisl. i obschch. red. V.S. Avtonomov A. [Capitalism, Socialism and Democracy, translation from English, foreword and general editing of Avtonomov V.S.] Moscow: Ekonomics Publ., 1995. P. 540.
3. Abbass H. A. Social Integration of Artificial Intelligence: Functions, Automation Allocation Logic and Human-Autonomy Trust// Cognitive Computation. New York: Springer US, 2019, Vol.11, N.2, P. 159–171.
4. Bandura A. Social learning theory. Englewood Cliffs, New Jersey: Prentice Hall, 1977.
5. Baum, S.D. On the promotion of safe and socially beneficial artificial intelligence // AI & Society. New York: Springer US, 2017, Vol. 32, P. 543–551.
6. Bostrom, N. Superintelligence: Paths, dangers, strategies. Oxford: Oxford University Press, 2014.
7. Burr C., Cristianini, N., Ladyman, J. An Analysis of the Interaction Between Intelligent Software Agents and Human Users Minds and Machines. New York: Springer US, 2018, Vol. 28, N 4, P. 735–774.
8. Cooley M. Human-centered Systems. Designing Human-centered Technology // The Springer Series on Artificial Intelligence and Society. New York: Springer US, 1989, P. 133–143.
9. Cooley M. On Human-Machine Symbiosis. In: Gill S. (eds) Cognition, Communication and Interaction // Human-Computer Interaction Series. London: Springer-Verlag London, 2008, P. 457-485.
10. Cunneen M., Mullins M., Murphy F. Autonomous Vehicles and Embedded Artificial Intelligence: The Challenges of Framing Machine Driving Decisions // Applied Artificial Intelligence. Abingdon: Taylor & Francis, 2019, Vol. 33, N 8, P. 706–731.
11. Danaher J. The rise of the robots and the crisis of moral patiency // AI & Society. London: Springer-Verlag London, 2019, Vol. 34, N 1, P. 129–136.
12. Duan Y., Edwards J.S., Dwivedi Y.K. Artificial intelligence for decision making in the era of Big Data – evolution, challenges and research agenda // International Journal of Information Management. Amsterdam: Elsevier, 2019, Vol. 48, P. 63-71.
13. Eslami S.M., Rezende D.J., Besse F.,Viola F.,Morcos A.S., Garnelo M., Ruderman A., Rusu A.A., Danihelka I., Gregor K., Reichert D.P., Buesing L.,Weber T.,Vinyals O., Rosenbaum D., Rabinowitz N., King H., Hillier C., Botvinick M., Wierstra D., Kavukcuoglu K., Hassabis D. Neural scene representation and rendering//Science. Ney-York: Science Publishing group, 2018, Vol. 360, N. 6394, P. 1204-1210.
14. Gelepithis P. A.M. AI and Human Society // AI & Society. London: Springer-Verlag London Limited, 1999, Vol. 13, P. 312-321.

15. Grace K., Salvatier J., Dafoe A., Zhang B., Evans O. When Will AI Exceed Human Performance? Evidence from AI Experts //Journal of Artificial Intelligence Research. Cambridge: AAAI Press, 2018, Vol. 62, P. 729-754.
16. Graves A., Wayne G., Reynolds M., Harley T., Danihelka I., Grabska-Barwińska A., Colmenarejo S.G., Grefenstette E., Ramalho T., Agapiou J., Badia A.P., Hermann K.M., Zwols Y., Ostrovski G., Cain A., King H., Summerfield C., Blunsom P., Kavukcuoglu K., Hassabis D. Hybrid computing using a neural network with dynamic external memory// Nature, London: Springer Nature, 2016, Vol. 538, N. 7626, P. 471-476.
17. Kachur A., Osin E., Davydov D. et al. Assessing the Big Five personality traits using real-life static facial images // Scientific Reports. Nature, 2020, Vol. 10, 8487, URL: https://www.nature.com/articles/s41598-020-65358-6 (дата обращения: 02.06.2020).
18. Kahneman D., Tversky A. Prospect Theory: An Analysis of Decision under Risk // Econometrica. Cleveland: The Econometric Society, 1979, Vol. 48, N 2, P. 263—291.
19. Kile F. Artificial intelligence and society: a furtive transformation // AI & Society. London: Springer-Verlag London, 2013, Vol. 28, P. 107–115.
20. Kosinski M., Stillwell D., Graepel T. Private traits and attributes are predictable from digital records of human behavior // PNAS. Washington: National Academy of Sciences, 2013, Vol. 15, N 110, P. 5802–5805.
21. Kurakin A., Goodfellow I. J., Bengio S. Adversarial Examples in the Physical World // Technical report. Google Inc., 2016.
22. Lake B. M., Salakhutdinov R., Tenenbaum J. B. Human-level concept learning through probabilistic program induction // Science. Washington DC: American Association for the Advancement of Science, 2015, Vol. 350, N6266, P. 1332–1338.
23. Lake B. M., Ullman T. D., Tenenbaum J. B., Gershman S. J. Building Machines That Learn and Think Like People // Behavioral and Brain Sciences. Cambridge: Cambridge University Press, 2017, P. 1-9.
24. Latour B. Technology is society made durable. In: Law J (ed) A Sociology of Monsters: essays on power, technology and domination //Sociological Review Monograph. Abingdon: Routledge, 1991, Vol. 38, P. 103–132.
25. Lo A. W. The adaptive markets hypothesis: Market efficiency from an evolutionary perspective // Journal of Portfolio Management. London: Euromoney Institutional Investor, 2004, Vol. 30, P. 15–29.
26. Matz S. C., Kosinski M., Nave G., Stillwell D. Psychological Targeting as an Effective Approach to Digital Mass Communication // Proceedings of the National Academy of Science. Washington DC: United States National Academy of Sciences, 2017.
27. McCarthy E. The dynamics of culture, innovation and organizational change: a nano psychology future perspective of the psycho-social and cultural underpinnings of innovation and technology // AI & Society. London: Springer-Verlag London, 2013, 28, P. 471–482.
28. Miller T. Explanation in artificial intelligence: Insights from the social sciences // Artificial Intelligence. Amsterdam: Elsevier, 2018, Vol. 267, P. 1-38.
29. Muller V. C., Bostrom N. Future progress in artificial intelligence: A survey of expert opinion // In V. C. Müller (Ed.) // Fundamental issues of artificial intelligence. Cham: Springer International Publishing, 2016, P. 553–570.
30. Russell S., Dewey D., Tegmark M. Research priorities for robust and beneficial artificial intelligence //AI Magazine. Cambridge: AAAI Press, Vol. 36, N 4, 2015, P. 105–114.
31. Rosenberg L. B. Human Swarms, a real-time method for collective intelligence. In Proceedings of the European Conference on Artificial Life. Artificial Life. London: MIT Press, 2015, P. 658– 659.
32. Shah H., Warwick K. Machine humour: examples from Turing test experiments // AI & Society. London: Springer-Verlag London, 2017, Vol. 32, P. 553–561.
33. Shank D.B., Graves C., Gott A., Gamez P., Rodriguez S. Feeling our way to machine minds: People's emotions when perceiving mind in artificial intelligence // Computers in Human Behavior. Amsterdam: Elsevier, 2019, Vol. 98, P. 256-266.
34. Silver D., Schrittwieser J.,Simonyan K., Antonoglou I., Huang A., Guez A., Hubert T., Baker L., Lai M., Bolton A., Chen Y., Lillicrap T., Hui F., Sifre L., Van den Driessche G., Graepel T., Hass D. Mastering the game of Go without human knowledge // Nature. London: Nature Research, 2017, Vol. 550, P. 354–359.
35. Simon H. A. A behavioral model of rational choice // The Quarterly Journal of Economics. Oxford: Oxford University Press, 1955, Vol. 69, N. 1, P. 99–118.
36. Thaler R., Sunstein C. Nudge: improving decisions about health, wealth, and happiness. London: Yale University Press, 2008.

37. Turkle S. Alone Together: Why We Expect More from Technology and Less from Each Other. New York: Hachette Book Group, 2011.
38. Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A.N., Kaiser L. Attention is all you need// NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems, 2017, P. 6000–6010.
39. Anonymous authors. Paper under double-blind review as a conference paper at ICLR 2021. An image is worth 16x16 words: transformers for image recognition at scale. URL: https://openreview.net/pdf?id=YicbFdNTTy
40. URL: https://futureoflife.org/ai-principles/