УДК 004.932

Д. В. Даровских

Искусственный и естественный интеллект: вызовы и этика

Аннотация:

Представлен взгляд на глобальные вопросы, связанные с развитием технологий искусственного интеллекта (ИИ). Он рассматривается как некий агент позитивных изменений. Взяв на себя часть нашей мыслительной деятельности, ИИ помогает понять ценность другой уникальной стороны человека, так называемого «человеческого фактора». Это не описывается механизмами или закономерностями, но в данной работе показаны значимость, возможность и своевременность практического применения «человеческого фактора», в частности, к исследованию и разработке ИИ. Для поступательного развития технологии необходим контроль над возможными рисками, которые должны быть выявлены с тем, чтобы предложить безопасные обходные решения. Этическая сторона рассматривается и в качестве инструмента фиксации предложенных решений (через нормативные документы), и с позиции совершенствования морали. Результатом работы выступает концептуальная модель.

Ключевые слова: Этика ИИ, Объяснимый ИИ, Сильный ИИ, когнитивные вычисления

Об авторе: Даровских Дмитрий Владимирович, Государственный университет «Дубна», магистрант кафедры системного анализа и управления института системного анализа и управления.

Мир охвачен глобальным трендом, связанным с разработкой Сильного искусственного интеллекта (далее – ИИ) и внедрением уже полученных результатов. Возможность технологического прорыва в этой области стимулирует конкурентную борьбу между государствами за лидерство. Не случайно более 30 стран разработали национальные стратегии развития ИИ. В такой ситуации возрастает роль проработки возможных рисков, связанных с активным становлением технологии. Это позволит спокойно и в полной мере осваивать преимущества ИИ и внедрять их в те области, где это действительно необходимо,

например, в достижении целей устойчивого развития, утвержденных большинством стран мира.

В России под ИИ понимается комплекс технологических решений, позволяющих имитировать когнитивные функции человека и получать сопоставимые или превосходящие результаты. ИИ может включать субъективную позицию своего разработчика, при этом по определению превосходить его в интеллектуальных способностях. Поэтому так важно обратиться к познанию и исследованию когнитивных функций человека до их внедрения в Сильный ИИ. Не меньшее значение имеют нормы социального взаимодействия. Даже нормы однозначно определенные, выраженные в законодательных нормативах, обнаруживают множество нестыковок, что мы покажем далее. Заявленное руководством страны повсеместное внедрение ИИ повышает значимость проработки законодательных норм.

В качестве объекта данной работы выступает система, которая включает в себя взаимодействие людей и ИИ. Исследование нацелено на формирование концептуальной модели разработки технологии ИИ, отвечающей достижениям общечеловеческих целей.

Мировое сообщество уже ведет активную работу и над этическими вопросами, и над их технологическими решениями. Мы предлагаем взглянуть на вызов ИИ с другого ракурса, фокусируясь больше на самом разработчике ИИ и его потребителях как носителях естественного интеллекта.

1. Когнитивный аспект

«Сильный ИИ» – пока гипотетическое понятие, но поиск решения ведется в ряде сквозных дисциплин, таких как когнитивная психология и коннекционизм. Отслеживание этого процесса важно для выявления возможных рисков и обнаружения путей их фиксации.

Когнитивная психология позволяет рассматривать и изучать работу мозга (не только человека, но и других живых существ, например, муравьев или улиток), начиная с крупных структур, и постепенно с помощью декомпозиции переходить к структурам более мелким. Поэтому вход в проблему со стороны когнитивных наук дает, в первую очередь, возможность увидеть стратегические принципы объединения и совместного функционирования более специализированных участков.

Если пытаться описывать мозг как машину, занимающуюся вычислениями, ясно, что она совершенно другого рода, чем все известные сегодня рукотворные технические устройства. Главное отличие состоит, несомненно, в эволюционном и онтогенетическом развитии, а также в обилии элементов — по некоторым данным общее число нейронов

головного мозга превышает 100 миллиардов, число же их специализированных соединений, синапсов оказывается на два-три порядка больше. Все это ведет к массивной параллельности нейрофизиологических процессов, сочетающейся, впрочем, с определенной анатомической дискретностью и функциональной специализацией мозговых структур [1, с. 153].

Проведенные исследования выявили, что различные нейроны человека имеют свою специализацию: даже при распознавании форм изображений часть откликается на квадраты, часть на круги. Но работают группы нейронов по нескольку сотен, решая определенную задачу. Глубокое машинное обучение действует похожим образом.

Согласно коннекционизму, интеллект возникает из вычислений в нейронах и нейронных сетях. Из этого следует, что ограничения глубоких нейронных сетей, если таковые имеются, несущественны, и они должны постепенно уменьшаться до нуля с появлением новых более глубоких и сложных архитектур. Первоначально эта точка зрения, казалось, подтверждалась очевидным успехом глубоких нейронных сетей, но в последнее время она ставится под сомнение. В нейробиологии появляется все больше свидетельств того, что знание мозга не коррелирует с пониманием поведения. Даже, если бы нам удалось полностью смоделировать человеческий мозг, мягко говоря, нет никаких гарантий, что симуляция будет делать именно то, что мы ожидаем. Некоторые из последних интерпретаций коннекционизма указывают на него, как на биологически правдоподобный способ объяснения и реализации интеллекта. Но обратное распространение, фундаментальный алгоритм для обучения, в первую очередь сделавший коннекционизм реализуемым, явно не является биологически правдоподобным, а элементы глубоких нейронных сетей, которые таковыми являются, имеют серьезные ограничения. Предлагаются новые подходы: символьные представления, алгоритмические репрезентативные модели, концептуальные представления о высших когнитивных функциях, комбинация подходов и т.д. Все эти предложения выступают попыткой подхода более высокого уровня к интеллекту, чем к нейронам, поскольку ученые все чаще признают, что коннекционизм не является ответом. Воспроизведение вычислительной способности мозга не приближает нас к модели настоящего интеллекта, как предполагалось ранее [16].

Что же выступает главным отличием человека от ИИ? По мнению ряда исследователей – мотивация. Именно она придает качество действиям. Пока у ИИ нет мотивации, он остается инструментом [4]. По словам Э. Юдковского, пока философия не дает безупречную этическую основу, попытки реализации мотивации ИИ могут допускать

множество потенциально вредных сценариев. Активное развитие технологии и вектор на моделирование когнитивных функций ведет в том числе к реализации ИИ «по своему образу и подобию». В Лаборатории когнитивных архитектур МФТИ занимаются разработкой модели искусственной психики роботов на основе реверс-инжиниринга принципов работы мозга. Прототип назвали ADAM. Вполне вероятны эксперименты с «человеческим фактором».

Попытки моделирования иррациональной стороны человека чреваты рисками. Могут неподконтрольно проявляться отрицательные следствия, такие как нецензурное общение или обман. Прецеденты существуют. По той же причине особый контроль нужен и для NLP как технологии, работающей с естественным языком – носителем человеческого иррационального в том числе. Другим подобным риском является компания Neuralink, созданная И. Маском с целью определенной превентивной защиты. Идея состоит в инвазивном внедрении в мозг неких усилителей когнитивных функций, а, в особенности, вариантов прямого взаимодействия с компьютером. Все вариации ИИ, в том числе «черный ящик» имеют некую модель, работающую на определенных математических принципах. А вот расширение возможностей мозга компьютерными технологиями может столкнуться с «человеческим фактором». Тогда сложно предсказать развитие ситуации.

Рассмотрим риски, сопутствующие становлению новой технологии.

Основным из них является попытка внедрения «человеческого фактора (иррационального)» в системы ИИ различными путями: отражением субъективной позиции разработчика, моделированием мотивации, обучением по созданным на естественном языке данным и даже инвазивным внедрением компьютерных систем в мозг (Neuralink). Стандартные механизмы регулирования могут не справиться с «человеческим фактором», усиленным искусственным интеллектом. И, в отличие от сложных, но все же основанных на математических принципах систем, здесь слаба вероятность что-либо предсказать.

2. В границах применимости

Существует мнение, что прорывы в сфере глубокого обучения привели к расцвету ИИ. В 2015 г. (рис. 1) на ежегодном турнире по точности распознавания изображений с помощью глубоких нейронных сетей было представлено несколько модификаций новой архитектуры, разработанной китайским подразделением компании Microsoft (ResNet, Inception ResNet, DenseNet).

Рисунок 1. Хронология архитектур DL

Они оказались настолько удачными, что превзошли возможности среднестатистического человека. С этого момента стало наращиваться индустриальное внедрение глубоких нейронных сетей наряду с другими разработками в этой области.

Успешные сети имеют сложные многослойные архитектуры. Но в основе лежат вполне определенные математические принципы, и простую сеть, например, реализованную для восстановления скрытой зависимости, можно настроить методом подбора. В больших сетях, конечно, подобное уже не получится. Там применяются математические методы градиентного спуска и Байесовской оптимизации. Последняя считается хорошим решением для оптимизации гиперпараметров (архитектур и опций учебного алгоритма), а также в специфических случаях (при прерывистых и недифференцируемых случаях) [2]. Другими словами, глубокие нейронные сети, называемые «черными ящиками», не приобретают каких-либо метафизических свойств, просто их уровень сложности выходит за границы, приемлемые для нашего осознания.

Результативность глубокого обучения стимулирует разработки в этой области, но сложность интерпретации устанавливает свои ограничения. Объяснимость критически важна для таких областей, как здравоохранение, клиническая и судебная работа, так как они потенциально имеют дело с человеческими жизнями, а не просто с анализом прибыли и затрат. Избежание риска происходит, когда ответственность возлагается на человека, профессионала в данной области. Подразумевается, что он может использовать ИИ в качестве помощника. Следовательно, модели, разработанные для экспертов (например, в здравоохранении), должны быть доступны для их понимания [12].

В случае с глубокими нейронными сетями довольно уместной выглядит аналогия с написанием программного кода. Если стиль понятен, например, обусловлен проверенными стандартами и снабжен необходимыми комментариями, ценность готового решения возрастает, поскольку повышаются возможности модификации, тестирования и

реинжиниринга. Объяснимый ИИ (XAI), помимо навеянного ключевого преимущества, также может быть интереснее и в долгосрочной перспективе. Чем лучше удастся понять код, тем качественнее будет реализовано достижение конечных целей. Для этих целей решение снабжается определенным числом пояснительных модулей.

Часть исследователей обращают внимание на тот факт, что изощренность и опыт пользователя системы ИИ будет влиять на уровень необходимого объяснения. Другие исследователи выступают за компромисс между полнотой описания и точностью работы системы. Здесь имеется в виду не просто попытка объяснить действия, а именно сделать их понятными для пользователя. Следует обратить внимание и на предвзятость людей к простым описаниям, чтобы не заставлять разработчиков вместо объяснимых систем делать убедительные [12]. А. Н. Аверкин выделяет 4 принципа объяснимого ИИ, основанных на проекте института NIST (от августа 2020 г.): 1) объяснение (способность системы ИИ его предоставить); 2) значимость (объяснения, понятные отдельным пользователям); 3) точность (суть процессов должна быть достоверно отражена, но зависит от контекста); 4) пределы знаний (система должна работать только в условиях, для которых была разработана, и отмечать внештатные ситуации).

Было показано, что интерпретируемость моделей машинного обучения обратно пропорциональна его гибкости (точности), а нейронные сети, вероятно, подвержены этому больше других моделей. Отладка такого алгоритма создает серьезные проблемы. Многие приложения используют каскады глубоких нейронных сетей для решения сложных задач, где выход одной передает значения на вход другой. Если вернуться к попытке имитировать функционал человеческого мозга, то стоит его воспринимать не как большую нейронную сеть, а, скорее, как сеть сетей, и разработки сильного ИИ могут потребовать исследования иерархий глубоких нейронных сетей. Такие системы, возможно, не удастся отладить [16].

В качестве одного из решений была предложена структура для ускорения объяснимого машинного обучения с использованием блоков тензорной обработки. Структура использует синергию между сверткой матриц и преобразованием Фурье. Реализуются преимущества естественной способности блоков тензорной обработки ускорять матричные вычисления. Разработчики поясняют, что подход применим к широкому набору алгоритмов машинного обучения и его эффективное использование может привести к интерпретации результатов в реальном времени. Экспериментальные результаты показывают, что ускорение во время классификации в среднем 25х, а во время интерпретации в среднем 13х по сравнению с аналогами [14].

В качестве другого решения могут выступать квантовые вычисления, либо их комбинация с машинным обучением. Прорывы в таких проблемах, как факторизация, задача «коммивояжера», сверхплотное кодирование, телепортация стали возможны благодаря квантовым вычислениям. Это уникальная технология, одним из свойств которой является ускорение.

Две подсистемы, каждая из которых находится в смешанном хаотическом состоянии (и с отличной от нуля энтропией) при слиянии в единую систему образуют чистое (с нулевой энтропией) состояние, обладающее высшим уровнем порядка (эффект квантовой самоорганизации). При этом количество информации в целой системе меньше, чем в каждой из ее составляющих подсистем, а взаимная энтропия имеет отрицательное значение. Квантовая суперпозиция, состоящая из двух классических взаимоисключающих логических состояний, позволяет образовать одно единое состояние, содержащее, например, логически противоречащие «да» и «нет» (Кот Шрёдингера) [10].

Глубокие нейронные сети в классической реализации оперируют дискретными данными. Внедрение математического аппарата квантовых вычислений снимает это ограничение благодаря принципу суперпозиции. Тогда квантовые нейронные сети могут быть более близкой аппроксимацией мозга, и взглянуть на проблему объяснимости можно под другим ракурсом. Бессознательная часть, которая работает и имеет свои настроенные связи между нейронами, соотносится с неинтерпретируемым ИИ (но точным), а сознание – с интерпретируемым ИИ (цена также может выражаться в точности, скорости…). Таким образом, решение вопроса объяснимости может быть снова скопировано у природы: на начальных этапах требуется интерпретируемость, но по реализации достаточного тестирования акцент может делаться на совершенствование. Согласно В. Б. Хозиеву, при освоении навыков ответственные за них нейронные связи сначала формируются в коре головного мозга (под контролем сознания), но при длительном совершенствовании уходят в подсознание, освобождая тем самым кору для новых задач.

3. Этический аспект

Как говорил Л. Д. Ландау, «не все можно представить, но все можно понять...». Сегодня трудно представить, к чему может привести повсеместное внедрение ИИ, поэтому стоит работать над пониманием этого вопроса. Конечно, в идеальном варианте – постепенное внедрение с проработкой возможных сценариев, моделированием и сопоставлением с реальной ситуацией при контроле над внедрением со стороны государств, но даже более существенно – со стороны обществ.

В 2019 г. В. В. Путин высказался о дискуссии в области социальных аспектов и последствий использования ИИ как о важной теме и предложил профессиональному сообществу, компаниям подумать над формированием свода этических правил взаимодействия человека с ИИ. Активная работа с Советом Европы, направленная на создание всеобъемлющей конвенции, призвана обеспечить установление общих принципов для разработки и применения ИИ на глобальном уровне во благо всего человечества.

ЮНЕСКО выступает с предложением о разработке всеобъемлющего глобального нормативного акта, призванного обеспечить ИИ прочную этическую основу, которая будет не только защищать, но и содействовать соблюдению прав человека и уважению человеческого достоинства. После принятия этот документ станет этическим ориентиром и глобальной нормативной основой, позволяющей обеспечить соблюдение принципа верховенства права в цифровом мире [9].

С учетом активного внедрения систем ИИ в различные сферы деятельности в IEEE запустили глобальную инициативу для исследований в области этики ИИ. Результатом таких исследований должны стать технические нормативные документы, регламентирующие разработку и внедрение систем ИИ с требованиями к их этическому поведению. Первым таким документом стал проект общих рекомендаций для разработчиков ИИ. Называется документ весьма примечательно – «Ethically Aligned Design» (в вольном переводе – «Этически обусловленное проектирование»). В нем собраны основные ближнесрочные угрозы, связанные с внедрением автономных систем на базе ИИ, которые сегодня отмечены в научной литературе. Помимо перечисления угроз IEEE обращает внимание на необходимость изменений в подготовке специалистов – разработчиков программных продуктов, использующих технологии ИИ. В целом представленный документ является одним из первых шагов к переносу рассуждений об этике ИИ из области научных исследований в практическое русло. Очевидно, что этот документ сам по себе пока обозначает круг проблем и дает только первичные идеи по их решению, однако это уже весомая основа, на которой будут строиться последующие (в том числе нормативные) документы, разрабатываемые IEEE [3]. Этически обоснованное проектирование включает восемь разделов, каждый из которых касается конкретной темы, связанной с ИИ/АС, а каждая тема подробно обсуждается конкретной рабочей группой Глобальной инициативы IEEE [15].

Таким образом, ведутся работы над нормативами в области ИИ – и на государственном, и на международном уровнях. Общество также активно реализует свою позицию, в частности, относительно правильности ответственных и открытых разработок

в области ИИ. Так, ряд разработчиков во главе с Беном Гертцелем основали OpenCog в качестве проекта с открытым исходным кодом, направленного на развитие инструментария для разработки ИИ. Илон Маск и Сэм Альтан создали OpenAI, некоммерческую исследовательскую организацию по ИИ для открытых исследований.

Соглашения и правила по безопасности и этике могут быть приняты заинтересованными сторонами, чтобы обеспечить их соответствие взаимно принятым стандартам и нормам. Однако опыт работы с рядом международных договоров, например, об изменении климата, соглашения о лесе и рыболовстве показал, что автономия и суверенитет вовлеченных сторон могут затруднить мониторинг и соблюдение нормативных требований. Таким образом, чтобы все могли пользоваться преимуществами безопасного, этичного и надежного ИИ, очень важно разработать и внедрить соответствующие стратегии стимулирования, чтобы обеспечить взаимную выгоду и соблюдение требований безопасности со всех сторон.

Хорошей идеей может стать моделирование последствий международных нормативов. Тут появляется большая вариативность, например, в качестве данных можно использовать законодательства разных стран, принимающих соглашение, или их судебную практику. Ряд исследователей предложили для моделирования эволюционную теорию игр (EGT). Большое значение могут иметь и различные формы стимулирования этического поведения [13].

В 2020 г. число документов в системе КонсультантПлюс превысило 200 миллионов. Следует учитывать, что большая часть является производной от нормативов (путеводители, комментарии, обзоры, тематические журналы), но документов с исполнительной силой не менее нескольких миллионов. Также происходит периодическое обновление самих нормативов и их связей. Даже в масштабах одной страны контроль и порядок в таком объеме информации, описанной естественным языком, представляется крайне сложной задачей. В такой ситуации неизбежны нестыковки. В частности, научная группа, под руководством З. А. Кучкарова, занимавшаяся концептуальным анализом существующих Российских нормативов с акцентом на экологическую нишу, выявила, что, например, такие краеугольные понятия как «устойчивость рубля», «ликвидность» и «инфляция» не имеют однозначных определений, что дает возможность их вольной трактовки [5; 6]. Некоторые подзаконные акты, на которые ссылаются другие нормативы, вообще не существуют.

Другой стороной вопроса является распределение полномочий. Существующая правовая система (РФ) сложилась в результате многократных изменений, что привело к множеству дефектов. Дефекты в распределении полномочий и сфер деятельности

вызывают конфликты интересов и неэффективность органов исполнительной власти. Например, природоохранные и природоресурсные функции государства направлены на достижение различных целей. В связи с этим, передача полномочий двух типов одному органу власти (сейчас – Министерству природных ресурсов и экологии РФ) неизбежно ведет к внутреннему конфликту и невыполнению одной из функций [7].

4. За границами применимости

В последнее десятилетие наблюдается расцвет глубокого обучения: увеличение объема размеченных данных, совершенствование методов их сбора и рост вычислительной мощности для их обработки. В некотором смысле современная революция машинного обучения больше связана с большими данными, чем с нейронными сетями, неспособными к качественному обобщению и экстраполяции.

Онтологический вывод – это способность человека делать выводы на основе чувственного восприятия. Падение яблока на голову робота никогда не заставит его открыть гравитацию. Хотя робот может регистрировать статистические закономерности падения предметов. Атомы и электроны никогда не были и, вероятно, никогда не будут напрямую наблюдаться, но их существование может быть выведено косвенно из других наблюдений. Эта способность онтологического обнаружения, которой обладают люди, вероятно, самый важный фактор для научного прогресса и процветания человечества. Сказать, что глубокие нейронные сети не имеют доступа к этой более высокой функции, было бы преуменьшением. Онтологический вывод настолько вне досягаемости любой математической формулировки и вычислительной модели, что это никогда не обсуждается в какой-либо современной науке, будь то когнитивная психология, нейрофизиология, информатика и даже современная философия [16].

Социальное взаимодействие человека с другими людьми подвержено постоянному изменению в ходе культурной эволюции, ведущей к неограниченному росту сложности. Понимание механизмов самоорганизации в этой сложной системе – одна из главных задач для ИИ. Возможно, стоит рассматривать ИИ как инструмент, способный помочь лучше понять человека [11]. В компаниях, занимающихся ИИ, возникает рекурсивный эффект: сотрудники развивают ИИ, а технология, в свою очередь, оптимизирует их организационную структуру.

С развитием технологий глубоких нейронных сетей появляются предпосылки их превосходства по критерию экономической целесообразности во многих областях деятельности. ИИ – инновационная технология, и для корректного освоения требуется внедрение инновационных бизнес-моделей, когда организации разворачиваются вокруг

выбранной эволюционной цели. Один из важнейших моментов – общие ценности сотрудников компании. Практикуется открытость и всеобщее вовлечение в задачи. При таком подходе естественным образом отсеиваются люди, не разделяющие цели компании. Финансовый фактор остается, но он перестает быть приоритетным. Самое главное, что индивидуальная энергия каждого сотрудника растет, когда он отождествляет себя с целью более масштабной, чем его собственная, личная [8, с. 359]. Тогда человек начинает вкладываться в работу не только из рациональных соображений, но подключает свой энтузиазм.

Современная наука недоверчива к эмоциям, предположительно, препятствующим рациональному мышлению. Альтернативная точка зрения порой чревата другой крайностью – полным отрицанием левополушарного, аналитического подхода в качестве метода принятия решений в пользу правополушарного наития. Предлагаемая инновационная парадигма обладает счастливой возможностью использовать все методы познания: и аналитические, и те, что находятся за пределами фактов и цифр [8, с. 67].

Заключение

Результат тематического исследования представлен в виде концептуальной модели (рис. 2).

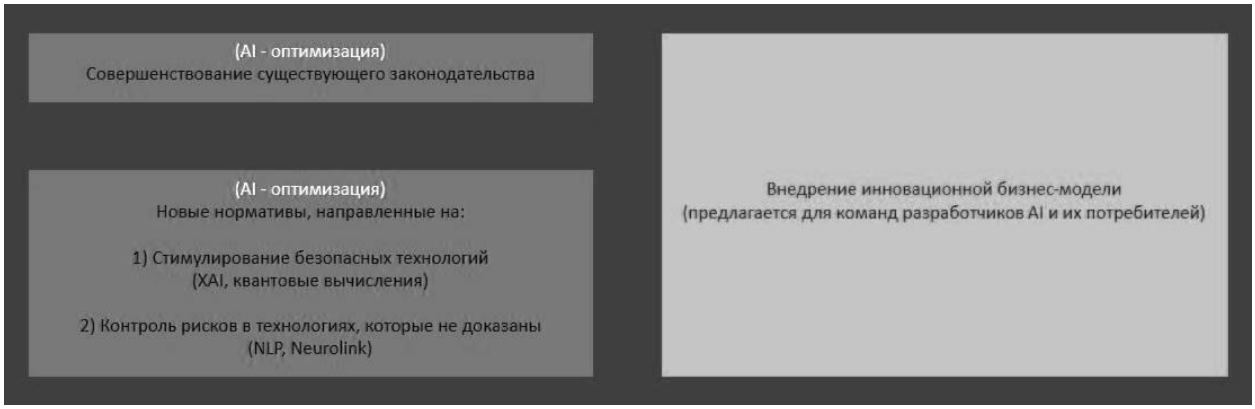

Рисунок 2. Концептуальная модель

Модель состоит из трех блоков, каждый из которых может быть реализован независимо, но в случае комплексного решения может появиться системное качество, поскольку блоки усиливают друг друга. Левая часть модели нуждается в нормативном регулировании. Здесь предлагается, в первую очередь, стимулировать безопасные технологические решения, такие как объяснимый ИИ (XAI) в различных модификациях, квантовые вычисления. Это повысит вероятность прорыва в данной области и естественный переход разработчиков на безопасные технологии. Вторая составляющая – это особый

контроль за непроверенными решениями, особенно попытками внедрения в ИИ «человеческого фактора». Уместно, если формированию нового законодательства будет предшествовать оптимизация существующего, удаление нестыковок и неоднозначностей. В этой задаче напрашивается помощь уже имеющихся и разрабатываемых решений ИИ. Правая часть модели направлена на работу с «человеческим фактором» в позитивном ключе и говорит о необходимости с приходом инноваций совершенствовать организационную структуру. Здесь уже не требуется нормативное регулирование.

Гипотеза заключается в том, что концептуальная модель решает обозначенную проблему, а именно: формирование наиболее подходящей позиции относительно разрабатываемой технологии ИИ для достижения общечеловеческих целей.

Библиографический список:

1. Величковский Б.М. Когнитивная наука: Основы психологии познания: в 2 т. Т. 1. М.: Смысл: Издательский центр «Академия», 2006. 448 с.
2. Глубокое обучение, используя байесовскую оптимизацию. ЦИТМ Экспонента [Электронный ресурс]. Режим доступа: https://docs.exponenta.ru/deeplearning/ug/deep-learning-using-bayesian-optimization.html (дата обращения: 01.04.2021).
3. Карпов В.Э., Готовцев П.М., Ройзензон Г.В. К вопросу об этике и системах искусственного интеллекта [Электронный ресурс] // Философия и общество. 2018. №2 (87). Режим доступа: https://cyberleninka.ru/article/n/k-voprosu-ob-etike-i-sistemah-iskusstvennogo-intellekta (дата обращения: 01.04.2021).
4. Колесникова Г.И. Искусственный интеллект: проблемы и перспективы [Электронный ресурс] // Видеонаука. 2018. №2 (10). Режим доступа: https://cyberleninka.ru/article/n/iskusstvennyy-intellekt-problemy-i-perspektivy (дата обращения: 01.04.2021).
5. Кучкаров З.А., Дербенцев Д.Д., Кузнецова Е.Б., Кузива Т.Д. Неопределенность понятий как источник проблем управления экономическими объектами в РФ. Пример «Устойчивости рубля» [Электронный ресурс] // УЭкС. 2017. №7 (101). Режим доступа: https://cyberleninka.ru/article/n/neopredelennost-ponyatiy-kak-istochnik-problem-upravleniya-ekonomicheskimi-obektami-v-rf-primer-ustoychivosti-rublya (дата обращения: 01.04.2021).
6. Кучкаров З.А., Дербенцев Д.Д., Лебедева В.А. Логический анализ понятия «Ликвидность» в нормативной правовой базе РФ [Электронный ресурс] // УЭкС. 2017. №7 (101). Режим доступа: https://cyberleninka.ru/article/n/logicheskiy-analiz-ponyatiya-likvidnost-v-normativnoy-pravovoy-baze-rf (дата обращения: 01.04.2021).
7. Кучкаров З.А., Шумилин Д.Е., Кузнецова Е.Б., Дербенцев Д.Д., Кузива Т.Д. Реинжиниринг полномочий органов власти в сфере экологического регулирования [Электронный ресурс] // УЭкС. 2017. №5 (99). Режим доступа: https://cyberleninka.ru/article/n/reinzhiniring-polnomochiy-organov-vlasti-v-sfere-ekologicheskogo-regulirovaniya (дата обращения: 01.04.2021).
8. Лалу Ф. Открывая организации будущего / пер. с англ. В. Кулябиной; науч. ред. Е. Голуб. М.: Манн, Иванов и Фербер, 2016. 432 с.
9. Разработка рекомендации об этических аспектах искусственного интеллекта. ЮНЕСКО [Электронный ресурс]. Режим доступа: https://ru.unesco.org/artificial-intelligence/ethics (дата обращения: 01.04.2021).
10. Ульянов С.В., Черемисина Е.Н., «Технологии интеллектуальных вычислений: Фундаментальные принципы и особенности применения в задачах системного анализа и

робастного управления» // Нечеткие Системы и Мягкие Вычисления. 2008. том 3. № 2. С. 7-13.
11. Eppe M., Oudeyer P.-Y. Intelligent behavior depends on the ecological niche // KI – Künstliche Intelligenz. 2021. V. 35. P. 103-108.
12. Gerlings J., Shollo A., Constantiou I. Reviewing the Need for Explainable Artificial Intelligence (xAI) // Proceedings of the 54th Hawaii International Conference on System Sciences. 2020. P. 1284-1293.
13. Han T.A. et al. Mediating artificial intelligence developments through negative and positive incentives // PloS one. 2021. V. 16. №. 1.
14. Pan Z., Mishra P. Hardware Acceleration of Explainable Machine Learning using Tensor Processing Units // arXiv preprint arXiv:2103.11927. 2021.
15. The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. IEEE SA [Electronic resource]. URL: https://standards.ieee.org/industry-connections/ec/autonomous-systems.html (Acces: 01.04.2021).
16. Tsimenidis S. Limitations of Deep Neural Networks: a discussion of G. Marcus' critical appraisal of deep learning // arXiv preprint arXiv:2012.15754. 2020.

Darovskikh D.W. **A challenge for learning about natural intelligence and improving ethical standards**

The view on global issues related to the development of AI technologies is presented from a different angle. He is seen as a kind of agent of positive change. By taking on a part of our thinking activity, AI helps to understand the value of another unique side of man, the so-called "human factor". This is not described by mechanisms or patterns, it is unlikely to predict, but this work will show the significance, possibility and timeliness of the practical application of the "human factor", in particular to the research and development of AI. For the smooth development of technology, it is necessary to control possible risks. As we go along, we will identify them and suggest safe workarounds. The ethical side is considered both as a tool for fixing the proposed solutions (through regulatory documents), and from the standpoint of improving morality. The result of the work is a conceptual model illustrating everything stated from a single position.

Keywords: AI Ethics, Explainable AI, Strong AI, Cognitive Computing, Quantum Computing