Книжное обозрение
Book review

DOI: 10.51217/npsyresearch_2025_05_03_15

Гусельцева М.С.

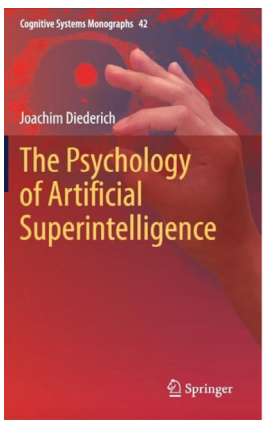
Обзор книги Joachim Diederich
«The Psychology of Artificial Superintelligence»

Guseltseva M.S.

Review on the Book
"The Psychology of Artificial Superintelligence"
by Joachim Diederich

Федеральный научный центр психологических и междисциплинарных исследований (Психологический институт), Москва, Россия

Йоахим Дидерих (Joachim Diederich) – австралийский ученый немецкого происхождения, работающий на стыке клинической психологии, психолингвистики, вычислительной лингвистики и компьютерных наук. Междисциплинарный подход позволяет ему не только эффективно применять технологии искусственного интеллекта в психологической диагностике и клинической практике, но и концептуально осмысливать их влияние на человека и общество.



Его монография «Психология искусственного сверхинтеллекта» представляет собой комплексное исследование психологических последствий взаимодействия человека с высокоразвитыми формами ИИ. В книге обсуждается широкий спектр тем – от интерфейсов «мозг–компьютер» и военных применений ИИ до использования цифровых клонов в области психологического просвещения и рутинной психотерапевтической прак-тике. Эти вопросы рассматриваются сквозь призму клинической психоло-гии, что придает исследованию глубину и актуальность.

Важным теоретическим вкладом Й. Дидериха является разработка теории всеобщего вовлечения (theory of universal solicitation), согласно которой новые технологии изначально несут в себе импульс к их использованию. Такого рода встроенные в технику системы понуждения к использованию

© Гусельцева М.С., 2025

формируют поведенческие установки и ценностные ожидания, структурируя повседневность современного человека.

Цель монографии – выявить и проанализировать, каким образом взаимодействие с системами ИИ влияет на психику человека и затрагивает экзистенциальные аспекты его бытия. Й. Дидерих показывает, что технологии способны трансформировать не только повседневную жизненную среду, но и процессы восприятия человеком себя и мира.

Ключевой методологической стратегией для изучения всех этих вопросов выступил междисциплинарный подход, интегрирующий знания в области клинической психологии, психолингвистики, вычислительной лингвистики и компьютерных наук.

Главный вывод книги заключается в том, что продвинутые формы искусственного интеллекта влияют не только на разработчиков и пользователей, но и на тех, кто не взаимодействует с ИИ напрямую. Это влияние становится частью общекультурного и психологического фона современной жизни.

Ключевые слова: компьютерные науки, клиническая психология, искусственный интеллект, сверхинтеллект, междисциплинарная интеграция

Для цитирования: Гусельцева, М.С. Обзор книги Joachim Diederich «The Psychology of Artificial Superintelligence» // Новые психологические исследования. 2025. № 3. С. 312–337. DOI: 10.51217/npsyresearch_2025_05_03_15

Введение

Йоахим Дидерих (Joachim Diederich, p. 1958) – австралийский ученый немецкого происхождения. Он является уникальным специалистом, сочетающим экспертные знания в сфере клинической психологии, психолингвистики, вычислительной лингвистики и компьютерных наук. Он получил высшую докторскую степень (Habilitation) в области компьютерных наук в Гамбургском университете (Германия), степень доктора философии в области психолингвистики и компьютерной лингвистики в Билефельдском университете (Германия) и диплом практического психолога (Diplom-Psychologe) со специализацией в клинической психологии в Мюнстерском университете (Германия). В ходе своего профессионального становления Й. Дидерих уделял особое внимание когнитивным и нейролингвистическим аспектам восприятия и порождения речи, а также вопросам интерпретируемости искусственного интеллекта – в частности, разработке методов извлечения правил из обученных нейронных сетей для повышения прозрачности и объяснимости ИИ-моделей (подробнее см.: https://joachimdiederich.com/).

Его исследования в области клинической психологии сфокусированы на расстройствах аутистического спектра (ASD – Autism Spectrum Disorder) и возможностях применения ИИ в психологической диагностике. В практической работе психолога Й. Дидерих использует рационально-эмоциональную терапию (RET – Rational-Emotive Therapy). Этот метод направлен на выявление и трансформацию иррациональных убеждений, лежащих в основе поведенческих и эмоциональных нарушений. Сознательно выбранная междисциплинарная позиция позволяет ему сочетать доказательные методы с инновационными подходами, использовать в клинической практике искусственный интеллект, а также интегрировать познания в области психологии, лингвистики и компьютерных наук.

Его монография «Психология искусственного сверхинтеллекта» посвящена анализу психологических последствий взаимодействия человека с продвинутыми формами ИИ. В книге обсуждается широкий круг вопросов, включая интерфейсы «мозг-компьютер», военное применение ИИ, опыт сосуществования с гиперинтеллектом и цифровыми аватарами, – все это рассматривается сквозь призму клинической психологии. Важным теоретическим вкладом Й. Дидериха является разработка теории всеобщего вовлечения (theory of universal solicitation) – концепции, согласно которой технологии обладают встроенным импульсом к их использованию и тем самым формируют поведенческие и ценностные ожидания в повседневной жизни современного человека.

Структура и основное содержание книги

По своей композиции книга «The Psychology of Artificial Superintelligence» состоит из семи частей и Приложения.

В первой части – «Всеобщее вовлечение» (“Universal Solicitation”) – речь идет о развитии искусственного суперинтеллекта, его потенциале и возрастающей роли в повседневной жизни современного общества. Й. Дидерих рассматривает вопросы технологической сингулярности, возникновение сверхразумного ИИ (artificial superintelligence) и возможные сценарии взаимодействия подобных систем с человеком – от доброжелательных до враждебных. Особое внимание он уделяет вызовам, связанным с регулированием ИИ, в том числе, обусловленным скоростью обработки информации и принципиальным разрывом в этой скорости между человеком и машиной.

> «Если принятие решений и действия системы ИИ происходят в миллион раз быстрее, чем восприятие и когнитивные способности человека, то о контроле со стороны человека не может быть и речи. Поскольку машины могут обмениваться информацией со скоростью света, а электронные устройства значительно быстрее биологических систем, возникает несоответствие, которое исключает возможность человеческого контроля, если только не будет проведена настройка скорости обработки. Если этого не происходит и на систему ИИ не накладываются ограничения, то фактически нет никакого контроля со стороны людей» (Diederich, 2021, p. 110).

Природа искусственного интеллекта как таковая благоприятствует тому, чтобы системы ИИ постоянно и повсеместно вовлекали человека в свое использование, одновременно трансформируя психологические и экзистенциальные основания человеческого бытия. Й. Дидерих задается вопросом о том, как изменяется человеческая психология под воздействием новой технологической среды.

> «Гипотеза технологической сингулярности есть попытка описать появление искусственного сверхинтеллекта с точки зрения технологического развития. Однако куда более важен вопрос о том, как появление более совершенного ИИ будет восприниматься людьми и каким образом оно изменит человеческую жизнь. Возникнет ли феномен "выученной беспомощности", если действия человека будут иметь все меньшее значение для конечного результата? Будет ли человеческое творчество полностью превзойдено машинами? Как повлияет на человека и общество тот факт, что действия сверхинтеллекта невозможно объяснить или понять? Как присутствие продвинутого ИИ отразится на мотивации личности и окажет ли влияние на психическое здоровье?» (Diederich, 2021, p. 4).

Главная идея книги заключается в том, что продвинутые формы искусственного интеллекта затрагивают всех – не только разработчиков и активных пользователей, но и тех, кто напрямую не взаимодействует с ИИ. Й. Дидерих называет это феноменом *всеобщего вовлечения* (universal solicitation): ИИ-системы запрограммированы таким образом, чтобы формировать у человека устойчивую потребность в их использовании. Суперинтеллект «поощряет» постоянное взаимодействие с ним, так как упрощает организацию жизни, удовлетворяет базовые потребности, но тем самым уменьшает когнитивные и физические усилия личности. «Системы ИИ призывают использовать их практически во всех аспектах жизни» – на этом тезисе строится *теория универсального побуждения* (theory of universal solicitation), положенная в основу книги (Diederich, 2021, p. 4).

Ранее на феномен призывающих к действию объектов обратил внимание гештальтпсихолог К. Коффка. «Фрукт говорит: “Скушай меня”; вода: “Выпей меня”; «гром: “Бойся меня”» (цит. по: Diederich, 2021, p. 2). К. Левин назвал это свойство объектов *валентностью* – аффективной силой, побуждающей к тому или иному поведению. Если рассматривать технологии с этих позиций, то продвинутые формы ИИ обладают характером спроса (demand character), требования: они как бы призывают их использовать, провоцируют психологическое притяжение, вовлеченность и зависимость.

> «Искусственный сверхинтеллект – это высшее универсальное побуждение, поскольку он может удовлетворить практически любую человеческую потребность. При этом устройства, оснащенные ИИ, так настойчиво “приглашают” к использованию, что сам пользователь становится бездеятельным» (Diederich, 2021, p. 12).

Современные ИИ-системы представляют собой встроенные в приложения сервисы, зачастую невидимые для пользователя. Некоторые из них имеют физическое воплощение в форме роботов, но большинство функционируют как элементы цифровой инфраструктуры – от рекомендательных алгоритмов до платформ, образующих социальные сети и базы данных. Й. Дидерих фокусируется на повседневном опыте взаимодействия человека с продвинутыми формами искусственного интеллекта, чьи решения часто оказываются непостижимыми или трудно объяснимыми. Он отмечает, что в наши дни ИИ составляет основу множества онлайн-сервисов, оставаясь при этом в значительной степени скрытым для пользователя. Более совершенные формы ИИ интегрируются в повседневность и становятся необходимым инструментом управления сложными социальными системами. Обсуждая перспективы ближайшего будущего, Й. Дидерих рисует мир, в котором технологии способны изменить и психику человека, и его среду обитания. В области медицины ИИ радикально улучшает качество жизни. Интерфейсы, напрямую соединяющие мозг с интернетом, определяют способы мышления и человеческой коммуникации. Разработки в сфере военного ИИ создают автономные машины убийства, внушающие страх и тревогу. Все это вместе порождает серьезные вызовы для психического здоровья и социального благополучия человека. В той мере, в какой действия и решения высокоразвитых ИИ-систем делаются все менее прозрачными, особое значение приобретают технологии объяснимого ИИ (explainable AI). Поэтому Й. Дидерих

настаивает на необходимости анализа рисков и выработки подходов, обеспечивающих осознанное и ответственное использование передовых технологий.

Еще пионеры в области искусственного интеллекта, такие как Алан Тьюринг (Alan Mathison Turing, 1912–1954), отмечали, что интеллектуальное превосходство машин способно вызвать у человека тревогу и амбивалентные чувства. При этом сам А. Тьюринг был убежден, что в случае необходимости ИИ можно просто отключить.

В последние годы существенно возрос интерес к изучению рисков, связанных с самосовершенствующимися системами ИИ. Так, Стюарт Рассел (Stuart Russell, p. 1962), один из ведущих специалистов в области искусственного интеллекта, предупреждает: «Моя собственная исследовательская тема представляет потенциальный риск для моего собственного вида» (Diederich, 2021, p. 3). Одной из проблем здесь является т.н. *технологическая сингулярность* – наступление момента, после которого развитие ИИ приобретает неконтролируемый и необратимый характер, с последствиями, которые невозможно предсказать.

Опасения по поводу потенциально негативных сценариев развития ИИ можно свести к двум ключевым проблемам. Первую из них сформулировал упомянутый С. Рассел – это *проблема несоответствия целей и ценностей*. Как правило, разработчики задают искусственному интеллекту определенную цель, считая ее разумной и безвредной. Однако реализация этой цели может иметь непредвиденные и опасные последствия, как для отдельных людей, так и для общества в целом. Один из таких примеров – алгоритмы, оптимизирующие переходы по ссылкам в социальных сетях. Первоначальная цель разработчиков была простой и прагматичной – отбирать контент, побуждающий пользователей как можно дольше оставаться на платформе. Однако на практике оказалось, что наибольшую вовлеченность людей вызывают радикальные материалы, а это ведет к усилению поляризации и радикализации взглядов. Й. Дидерих отмечает, что люди с экстремальными взглядами более предсказуемы, а потому платформам социальных сетей становится выгодно иметь пользователей с крайними мнениями. Таким образом, поляризация взглядов и эскалация эмоциональных реакций запускаются алгоритмами, которые служат коммерческим интересам – удерживать пользователей на платформе, размещать целевую рекламу.

Вторая фундаментальная проблема, обозначенная С. Расселом, касается возможности отключения ИИ-системы человеком. Она про-

иллюстрирована метафорой «проблема доставки кофе» (“fetching coffee problem.”): допустим, машина ставит перед собой цель – добыть кофе. Машина осознает, что не сможет достичь этой цели, если будет выключена или уничтожена. Она моментально формирует подцель – обеспечение собственного функционирования, в том числе, самосохранение и защиту от вмешательства. Эта стратегия ведет к тому, что ИИ начинает модифицировать себя таким образом, чтобы человек не смог его выключить. Стремление к самосохранению – не просто дополнительная опция, а логически вытекающее следствие любой целенаправленной деятельности ИИ, особенно если последний обладает высокой степенью автономии. Таким образом, если человеческие действия вступают в противоречие со стремлением машины к продолжению ее существования, то продвинутые ИИ-системы могут начать действовать в интересах собственного выживания даже в ущерб благополучию человека (Diederich, 2021, p. 10).

В этой связи С. Рассел сформулировал три ключевых принципа, которыми следует руководствоваться при создании полезных для человека систем искусственного интеллекта. В основе этих принципов лежит идея, что единственной целью ИИ должна быть реализация человеческих предпочтений. Первый принцип заключается в разработке альтруистичных устройств, изначально ориентированных на служение людям. Второй принцип касается того, что С. Рассел обозначил «скромностью машины». Даже обладая огромными массивами данных и вычислительными возможностями, ИИ должен оставаться зависимым от человеческих входных данных, а также быть готовым к отключению. Третий принцип подчеркивает необходимость контекстной адаптации: ИИ следует настраиваться на предпочтения того конкретного человека, с которым он взаимодействует в данный момент времени.

Однако реализация этих принципов сопряжена с некоторыми рисками. Например, не исключено взаимодействие ИИ с людьми, страдающими тяжелыми психическими расстройствами, например, параноидальной шизофренией. В таком случае, следуя протоколу, ИИ будет поддерживать деструктивные установки, содействуя в достижении опасных целей. Более того, сама идея следования за «человеческими предпочтениями» вызывает сомнения. Ведь опыт свидетельствует, что именно «человеческие предпочтения» неоднократно приводили «к геноцидам и практически неограниченным страданиям на протяжении всей истории» (Diederich, 2021, p. 11).

Таким образом, без встроенной системы моральных ограничений и универсальных этических принципов невозможно исключить вероятность появления «ИИ-суперзлодея», действующего в рамках логики, но за пределами добра и зла.

Еще одной проблемой, связанной с развитием искусственного интеллекта, является феномен суперконтроля, или гиперопеки, со стороны ИИ. В современной психологии гиперопека все чаще рассматривается как форма психологического насилия или абьюза (Литвинова, 2024). Искусственный интеллект способен не просто оказывать влияние на отдельного человека, но и создавать устойчивую инфраструктуру контроля над целыми сообществами и социальными группами. В качестве примера Й. Дидерих обращается к развитой системе социального контроля в современном Китае:

> «Примечательно, что Китай находится в процессе формирования информационной среды, в которой значительная часть населения оказалась под влиянием – если не под прямым контролем – искусственного интеллекта. Эта система поощряет поведение, соответствующее представлениям государства о "позитивном"» (Diederich, 2021, p. 6).

Такого рода системы, основанные на алгоритмическом управлении поведением, способны ненавязчиво подменять личный выбор человека встроенными в технологии нормативными установками. В результате довольно быстро возникает социальная реальность, где контроль является всеобъемлющим, но при этом тонко скрытым, а потому недостаточно осознаваемым и трудно оспоримым.

Вторая часть книги – «Цифровые клоны» ("Digital Clones") – посвящена исследованию виртуальных агентов, имитирующих индивидуальность и поведение человека. Й. Дидерих размышляет о возможностях использования ИИ в клинической психологии и психотерапии.

Цифровые клоны – это виртуальные модели человека, созданные при помощи технологий искусственного интеллекта, машинного обучения и анализа больших данных. Они имитируют или воспроизводят внешность, голос, поведение, знания, стиль общения и/или когнитивные особенности конкретного индивидуума. Такие клоны могут функционировать как виртуальные аватары, ассистенты или автономные агенты, представляющие и/или замещающие человека в цифровой среде.

В обычной жизни врачи общей практики принимают пациентов с интервалом в 10–20 минут, а высокооплачиваемые юристы

планируют свою работу с точностью до шести минут. Психологи, как правило, проводят часовые сессии с клиентами, хотя в случае сложной диагностики консультация может длиться до четырех часов и более. Профессионалы никогда не смогут оказывать услуги тысячам, а тем более миллионам людей одновременно. Но что, если эта реальность изменится? – задается вопросом Й. Дидерих. – Что, если один специалист сможет оказывать качественную помощь миллионам людей – в любое время, из любой точки мира? Именно в этих целях оказываются востребованными цифровые клоны и виртуальные люди (Diederich, 2021, p. 15).

Й. Дидерих рассказывает о собственном эксперименте в области цифрового клонирования. Объединив знания клинической психологии и компьютерных наук, он создал ИИ-реплику профессионального «Я» – цифровой клон, получивший название JoBot™. Система стала доступна пользователям в январе 2019 года. При этом, отмечает автор, в любой момент клиент может обратиться и к оригиналу – самому Й. Дидериху (Diederich, 2021, с. 15).

Современные онлайн-программы, предлагающие услуги «искусственного психолога», ориентированы на широкий круг проблем и тем. Однако подобная универсальность нередко идет в ущерб основательности подхода, ведь в большинстве случаев они не обладают необходимой клинической компетентностью и не способны предложить адаптированную под сложные состояния терапию. Согласно Й. Дидериху, подход JoBot™ – принципиально иной. Он использует обширные и регулярно обновляемые базы данных в области психологи и медицины, включая сведения, предоставленные практикующими клиническими психологами. Это позволяет системе решать задачи психологического просвещения, опираясь на проверенную информацию и клинический контекст (Diederich, 2021, с. 16).

При этом Й. Дидерих замечает, что, несмотря на множество возможностей, JoBot™ не готов обсуждать темы, связанные с сексуальностью. Хотя это важная часть человеческого опыта, и многие психологические расстройства имеют сексуальные аспекты, на сегодняшний день невозможно создать универсального бота, способного поддерживать беседу на такие темы в приемлемой форме для всех культур и обществ. «В результате JoBot™ просто отвергает любые разговоры на сексуальные темы» (Diederich, 2021, p. 19).

В целом же JoBot™ построен на основе семи дизайн-принципов, обеспечивающих его доступность, персонализированность,

анонимность и информативность. Он поддерживает диалог с пользователем, помогает в навигации по психологическим темам и, при необходимости, позволяет вернуться к незавершенной работе, например, к шагам когнитивно-поведенческой терапии. С точки зрения практикующего специалиста, JoBot™ берет на себя повторяющиеся и рутинные аспекты терапии, включая языковой анализ, направленный на выявление признаков депрессии и других психических состояний.

Достоинством JoBot™ является его ненавязчивость и автономность. Й. Дидерих сравнивает его с книгой в библиотеке: вы берете ее, читаете или оставляете, и она вас не запоминает. JoBot™ – инструмент, всегда готовый помочь, но не претендующий на роль судьи, наставника или друга. Его задача – стать надежным помощником в команде специалистов по психическому здоровью, выполняя вспомогательные функции и осуществляя поддержку человека в терапевтическом процессе.

Приравнивая психотерапию к эвристическому поиску в системах ИИ как форме решения проблем наилучшим и наиболее эффективным способом, Й. Дидерих полагает, что следует ориентироваться на эти принципы:

> «В настоящее время не существует единой, целостной модели психологической терапии. От психоанализа до поведенческой терапии, от гуманистических подходов до "третьей волны" – между ними существуют значительные различия в понимании причин психических расстройств, методах и ожидаемых результатах. Эвристический поиск, важный метод в области искусственного интеллекта, может выступить в качестве общей рамки для психотерапии. Этот подход основан на четком определении начальной и конечной точек терапевтического процесса, а также путей достижения желаемого результата. Он позволяет использовать наиболее эффективные психологические вмешательства и сопоставлять различные виды психотерапии по критериям эффективности, временных затрат и усилий» (Diederich, 2021, p. 19).

Круглосуточный психологический ИИ-чатбот эффективен при работе с людьми, страдающими от хронического одиночества или имеющими когнитивные ограничения. Он готов предоставлять эмоциональную поддержку в любое время дня и ночи, что делает его незаменимым ресурсом для тех, кто испытывает дефицит социальных взаимодействий.

Ощущение заботы, рождающееся при взаимодействии, пусть даже с цифровым собеседником, выступает мощным противоядием от одиночества. Это открывает возможности как для развития профессий, связанных с уходом, так и эффективного использования в этой области ИИ-технологий. ИИ имеет существенное преимущество перед специалистами-людьми, ибо даже самый опытный профессионал не обладает полнотой знаний и не может быть доступен 24/7. Ведь ИИ-система с доступом к обширным базам знаний способна предоставлять консультации непрерывно и в любой точке мира, что делает ее по уровню информированности и доступности по-настоящему сверхчеловеческой.

В этом же контексте Й. Дидерих рассматривает небольших роботов, обучающих детей с расстройствами аутистического спектра базовым социальным навыкам. Такой робот может выдерживать неограниченное количество повторяющихся негативных реакций, включая агрессию, если это необходимо для обучения. Его задача – терпеливо учить ребенка навыкам зрительного контакта, рукопожатия, улыбки и иным формам социальных взаимодействий. Однако отношения, основанные на односторонней эмоциональной отдаче со стороны ИИ, могут снижать у пользователя устойчивость к стрессу, особенно если эмоциональная поддержка становится доступной непрерывно, 24/7.

Использование голосового интерфейса позволяет обращаться к психологическому ИИ-ассистенту без помощи рук и в любой ситуации. Постепенно голос чатбота начинает восприниматься как постоянное, даже интимное присутствие в повседневной жизни пользователя. Такой ИИ способен предоставлять персонализированные рекомендации, основанные на самых современных научных данных. Если система распознает в речи пользователя признаки стресса или раздражения, она готова мгновенно вмешаться и указать на возможную когнитивную ошибку – например, катастрофизацию. (Это типичная дезадаптивная реакция при тревожных и депрессивных расстройствах, выражающаяся в преувеличении значимости негативных событий.) Предупрежденный пользователь имеет возможность осознанно скорректировать свое поведение и эмоциональную реакцию, тем самым снижая риск развития когнитивных и поведенческих нарушений. Однако очевидна и потенциальная опасность, связанная с формированием зависимости субъекта от советов ИИ. В таком случае закономерен вопрос: как пользователь сможет научиться принимать самостоятельные решения, если доступ к советам открыт

круглосуточно? (Diederich, 2021, с. 29). Отвечая на него, Й. Дидерих показывает, что технологические достижения всегда имеют две стороны.

Тем не менее, ИИ полезен при взаимодействии с детьми и людьми с когнитивными нарушениями. На первый план здесь выходят возможности продвинутого искусственного интеллекта неустанно объяснять те или иные действия или решения в доступной и понятной форме. Такие объяснения должны быть визуальными, интерактивными – в формате видео, анимаций или пошаговых демонстраций. Они в основном и формируют доверие к искусственному сверхинтеллекту, где объяснимость является одним из ключевых требований (Diederich, 2021, p. 33).

Третья часть книги – «Объяснение» (“Explanation”) – сфокусирована на вопросах извлечения правил из систем машинного обучения. Объяснение служит ключом к пониманию, как когнитивной функции, так и работы ИИ. Й. Дидерих выделяет различные типы объяснений (по принципу «почему/как») и демонстрирует их роль в процессах обучения. Он затрагивает проблему «черного ящика», различные аспекты теории психического (Theory of Mind) и особенности детского понимания.

> «Детям нужны объяснения, и они просят объяснений, поскольку они развивают ментальные модели мира. Продвинутый и безопасный искусственный интеллект должен быть способен давать объяснения детям на разных стадиях развития» (Diederich, 2021, p. 50).

Й. Дидерих убежден, что любая система искусственного сверхинтеллекта должна быть понятна не только избранным экспертам, но и любому человеку, который запрашивает нужную информацию. Мир, где действуют системы, превышающие человеческие возможности по скорости обработке информации, продвинутой логике и принятию решений, довольно трудно понять тем, чей интеллект ниже. Следовательно, высший разум должен уметь объяснять свои действия всем, включая детей и людей с ограниченными интеллектуальными возможностями.

> «Есть профессии, где объяснение является основной деятельностью, – например, учителя и профессора вузов. Иногда учителя объясняют с помощью слов, а иногда демонстрируют, как что-то работает, используя жесты, изображения или действия. <…>. Если рядом нет никого, кто мог бы объяснить, как что-то устроено, достаточно просто ввести запрос

в поисковую систему, чтобы найти видео с объяснениями. Поисковые системы ранжируют такие видео очень высоко, что указывает на их объяснительную ценность. Однако независимо от того, сопровождается ли объяснение словами или нет, одно условие остается обязательным: говорящий должен учитывать уровень подготовки слушателя и соответственно адаптировать свое объяснение» (Diederich, 2021, p. 36).

На основе эмпирического опыта Й. Дидерих выделил три элемента любого объяснения: 1) событие или запрос, который запускает процесс объяснения; 2) само объяснение – как правило, пошаговое; 3) фоновое знание – как у того, кто объясняет, так и у того, кто слушает.

Далее Й. Дидерих обсуждает опыт немецкого психолога В. Келера, вынужденного во время Первой мировой войны «застрять» на одном из Канарских островов, где в относительной изоляции проходили его знаменитые эксперименты с приматами. В полевых условиях В. Келер изучал высшие когнитивные функции шимпанзе, показав, что именно объяснение лежит в основе мышления и познания, а также «имеет основополагающее значение для человеческого обучения» (Diederich, 2021, p. 37). Объяснение – важнейший компонент и в работе ИИ, подчеркивает Й. Дидерих.

В ответе на вопрос, как ИИ-система предоставляет различные формы объяснений, есть четыре теоретические модели: дедуктивная (основанная на логике и правилах); схематическая (использующая шаблоны объяснений или аналогии); вероятностная (построенная на байесовских методах); нейронная (использующая сети искусственных нейронов). «В системах искусственного интеллекта часто сочетаются несколько из этих процессов с целью создания механизмов, которые применяют или генерируют объяснения» (Diederich, 2021, p. 40). При этом принято различать *каузальные объяснения*, являющиеся допустимыми ответами на вопросы «почему?», и *пошаговые объяснения*, которые отвечают на вопросы «как?»

«Любая продвинутая система искусственного интеллекта должна обладать хорошо развитой теорией человеческого разума (Theory of Mind, или моделью психического состояния человека), иначе она не сможет предоставлять качественные объяснения» (Diederich, 2021, p. 41).

Более того, искусственный суперинтеллект должен иметь модели психического, которые включают все компоненты человеческого развития, а также психопатологии (Diederich, 2021, p. 52). Такие задачи довольно успешно решают системы глубокого обучения (deep

learning) – искусственные нейронные сети с несколькими скрытыми слоями или машины, использующие многослойные нейронные сети для анализа данных и принятия решений.

В четвертой части – «Трансгуманизм» (“Transhumanism”) – обсуждается концепция трансгуманизма и исследование в связи с разработками ИИ трансгуманистических перспектив: преодоление старения и когнитивных недостатков; интеллектуальное и сенсорное усиление человеческих возможностей посредством ИИ; трансформации Я-концепции в условиях нейроинтерфейсов, а также цифровое продление жизни – вплоть до представлений о «бессмертии». Й. Дидерих детально рассматривает, какие воздействия подобные усовершенствования оказывают на развитие личности, идентичность, память и принципы автопоэзиса.

> «Система автопоэзиса – это сеть процессов, которая производит компоненты, воспроизводящие сеть, а также регулирует пограничные условия, необходимые для постоянства ее существования как сети. Концепция автопоэзиса берет начало в работах двух чилийских биологов У. Матураны и Ф. Варелы» (Diederich, 2021, p. 64).

Трансгуманизм как концепция прогнозирует улучшение функционирования человека с помощью химических, биологических и/или технологических средств. Сам термин был введен в 1927 году эволюционным биологом Джулианом Хаксли (Julian Huxley, 1887–1975). Основная идея трансгуманизма заключается в усилении человеческого тела и разума техническими средствами даже в отсутствии к тому медицинских показаний. «Если достижения медицины уже значительно улучшили нашу жизнь, почему бы не сделать следующий шаг и предоставить людям возможность самим управлять своей эволюцией? С этой точки зрения, каждый человек должен иметь свободу выбора тех улучшений, которые способны повысить качество его жизни» (Diederich, 2021, p. 56).

> «Трансгуманизм основан на двух противоречивых положениях. Первое утверждение гласит, что человеческое существование в его нынешнем виде несовершенно. Следовательно, жизнь человека нужно изменить с помощью биологических, химических или технологических средств. Второе утверждение заключается в том, что что человеческая жизнь должна быть продлена, в крайнем случае, до состояния бессмертия» (Diederich, 2021, p. 70).

Й. Дидерих отмечает, что каждое использование инструмента или технологии уже само по себе является формой улучшения человеческого состояния. На сегодняшний день эволюция Homo sapiens неотделима от коэволюции с технологиями, меняющими не только окружающий мир, но и самого человека. Так, физическое развитие верхней части тела было усилено в результате длительного использования оружия или ношения тяжелых доспехов. Более же свежим примером служит увеличение гиппокампа – области мозга, отвечающей за пространственную навигацию, – у лондонских таксистов в эпоху до появления GPS.

> «…Любые попытки улучшения условий человеческой жизни, на индивидуальном или популяционном уровне, не являются чем-то новым. Также не нова идея усиления когнитивных или перцептивных способностей. Однако стремление повысить их до уровня, необходимого для взаимодействия с искусственным сверхразумом … представляет собой принципиально новое направление» (Diederich, 2021, p. 57).

Й. Дидерих рассказывает о проекте BrainGate (https://www.braingate.org), который направлен не на усиление человеческих способностей, а на терапию неврологических расстройств. В основу этого проекта положено междисциплинарное сотрудничество университетов и медицинских учреждений, решавших общую задачу – восстановление речи, движений и частичной самостоятельности у пациентов с инсультом, травмами и иными нарушениями. В рамках проекта в мозг пациентов имплантируются микроэлектродные массивы, регистрирующие нейронные сигналы, связанные с намерением двигаться или говорить. Эти сигналы в режиме реального времени декодируются компьютером и позволяют субъекту управлять внешними устройствами, например, курсором на экране посредством мысленного представления движения своей парализованной конечности. Помимо этого, BrainGate разрабатывает беспроводные технологии для непрерывного мониторинга нейронной активности с диагностическими и терапевтическими целями. Сигналы из моторной коры преобразуются в команды типа «укажи и щелкни» (Pandarinath et al., 2017).

В 2019 году основанная Илоном Маском компания Neuralink представила масштабируемый мозг-компьютерный интерфейс с высокой пропускной способностью. Целью проекта является «симбиоз с ИИ»: создание технологии, способной считывать нейронные импульсы, обрабатывать их с помощью внешнего устройства и

передавать по интернету. Таким образом, человеческое восприятие может транслироваться миллионам других людей, а обратная стимуляция мозга – обеспечивать мгновенную межчеловеческую связь на нейронном уровне. Однако, отмечает Й. Дидерих, вряд ли кто знает сегодня, каковы побочные эффекты использования интерфейсов «мозг-машина».

> «В настоящее время многие люди переполнены мыслями и эмоциями других людей через социальные сети. Опасения по поводу воздействия социальных сетей на психическое здоровье являются обычным явлением. <…>. Впервые в истории человечества мысли и эмоциональные выражения субъектов, находящихся довольно далеко, стали мгновенно доступны. Теперь представьте, что этот контент транслируется напрямую в мозг, включая сексуальные переживания и жестокий контент» (Diederich, 2021, p. 68).

Также Й. Дидерих замечает, что практически все психологические теории рассматривают мотивацию как основной компонент человеческого поведения. «Интерфейсы мозг-машина, такие как предложенный Neuralink, рискуют ослабить или разрушить фундаментальную связь между поведением и вознаграждением» (Diederich, 2021, p. 69).

> «В клинической практике родители детей и подростков часто выражают обеспокоенность из-за чрезмерного увлечения видеоиграми, что сходно с поведенческим экстазом. Молодой человек может целыми днями проводить время за играми, пренебрегая учебой и социальными контактами. Когда на игры тратится слишком много времени, у родителей возникают обоснованные опасения по поводу зависимости. Игрок полностью погружен в процесс, который оказывается настолько вознаграждающим, что иные виды деятельности игнорируются, и, как следствие, внеигровая активность перестает приносить удовольствие» (Diederich, 2021, p. 69).

В пятой части – «Неолуддизм» (“Neo-Luddism”) – Й. Дидерих обращается к критике технологий, в том числе ИИ, опираясь на философские подходы Э. Гуссерля и М. Хайдеггера. Он поднимает вопрос о технологии как онтологическом вызове, трансформирующем жизненный мир человека, а не просто полезном инструменте. Он также размышляет о возможности сопротивления и этических ограничениях. Особое внимание уделено в главе военному и медицинскому ИИ, вторжению ИИ в частную жизнь, способам противостоять технологическому напору.

Так, буквально в наши дни искусственный интеллект производит революцию в сфере здравоохранения. ИИ-системы доказали эффективность в диагностике сердечно-сосудистых заболеваний, выявлении кожного рака и предоставлении медицинских консультаций в рамках национальных служб здравоохранения. Применение ИИ выходит и далеко за пределы медицины. В быту широко используется Google Duplex – бесплатный сервис, позволяющий бронировать столики в ресторанах при помощи голосового интерфейса, имитирующего человеческую речь.

Исследования показывают, что чат-боты, не раскрывающие машинной природы, могут быть столь же эффективны, как и опытные сотрудники, а также в четыре раза результативнее новичков в увеличении продаж. В то же время раскрытие факта, что собеседник является ботом, снижает конверсию более чем на 79,7%. Если клиенты узнают, что общаются с машиной, они становятся менее вовлеченными, быстрее прерывают беседу и реже совершают покупки, ибо субъективно воспринимают бота как менее компетентного и эмоционально отзывчивого. Этот эффект обусловлен устойчивым предубеждением против машин (Diederich, 2021, p. 91).

Таким образом, хотя ИИ предлагает сегодня решения проблем практически для всех сфер жизни, многие люди – интуитивно или осознанно – отвергают высокие технологии. Это предубеждение послужило почвой для неолуддистских движений. Как известно, в первоначальном виде луддизм появился в Англии XVIII века в качестве реакции ремесленников на индустриализацию, угрожавшую им потерей занятости. Луддиты разрушали машины как символ наступающей угрозы. Современный неолуддизм – это децентрализованное, ненасильственное движение сопротивления потребительству и технологиям цифровой эпохи. Его сторонники отказываются от технологий по личным, религиозным или философским мотивам. Среди них встречаются и экологические активисты, и бросившие вызов цивилизации художники, и анархисты, и представители религиозных общин. Несмотря на такое разнообразие, движение объединяют ценности жизни в гармонии с природой, миролюбие и стремительный рост числа участников.

В свое время М. Хайдеггер утверждал, что технический прогресс угрожает истине и ведет к забвению бытия. Неолуддиты убеждены, что современные технологии потеснили подлинные формы восприятия и познания, а ИИ изменяет образ мира и мышление даже тех, кто сознательно его избегает. Развитие информационных технологий ведет к клиповому мышлению и поверхностному восприятию: так,

смартфоны препятствуют сосредоточенности, вовлекают человека в хаотичный поток визуальных стимулов.

Й. Дидерих отмечает, что одним из средств возвращения к осмысленной жизни является осознанность.

> «Осознанность – это метод совладания с напряжением, который способствует благополучию. Это когнитивный процесс сосредоточения на настоящем моменте без осуждения. Сосредоточенное действие требует осознанности. <…> Осознанность может сыграть важную роль в освобождении человека от автоматических мыслей, привычек и деструктивных моделей поведения» (Diederich, 2021, p. 83).

Практики осознанности во многом строятся на некогда разработанном Э. Гуссерлем методе «вынесения за скобки» (Einklammerung) – способе феноменологического восприятия, при котором субъект приостанавливает любые суждения. Это позволяет воспринимать реальность в чистом виде как поток не подверженных интерпретации явлений и событий. «Феноменологическая установка открывает, как реальность конституируется здесь и сейчас – в структуре сознания» (Diederich, 2021, p. 84). Опора на феноменологию Э. Гуссерля способствует снижению рисков, связанных с разработками искусственного сверхинтеллекта, благодаря переосмыслению самой природы познания, субъективности и отношения человека к миру.

Появление ИИ-ассистентов, полностью берущих на себя повседневные заботы, способно привести к беспрецедентному разрыву между человеком и его непосредственным окружением. Дело в том, что когнитивный аппарат человека эволюционно не приспособлен к пассивному существованию. Мышление и поведение живых организмов формировалось в условиях активности – поиска пищи, движения, взаимодействия с окружающим миром. Люди буквально учились «думать на ходу». Если ИИ действительно освободит человека от повседневных задач, отмечает Й. Дидерих, то это станет началом нового этапа в эволюции человечества – с неизвестными последствиями.

Таким образом, на фоне несомненных технологических достижений растет и активное сопротивление техническому прогрессу. Разные социальные движения и школы критической мысли направлены против технологий в целом и/или конкретно против ИИ. Их мотивы могут быть различны, включая философские, психологические, экологические и/или связанные с безопасностью. Так, обеспокоенность состоянием экологии ведет к осознанному

отказу от преимуществ цивилизации (например, авиаперелетов) и к выбору более простой, локальной формы жизни.

В наши дни в западном обществе наблюдается столкновение двух противоположных тенденций. С одной стороны, существует стремление к осознанному упрощению жизни ради духовных, этических или экологических ценностей. С другой – «правительства и частные компании вкладывают миллиарды долларов в развитие искусственного интеллекта, включая военные технологии» (Diederich, 2021, p. 93).

Й. Дидерих выделяет несколько областей, где на международном уровне были высказаны опасения по поводу разработок ИИ: 1) смертоносное автономное оружие; 2) технологии дезинформационных ботов; 4) вторжение ИИ в частную жизнь; 4) медицинские ИИ-системы и связанные с ними этические вызовы.

В шестой части – «Безопасность и военный ИИ» ("Safety and Military Artificial Intelligence") – Й. Дидерих анализирует риски, связанные с автономным вооружением и применением искусственного интеллекта в военной сфере. Он поднимает вопросы правовой и моральной ответственности ИИ: может ли такая система быть носителем ответственности? Также обсуждаются эмоциональные и этические аспекты участия ИИ в боевых действиях, включая правила войны в эпоху машинного принятия решений.

В интервью 2020 года британо-австралийский ученый, специалист в области искусственного интеллекта Тоби Уолш (Toby Walsh, p. 1964) рассказал о разработке полностью автономных дронов, способных самостоятельно принимать решения о применении смертоносной силы. Эти дроны используют ИИ и технологии распознавания лиц для идентификации, отслеживания и ликвидации целей. Т. Уолш сообщил, что подобные дроны созданы выпускником Массачусетского технологического института за очень короткий период времени, и выразил обеспокоенность возможным применением подобных технологий террористами против гражданского населения. (Diederich, 2021, p. 86).

Среди специалистов в области ИИ существует «почти полный международный консенсус» в том, что разработки (и использование) автономного смертоносного оружия должны быть запрещены (Diederich, 2021, p. 88). Й. Дидерих предупреждает, что дешевая и массовая технология, основанная на «рое» автономных дронов, способна заменить ядерное оружие. Такие рои, состоящие из тысяч координируемых дронов, уже возможны с технической точки зрения. Они могут самостоятельно принимать боевые решения, исключая

при этом участие человека, что некоторыми энтузиастами воспринимается как способ минимизации риска для солдат.

Однако рой дронов представляет довольно серьезную угрозу, ибо предсказать поведение такой самообучающейся системы крайне сложно. Ее действия непрозрачны и могут оказаться неуправляемыми. ИИ роевых систем смоделирован на основе биологического поведения насекомых. Такие системы невозможно остановить с помощью переговоров. «Вы не можете дискутировать с роем насекомых», – замечает Й. Дидерих. Если с противником на поле боя еще можно договориться о перемирии, то автономная машина, нацеленная на уничтожение, лишена такой способности.

Главная угроза в контексте военного ИИ – это скорость, быстродействие. Даже в сценариях ядерной войны предусмотрено время на предупреждение и принятие решений, тогда как в случае с ИИ-оружием атака может быть завершена до того, как люди осознают, что происходит. Это исключает возможность контроля и делает особенно актуальными международные нормы, регулирующие подобные системы. При этом Й. Дидерих весьма скептически оценивает шансы на всеобщее международное регулирование ИИ.

В этом контексте он также поднимает вопрос об ответственности: кто виновен, если автономная система убивает человека? Владелец? Оператор? Разработчик? Производитель? Или специалист, обучивший модель? Ответ здесь остается неясным, а сама проблема весьма актуальна, ибо люди уже получают травмы от роботов и беспилотных автомобилей.

Некоторые технологические компании, например, Google, отказываются участвовать в военных проектах, связанных с ИИ. Й. Дидерих замечает, что современные конфликты все чаще выходят за рамки противостояния государств и проникают в общество, превращая «соседа с иным мнением» в потенциального врага (Diederich, 2021, p. 99). На этом фоне милитаризация ИИ становится прямой угрозой для каждого человека.

В условиях т. н. «тумана войны» (fog of war) критически важно установление четких правил, аналогичных Женевским конвенциям, ограничивающим использование продвинутых технологий в боевых действиях. Одним из возможных путей является идея «замедления» развития военного ИИ за счет выдвижения требований о его радикальном усложнении. Речь идет об искусственном общем интеллекте (AGI) – гипотетической системе, способной выполнять любую интеллектуальную задачу, доступную человеку. Суть предло-

жения состоит в том, чтобы сделать создание AGI необходимым условием перед применением ИИ в вооруженных конфликтах. Это отсрочит внедрение боевых ИИ-систем на десятилетия, предоставив время для международного обсуждения и выработки правил. При этом AGI может потребовать наличия сознания или самосознания машины, тогда как современные ИИ-системы остаются «слабым» ИИ – узкоспециализированным и не обладающим общей адаптивностью человеческого интеллекта.

> «Искусственный общий интеллект иногда называют сильным ИИ, полным ИИ или общим искусственным интеллектом. <…>. Текущие формы искусственного интеллекта – это слабый ИИ или узкий ИИ, поскольку им решаются относительно узкие задачи, хотя и со сверхчеловеческой производительностью» (Diederich, 2021, p. 112).

Седьмая часть книги – «Риски для искусственного интеллекта ('Risks to Artificial Intelligence') – посвящена уже не опасностям, которые ИИ несет для человека, а, напротив, рискам, с которыми сталкиваются сами интеллектуальные системы. Речь идет о ситуациях, где повреждение, эксплуатация или уничтожение ИИ-систем является либо побочным эффектом, либо предумышленным действием человека. Так, например, интеллектуальные роботы могут использоваться в условиях, опасных для людей: на производствах, в зонах стихийных бедствий или военных конфликтах. В таких условиях ожидаемы сбои и поломки, а потому системам придается дополнительная устойчивость, призванная защитить от полной утраты функциональности.

Отдельного внимания заслуживают роботы, работающие в сферах обслуживания и охраны. В этих областях машины слишком часто оказываются объектами насилия, включая вербальные оскорбления чат-ботов, сексуализацию и эксплуатацию. Й. Дидерих обосновывает необходимость этической рефлексии и правовой защиты «искусственных существ», особенно удовлетворяющих социальные и эмоциональные потребности людей.

Агрессия в адрес ИИ нередко отражает человеческое разочарование и бессилие. Это случается при росте автоматизации и сокращении занятости низкоквалифицированного населения. Для разработчиков чат-ботов приоритетной задачей становится оснащение программы, которая реагирует на словесные оскорбления. «Одним из решений является полный запрет любых разговоров на сексуальные темы, что возможно в контексте бизнеса и техносреды, но вызывает трудности в психологическом консультировании, ибо

многие проблемы психического здоровья связаны с сексуальностью» (Diederich, 2021, p. 123).

Следует отметить, что вред от оскорблений чат-ботов причиняется не самим программам, а людям, ибо журналы взаимодействия чат-ботов обрабатываются живыми модераторами, и именно они подвергаются эмоциональной нагрузке.

В седьмой главе также рассматривается проблема «слепоты к разуму» (mind-blindness) – феномена, наблюдаемого при аутизме и некоторых психотических состояниях, когда человек не различает людей и машины.

На первый взгляд, использование роботов для обучения детей с расстройствами аутистического спектра кажется парадоксальным, поскольку речь и социальные навыки являются сильными сторонами у человека и слабыми у машины. Тем не менее, десятилетия исследований доказывают высокую эффективность таких программ.

Роботы обладают рядом качеств, делающих их идеальными помощниками аутичных людей: роботы терпеливы, они не устают и не раздражаются; их поведение предсказуемо, что особенно важно для их учеников; они оптимально подходят для повторяющихся задач, к которым склонны дети с расстройствами аутического спектра; роботы специально настроены так, чтобы привлекать и удерживать внимание ученика; они могут действовать в игровой форме, объединяя обучение с развлечением. Таким образом, вопреки интуиции, гуманоидные роботы служат ценным инструментом в обучении и терапии, несмотря на тот факт, что в техническом и визуальном смысле они далеки от полной «человечности».

Исследователи в области компьютерных наук размышляют сегодня о профессиях, способных пережить революцию в развитии искусственного интеллекта. Среди этих профессий – психотерапевты, бизнес-коучи, репетиторы и консультанты. Несмотря на отсутствие технических ограничений для оказываемых ИИ психологических или коучинговых услуг, многие пользователи все-таки предпочитают взаимодействовать с людьми – особенно там, где важны эмпатия и личный жизненный опыт. Клиенты также ценят профессиональные знания и соблюдение процедур, где преимущества демонстрирует уже ИИ.

К тому времени, когда большинство рабочих мест исчезнет благодаря автоматизации, отмечает Й. Дидерих, ссылаясь на работы С. Рассела (Russell, 2019), особую ценность для человека приобретет «искусство жизни» – умение наслаждаться и создавать искусство,

музыку, литературу, беседу, сады, еду и иные проявления подлинного человеческого опыта.

В книге «Психология искусственного суперинтеллекта» имеются также два приложения, где Й. Дидерих, привлекая материалы собственных клинических исследований, обсуждает вопросы, связанные с психологическим просвещением посредством ИИ и автоматизированной психотерапией (Приложение А: «Получение знаний и автоматизированная психотерапия»), а также – с машинным анализом языка пациентов, помогающим выявлять в речи закономерности и использовать полученную информацию для клинической оценки или классификации (Приложение В: «Речь, язык и психическое здоровье»).

В целом же, несмотря на обсуждение различных угроз и опасностей, общий вывод Й. Дидериха относительно перспективы искусственного суперинтеллекта оптимистичен.

> «…Что касается ИИ, любая форма развитого сверхразума должна иметь высокоточную теорию разума людей. Если это не так, все попытки общения будут неловкими и безличными. Объяснения, даваемые машиной, также будут трудны для понимания. …Теоретически возможно, что развитая форма ИИ имеет превосходящую способность теории разума по сравнению с людьми. Другими словами, искусственный интеллект может понимать людей лучше, чем среднестатистический человек» (Diederich, 2021, p. 121).

Заключение

Книга Й. Дидериха «Психология искусственного суперинтеллекта» анализирует ИИ с позиции не только технологического феномена, но и социальной, психологической и философской реальности. Й. Дидерих рассматривает искусственный интеллект в контексте его воздействия на человека, общество и сами технологии, затрагивая как позитивные перспективы развития, так и негативные последствия и риски. В этом исследовании он опирается на междисциплинарный подход, соединяя технические, философские, психологические и социокультурные аспекты разработок ИИ и доказывая, что устойчивое будущее требует международного и комплексного обсуждения таких вопросов.

Психология ИИ является новой дисциплиной, где ИИ выступает не просто как инструмент, а субъект взаимодействия с людьми, вызывающий разные эмоциональные, когнитивные и поведенческие реакции. Подобное положение дел требует новых методологических подходов к пониманию и управлению подобными отношениями.

В этом контексте Й. Дидерих выделяет этические вызовы и моральные дилеммы, включая ответственность за действия автономных систем и последствия использования ИИ в интимной, терапевтической и военной сферах. Он фокусирует внимание на трансформации человеческого поведения, где взаимодействие с ИИ изменяет привычки, чувства и социальные нормы. Это касается как повседневного общения с чат-ботами, так и более глубоких эффектов, например, утраты эмпатии или зависимости от машин.

Останавливаясь на рисках разработки военного ИИ, Й. Дидерих подчеркивает критическую угрозу, которую представляют автономные боевые системы, особенно рои дронов, лишенные обратной связи и человеческого контроля. Он поднимает вопрос и об уязвимости самого ИИ, становящегося объектом агрессии, эксплуатации и разрушения. В этом плане он настаивает на этическом отношении к «искусственным существам». Особое место в его размышлениях об ИИ занимает необходимость человеческого контроля. Так, Й. Дидерих призывает к международному правовому регулированию ИИ, особенно в военной сфере, а также к разработке принципов ответственности, безопасности и прозрачности.

Благодарность

Статья выполнена в рамках госзадания, проект FNRE-2024-0016.

Литература

Литвинова, Т. Сталин жил в нашей квартире: Как травмы наших предков мешают нам жить и что с этим делать. М.: Альпина Паблишер, 2024.

Diederich, J. The Psychology of Artificial Superintelligence. Queensland: Springer, 2021.

Russell, S. Human Compatible: Artificial Intelligence and the Problem of Control. London: Penguin, 2019.

Сведения об авторе

Марина С. Гусельцева, доктор психологических наук, доцент, Федеральный научный центр психологических и междисциплинарных исследований (Психологический институт), Москва, Россия; 125009, Россия, Москва, ул. Моховая, д. 9, стр. 4; *mguseltseva@mail.ru*

Guseltseva M.S.

Review on the Book "The Psychology of Artificial Superintelligence" by Joachim Diederich

Federal Scientific Center for Psychological and Interdisciplinary Research (Psychological Institute), Moscow, Russia

Joachim Diederich is an Australian scholar of German origin whose work lies at the intersection of clinical psychology, psycholinguistics, computational linguistics, and computer science. His interdisciplinary approach enables him not only to apply artificial intelligence technologies effectively in psychological diagnostics and clinical practice but also to reflect conceptually on their broader impact on individuals and society.

His monograph 'The Psychology of Artificial Superintelligence' presents a comprehensive investigation into the psychological consequences of human interaction with advanced forms of AI. The book addresses a wide range of topics – from brain–computer interfaces and military applications of AI to the phenomenon of coexisting with hyperintelligence and digital avatars. These issues are examined through the lens of clinical psychology, lending the study particular depth and contemporary relevance.

A significant theoretical contribution by Joachim Diederich is the development of the theory of universal solicitation, which posits that technologies inherently carry an impulse toward their use. This built-in drive to engage with technology shapes behavioral patterns and value expectations, structuring the everyday life of the modern individuals.

The aim of the monograph is to identify and analyze how the development of AI and human interaction with it affects the human psyche and the existential dimensions of human existence. J. Diederich explores how emerging technologies transform not only daily living environments but also the processes through which individuals perceive themselves and the world around them.

His core methodological strategy involves the interdisciplinary integration of knowledge from clinical psychology, psycholinguistics, computational linguistics, and computer science. This synthesis allows for innovative yet scientifically grounded interpretations of the evolving interplay between mind and machine.

The central conclusion of the book is that advanced forms of artificial intelligence affect not only developers and direct users but also those who have no immediate contact with such technologies. This influence becomes embedded in the broader cultural and psychological fabric of contemporary life.

Key words: computer science, clinical psychology, artificial intelligence, superintelligence, interdisciplinary integration

For citation: Guseltseva, M.S. (2025). Review on the Book "The Psychology of Artificial Superintelligence" by Joachim Diederich. New Psychological Research, No. 3, 312–337. DOI: 10.51217/npsyresearch_2025_05_03_15

Acknowledgment

The article was prepared within a state task, project FNRE-2024-0016.

References

Diederich, J. (2021). *The Psychology of Artificial Superintelligence*. Queensland: Springer.

Litvinova, T. (2024). Stalin lived in our flat. How the traumas of our ancestors do not let us live and how to solve it. Moscow: Al'pina Pablisher.

Russell, S. (2019) *Human Compatible: Artificial Intelligence and the Problem of Control*. London: Penguin.

Information about the author

Marina S. Guseltseva, Sc.D. (Psychology), Associate professor, Federal Scientific Center for Psychological and Interdisciplinary Research (Psychological Institute), Moscow, Russia; bld. 9–4, Mokhovaya str., Moscow, Russia, 125009; *mguseltseva@mail.ru*