ИНЖЕНЕРНАЯ ПСИХОЛОГИЯ И ЭРГОНОМИКА

УДК 159.9

ГРНТИ 15.81.31

ВЗАИМОДЕЙСТВИЕ ЧЕЛОВЕКА-ОПЕРАТОРА С ИСКУССТВЕННЫМ ИНТЕЛЛЕКТОМ: ПРОБЛЕМА ДОВЕРИЯ

© 2022 г. В.М. Дозорцев*, А.Л. Венгер**

** Доктор технических наук, директор по развитию ООО «Центр цифровых технологий», Московский физико-технический институт (Национальный исследовательский университет), г. Москва, Россия*
E-mail: victor.dozortsev@mipt-cdt.ru

*** Доктор психологических наук, профессор кафедры психологии Государственного университета «Дубна»; г. Москва, Россия*
E-mail: alvenger@gmail.com

Искусственный интеллект (ИИ) стремительно проникает в промышленную автоматизацию, в частности, в составе рекомендательных систем и интеллектуальных систем принятия решений, дающих человеку-оператору советы по управлению сложными технологическими системами. Это качественно усиливает человеко-машинные системы управления, но и порождает новые вызовы обеспечения безопасности работников и производственных активов, в центре которых находится проблема доверия/недоверия человека-оператора алгоритмам, основанным на ИИ. Указанная проблема рассматривается в работе в общем контексте доверия/недоверия технике на фоне ее поступательного усложнения и интеллектуализации. На практических примерах обсуждается специфика применения алгоритмов ИИ в задачах управления сложными технологическими системами. Исследуются психологические (в том числе – индивидуально-психологические) составляющие проблемы доверия/недоверия операторов искусственному интеллекту. Предлагается и анализируется модель принятия операторских решений в потенциально опасных ситуациях. Приводится содержательная интерпретация модели в терминах субъективно допустимого уровня риска и личностной тревожности на примере задачи предиктивного анализа состояния технологического оборудования, когда игнорирование правильного совета остановить оборудование,

полученного оператором от алгоритмов ИИ, чревато катастрофическими потерями, а принятие неправильного совета – неоправданными затратами материальных ресурсов и времени. Обсуждаются меры повышения доверия и снижения недоверия алгоритмам ИИ, включая поддержание постоянной активности оператора в системе управления «человек-оператор – ИИ», учет мотивационной составляющей принятия операторских решений в присутствии советов ИИ, обеспечение прозрачности и интерпретируемости советов, разработку обоснованных регламентов взаимодействия оператора и ИИ в рамках интеллектуальных систем управления сложным технологическим оборудованием.

Ключевые слова: человек-оператор, доверие/недоверие искусственному интеллекту, рекомендательные системы, интеллектуальные системы принятия решений, субъективно допустимый уровень риска, предиктивный анализ состояния оборудования

ВВЕДЕНИЕ

В контексте современной промышленной автоматизации, по мере появления соответствующих научных методов и технологических инструментов, искусственный интеллект (ИИ) все шире используется для решения задач управления производственными объектами. Пока это происходит в форме предлагаемой человеку-оператору рекомендации или совета, который может быть принят или отвергнут человеком-оператором в соответствии с принятым регламентом работы. Несмотря на ведущую роль человека-оператора в такой связке (а, возможно, и вследствие этой роли), проникновение ИИ в системы управления порождает серьезные вызовы, связанные с безопасностью работников и производственных объектов, мотивацией персонала, этикой производственных отношений и пр. Ключевой аспект этих проблем – доверие человека-оператора искусственному интеллекту. Представляется, что достижение достаточного уровня такого доверия является необходимым условием безопасного и эффективного функционирования современного высокоавтоматизированного производства. Тем острее становится необходимость психологического обеспечения человеко-машинного взаимодействия при применении ИИ.

Цель работы – рассмотреть заявленную проблему в контексте доверия человека технике вообще на фоне поступательной интеллектуализации технических систем.

ИССЛЕДОВАТЕЛЬСКИЙ ЛАНДШАФТ ПРОБЛЕМЫ ДОВЕРИЯ СЛОЖНОЙ ТЕХНИКЕ

Доверие человека ИИ нельзя рассматривать в отрыве от его доверия технике вообще – темы, по-видимому, столь же старой, как и сама техника. Еще в пору начальной механизации (рубеж XVIII-XIX вв.) луддиты уничтожали ткацкие станки, не без оснований подозревая, что техника отнимет работу у ткачей, наиболее квалифицированных на тот момент работников (Jones, 1959). Жестоко подавленные экономические бунты позже получили продолжение в протестах против вытеснения человека из производства и неприятии всего технического, ярко представленных в литературе романтизма (М. Шелли, Э.-Т. Гофман). По мере развития техники проблема доверия становилась все более многоаспектной, вслед за литераторами в нее вовлекались философы, социологи, культурологи, психологи, фантасты. На каждом витке усложнения техники тема доверия затрагивала все новые категории пользователей – операторы промышленных систем, наделенные соответствующими профессиональными полномочиями и ответственностью; эксплуатанты, использующие технику вне профессиональной деятельности; владельцы, санкционирующие использование техники; клиенты, пассажиры, «юзеры» и др. Начиная с определенного момента, человек-оператор на производстве работал уже не с техникой как таковой, а с автоматикой, ею управляющей. Причем во все большей степени человек-оператор «видел» не сам объект управления, а его представление в подсистеме визуализации системы управления. Это справедливо не только для технологических объектов, но и для крупных подвижных объектов (лайнеров, танкеров), а также отчасти для пассажирского транспорта, где за пилотом или машинистом все же остается непосредственное наблюдение за внешней обстановкой. Далее рассматриваются факторы, определяющих доверие человека-оператора именно в автоматизированной технике.

Проблема доверия технике – традиционная для психологических исследований (Lee, See, 2004). Принципиальное положение – уровень доверия должен

соответствовать возможностям техники. Если такого соответствия нет, могут возникать сверхдоверие или сверхнедоверие, чреватые либо снижением безопасности, либо неоправданным отказом от преимуществ современных систем автоматизации. При этом. помимо технических параметров самой автоматики (надежность, безопасность, удобство использования, история применения), на уровень доверия влияют характеристики человека-оператора (личный опыт, профессионализм, уровень самооценки, другие личностные характеристики). Подчеркнем, что в отечественной психологии доверие или недоверие технике понимаются как осознанные психологические отношения, описываемые когнитивными, эмоциональными и поведенческими характеристиками (Акимова, Обознов, 2016).

Важно отметить, что пара «доверие-недоверие» в отношении оператора к технике несимметрична (Lewicki et al., 1998; Купрейченко, 2008; Акимова, Обознов, 2017; Акимова, 2020). В исследовании А.Ю. Акимовой и А.А. Обознова выделены и экспериментально обоснованы специфические факторы, несимметрично влияющие на уровень доверия и недоверия: среди них – оправдание ожиданий правильных действий техники в критических ситуациях (повышение доверия) или преодоление собственных негативных переживаний (снижение недоверия). Определены общие факторы, одновременно повышающие доверие и снижающие недоверие или наоборот. Пример таких факторов – поддержание хорошего технического состояния или оценка надежности техники. Воздействуя на отдельные факторы, можно разнообразными способами влиять на уровень доверия/недоверия технике. Показано, что высокий уровень доверия снижает для оператора когнитивную сложность ситуации, оптимизирует ресурсы внимания, предотвращает психосоматические заболевания, а низкий – наоборот, увеличивает эмоциональную напряженность работы и способствует психосоматическим расстройствам. Таким образом, доверие технике представляет собой важный операторский ресурс. (Акимова, Обознов, 2017; Акимова, 2020).

Сложную автоматику рассматривают в контексте т.н. агентных систем, в первую очередь, интеллектуальных и когнитивных агентов, причем доверие к таким агентам предполагает их восприятие как обладающих намеренностью и понимается как отношение человека-оператора к тому, поможет ли ему агент в ситуации неопределенности и уязвимости (Lee, See, 2004). Именно для сложной техники, обладающей свойствами целеустремленности, саморегуляции, непредсказуемости, ранее относимыми исключительно к людям (Голиков, 2003), направленности на диалог на естественных языках и даже многовариантности и оптимальности, ярко проявляется смещение феномена доверия/недоверия от отношения «субъект-объект» к отношению «субъект-субъект» (Акимова, Обознов, 2017).

В этом контексте интересна тема семантизации техники (Хорошилов, 2009). Символическое представление техники разворачивается по нарастающей в цепочке «витализация-анимация-антропоморфизация-субъективизация-персонализация»; оно повышает мотивацию человека-оператора, дополняет и усиливает абстрактное концептуальное восприятие им технической системы. Указанная тенденция подпитывается все большей похожестью техники на живой организм (таков базовый принцип бионики; теми же чертами обладают управляемые технологические процессы, описываемые как открытые системы). Представление машины как партнера/соперника, друга/врага мобилизует продуктивные мотивы человеко-машинного взаимодействия, формирует когнитивные и поведенческие установки в отношении объекта, дает интегрированную когнитивную мета-схему объекта как сложной, динамической, самореализующейся активной, целостной системы, как некоего существа (Хорошилов, 2009). Этот механизм особенно важен в потенциально опасных ситуациях управления техникой, но он же может привести к излишнему доверию, избыточным затратам на коммуникацию, переоценке возможностей автоматики. Семантизация помогает и в решении коммуникационных (поведенческих) проблем. Сложную технику, все больше представляемую современными иммерсивными интерфейсами на базе виртуальной и

дополненной реальности, «не погладишь и не пнешь» (Акимова, Обознов, 2016). нужно разрешать ситуацию средствами субъективизации.

СПЕЦИФИКА ДОВЕРИЯ/НЕДОВЕРИЯ В КОНТЕКСТЕ ИИ

Искусственный интеллект радикально меняет современную экономику. McKinsey Global Institute прогнозирует, что к 2030 г. ИИ породит новую экономическую активность на 1 триллион долларов в год, что увеличит совокупный мировой ВВП на 16%. Уже сейчас более половины опрошенных работодателей готовы применить ИИ вместо приема на работу нового сотрудника. Примерно столько же работодателей ожидают, что на горизонте пяти лет отдельные активы большой ценности будут автономно управляться ИИ. Собственно в промышленности 11% компаний уже используют ИИ, почти половина компаний приступит к этому в течение нескольких следующих лет (Мельникова, 2021).

Что же ожидает персонал промышленных предприятий с приходом искусственного интеллекта? Ясности у практиков нет, да и в исследовательской среде не утихают дискуссии о том, в какой форме ИИ проникнет в промышленные системы управления и к чему это приведёт[1]. На сегодня принято следующее прагматическое определение: ИИ – это комплекс технологических решений, позволяющий имитировать когнитивные функции человека (включая самообучение и поиск решений без заранее заданного алгоритма) и получать при выполнении конкретных задач результаты, сопоставимые, как минимум, с результатами интеллектуальной деятельности человека (см. Национальную стратегию развития искусственного интеллекта на период до 2030 г. (Развитие…,

[1] Исторически в ИИ выделяют *восходящую* парадигму (Минский и Пейперт, 1971), т.е. от моделирования базовых биологических процессов к решению интеллектуальных «человеческих» задач (искусственные нейронные сети, эволюционные алгоритмы, пр.) и *нисходящую* парадигму (McCarthy, 1998), т.е. имитацию работы естественного интеллекта на основе символьных и семиотических систем (экспертные системы, обработка знаний и логический вывод, агентные системы). Представляется, что в промышленности будет востребован гибридный подход: оптимизация, поиск скрытых закономерностей – от восходящей парадигмы, возможность объяснить принятые решения – от нисходящей. Последнее чрезвычайно важно в контексте доверия ИИ.

2022). Отдельные компоненты ИИ и связанные с ними технологии находятся на разных уровнях зрелости. Человеку-оператору автоматизированных производственных систем уже приходится иметь дело с рекомендательными системами и интеллектуальными системами поддержки принятия решений. Эти компоненты, составляющие особый кластер в направлении ИИ, на сегодня достаточно развиты и по шкале зрелости технологий они находятся между этапом II (расцвет технологии) и этапом III (зрелость технологии)[2]. Высока публикационная и патентная активность по этому компоненту: число соответствующих научных публикаций в мире составило 593 (в 2010 г.), 956 (2015 г.) и 3128 (в 2020 г.). В том же темпе росло число патентных заявок (от 993 в 2010-м, до 1814 в 2020 г.)[3]. Очевидно, настало время обеспечить условия (в т.ч. – психологические) практической реализации этой активности. Важнейшее из этих условий – доверие оператора ИИ.

Приведем несколько реальных примеров использования ИИ только в металлургической отрасли:

– при лазерном спекании металлических изделий (одна из наиболее эффективных аддитивных технологий) возможные локальные перегревы приводят к выпуску брака. Обученная на экспериментальных данных система ИИ определяет точные локальные тепловые профили, прогнозирует области перегрева в заготовке и позволяет с первого раза правильно напечатать изделие;

– похожая модель распознает износ конвейерной ленты, определяет причины ускоренного износа (например, колебания и биения из-за неправильной настройки и эксплуатации конвейера), позволяет осуществлять предиктивное обслуживание и ремонт ленты;

– важнейшая проблема отрасли состоит в своевременной и надежной диагностике состояния футерованного оборудования (литейные ковши, транспортные тележки, др.),

[2] Этап I на шкале зрелости соответствует зарождению технологии, а этап IV – плато.

[3] РФ слабо представлена в этом движении: ее публикационная доля выросла за рассматриваемое десятилетие с 0.3 до 0.8%, а число патентов с нуля до трех заявок в год.

поскольку в результате длительного воздействия механических и температурных перегрузок возрастает риск разрушения этого оборудования, что чревато многомиллионными ущербами и угрозой здоровью и жизни персонала. Оператор в таких ситуациях часто действуют по принципу "лучше перестраховаться, чем потом сожалеть". Соответствующие алгоритмы ИИ, используя технологию глубоких нейронных сетей, позволяют обнаруживать зоны выгорания футеровки и дают надежную рекомендацию по ее замене;

– методам ИИ нет альтернативы в задачах обслуживания оборудования при высокой стоимости поломки (например, оборудования для сварки стальных труб или турбокомпрессорные агрегаты для подачи воздуха в доменные печи). Использование ИИ-технологий позволяет не только заранее выявить критические отказы и тем самым обезопасить дорогостоящее оборудование, но и изменить принцип профилактического технического обслуживания, достигая существенного снижения электропотребления и продолжительности простоев. Только ИИ делает возможным переход от планово-предупредительного ремонта к ремонту по фактическому состоянию оборудования, что принципиально меняет экономику производства.

Во всех приведенных примерах решение ИИ предоставляется в режиме совета, который человек-оператор вправе принять или отвергнуть. В некоторых случаях (замена футеровки или ленты конвейера) у него достаточно времени для анализа предложенного совета (часы, дни); в других (поломки турбины или компрессора) сигнал может поступить за четверть часа до возможного отказа, оставляя возможность спасти критически важное оборудование и предотвратить длительные, иногда многомесячные и дорогостоящие простои, а также требуя от человека-оператора очень быстрого ответственного решения.

ФАКТОРЫ, ВЛИЯЮЩИЕ НА ДОВЕРИЕ/НЕДОВЕРИЕ ЧЕЛОВЕКА-ОПЕРАТОРА ИИ

Выделим особенности ИИ, влияющие на доверие/недоверие операторов в ситуации выбора.

Мотивационная ловушка. Если отклонить правильный или принять неправильный совет ИИ, человек-оператор рискует продемонстрировать неполное профессиональное соответствие. Если принять правильный совет, то ИИ окажется, как минимум, расторопнее, а то и «умнее» человека-оператора. Только отклонение неверного совета – в плюс человеку-оператору. Необходимо четкое разделение штрафов и выгод от использования ИИ. Во всех применениях, в том числе промышленных, должен быть определен ответственный за негативные последствия – никто (форс-мажор), прямой виновник (производитель, разработчик, владелец, пользователь). То же с бенефициарами ИИ, которыми могут выступать производитель, разработчик, покупатель, пользователь, сам ИИ (если когда-нибудь будет наделен субъектностью). С учетом формальной и часто неформальной ответственности человека-оператора и его статуса наемного работника, такая «несимметричная» мотивация вряд ли повышает уровень доверия человека-оператора ИИ. Эта тема может рассматриваться в контексте этики ИИ, предполагающей помимо справедливых наказаний и поощрений, проблему предвзятости ИИ, запрета на его использование в определённых задачах и предотвращения индивидуальных и общественных предубеждений (Этичное применение…, 2020).

Прозрачность ИИ. Человеку-оператору необходимо обоснование решения, выработанного ИИ. Лозунг «компьютер сказал нет» теперь не проходит – алгоритмы ИИ должны быть транспарентны, как на уровне специалистов, анализирующих работу ИИ, так и для собственно операторов, в том числе – в режиме реального времени, насколько это возможно по ходу производственной ситуации. Совет ИИ не должен выглядеть как выход черного ящика (например, искусственной нейросети), а должен

быть объяснен/обоснован или хотя бы прокомментирован пользователю на понятном ему языке. Разумеется, прозрачность не должна противоречить безопасности (сошлемся на состязательные атаки на системы распознавания лиц, саботаж и диверсии на критических объектах производства и инфраструктуры). Институт инженеров электротехники и электроники (IEEE) работает над стандартом прозрачности ИИ. Прозрачность и необходимость комментировать решения ИИ также можно рассматривать как часть проблемы этичного применения ИИ (Этичное применение…, 2020).

Неготовность использовать ИИ. Тема неготовности к использованию новой техники также из разряда вечных. Один из трагических примеров — нападение японской авиации на американскую военно-морскую базу Перл-Харбор 7 декабря 1941 г. Нечеткие сигналы на мониторах радаров, весьма несовершенных на тот момент, были восприняты как очередные помехи от птичьих стай и проигнорированы операторами. Внезапность нападения не была компенсирована вследствие неготовности операторов и низкого уровня их доверия новой технике. В противоположность этому оправданное недоверие технике спасло СССР и США от ядерного столкновения в разгар холодной войны. Так, за первое полугодие 1980 г. только на американской стороне зарегистрировано 2159 случаев ложного срабатывания автоматики (больше 10 за сутки, 69 из них были расценены как экстренны). Все они были результатом неисправностей или сбоев аппаратных средств и программного обеспечения, а также естественных помех (световые блики и температурные перепады) (*Hart, Goldwater*, 1980).

Очевидный инструмент повышения готовности персонала использовать новую технику — обучение и перепрофилирование пользователей и специалистов, однако в случае ИИ ситуация усугубляется рядом дополнительных факторов:

– инструменты ИИ все еще достаточно «молоды», не застрахованы от «детских» болезней, при наступлении неучтенных на стадии обучения ситуаций могут повести себя непредсказуемо. В сравнении с решениями традиционной автоматики

операторам объективно сложнее понять логику ИИ и отделить разумные, но необычные решения от ошибочных;

– опасения потерять работу обострялись в каждой критической точке развития средств автоматизации: при переходе от аналогового к цифровому компьютеризованному управлению, при внедрении первых «производственных автопилотов» и, наконец, при проникновении в автоматику инструментов ИИ. Как правило, эти опасения не оправдывались или сильно завышались, хотя резкие изменения характера труда и профессиональных требований к оператору всегда были необходимы. Согласно прогнозам, в ближайшее время сокращения персонала при внедрении ИИ ожидались на уровне 75 млн человек по всему миру при создании 133 млн новых позиций. Но следующие волны сокращений могут затронуть до 30% всех работников. Наблюдаемый сейчас дефицит кадров, готовых работать с ИИ, сменится переизбытком. Необходимы проактивные публичные действия сообщества разработчиков и пользователей ИИ, чтобы показать социальные преимущества новых технологий;

– в отечественной практике объективные опасения операторов по поводу сложной автоматики (неверие в возможности компьютеров и передовых технологий вообще, неуверенность в своей способности овладеть новыми средствами управления, страх потерять работу) часто маскировались «правильными» лозунгами об экономии народных средств и «оторванности науки от практики». В случае ИИ эти причины недоверия дополняются признаками ущемленной гордости хомо-сапиенс. Теперь автоматика претендует не только на недоступные человеку-оператору быстрые рутинные действия, но и на оптимизацию, планирование, предиктивный анализ, то есть все больше посягает на «святая святых», доступное, по мнению большинства людей, только естественному человеческому интеллекту. Здесь работает та же семантизация техники, только уже не как дополнительный ресурс человека-оператора, а как механизм отчуждения человека-оператора и техники. ИИ воспринимается как чужой, в

определенной степени даже враг[4]. Представляется, что с приходом сильного ИИ эту проблему нельзя будет игнорировать.

КАК ПОВЫСИТЬ ДОВЕРИЕ ИСКУССТВЕННОМУ ИНТЕЛЛЕКТУ

По мере повышения надежности алгоритмы ИИ будут брать все больше инициативы, а человек-оператор сможет возвращать управление себе, только когда сам сочтет нужным или когда ИИ запросит об этом. Эта тенденция прослеживается на всех этапах автоматизации и сейчас многие операции безоговорочно остались за автоматикой. Так, в управлении технологическими процессами к ней отошло не только очень «быстрое» базовое регулирование, но и более медленное программное управление и даже элементы многосвязного оптимального управления.

На заре автоматизации производства Б.Ф. Ломов сформулировал принцип «активного оператора», согласно которому человек-оператор не должен исключаться из цепи управления, сохраняя в ней активные функции (Ломов, 1966). В новое время человек-оператор — активный элемент человеко-машинной системы, даже в автоматическом режиме управления он всегда должен иметь в виду конечную цель своих взаимодействий с машиной. В этом случае его переход к активным действиям пройдет с меньшими затратами когнитивных и психологических ресурсов, эффективность его деятельности будет выше, а психофизиологические затраты меньше[5]. Развитием этой идеи стал разработанный А.Н. Костиным принцип оптимального взаимного резервирования оператора и автоматики (Костин, 2021). Согласно этому подходу, критерии перевода управления на себя у операторов преимущественно качественные, а у

[4] Согласно опросам, люди, теряющие работу, существенно чаще хотят, чтоб она досталась другим людям, а не «роботам». В то же время известен феномен т.н. «уклона автоматизации», когда в некоторых задачах люди изначально склонны больше доверять автоматике, чем другим людям. например, при проверке багажа в аэропортах (Dzindolet et al., 2002).

[5] Обратим внимание на схожесть описанного подхода с результатами теории активных систем, полученными в пору появления первых советчиков оператора. В работе (Большие системы…, 1989) решения оператора поддерживаются премиями за принятие верного и отклонение неверного совета; штрафами — за игнорирование верного и принятие неверного совета. Там же обосновывается алгоритм, оптимально мотивирующий оператора на эффективное взаимодействие с автоматикой.

автоматики (точнее, у ее разработчиков) – количественные. Резервирование человека-оператора автоматикой обуславливается превышением предельно допустимой субъективной сложности операторской деятельности, когда для выполнения профессиональных функций от оператора требуется слишком высокий уровень психической регуляции. Указанный переход может осуществляться, например, по измерениям межсаккадических интервалов методом электроокулографии. Обратное резервирование (от автоматики к оператору) происходит при наступлении отказов или по инициативе оператора. Помимо прочего это гармонизирует отношения операторов (изначально склонных не доверять автоматике и брать как можно больше управления на себя) и разработчиков, опасающихся человеческих ошибок и стремящихся все автоматизировать.

Взаимное резервирование хорошо себя зарекомендовало в управлении подвижными объектами (особенно в авиации и космонавтике); перенесение этого успеха на другие операторские профессии выглядит не столь очевидно. К тому же в контексте доверия автоматике взаимное резервирование работает, скорее, на пару «оператор-разработчик» и только затем – опосредованно и частично – на пару «оператор-автоматика». Представляется, однако, что проблема значительно шире; особенно в плане влияния на доверие/недоверие личностных особенностей операторов. Остановимся на некоторых из них подробнее.

ИНДИВИДУАЛЬНО-ПСИХОЛОГИЧЕСКИЕ ФАКТОРЫ ДОВЕРИЯ ИИ

Уровень доверия искусственному интеллекту и итоговый результат – принятие или отклонение его рекомендаций – может в значительной мере зависеть как от прежнего опыта, так и от личностных особенностей человека- оператора. В частности, можно полагать, что большую роль играет степень успешности его предшествующей деятельности с использованием не только ИИ, но и других высокотехнологичных систем. Высокая успешность должна повышать доверие ко всем таким системам и, в частности, к

ИИ; напротив, многочисленные неудачи, отказы сложной техники могут существенно снизить доверие ИИ. Это предположение доступно экспериментальной проверке на модели краткосрочного изменения уровня доверия. В таком эксперименте, предложенном нами ранее (Дозорцев, Венгер, 2022), следует предпослать выявлению уровня доверия ИИ выполнение человеком-оператором нескольких заданий с использованием достаточно сложных компьютерных программ. Далее будет сопоставляться уровень доверия ИИ при бесперебойной работе этих программ и при многочисленных сбоях в их работе (разумеется, испытуемый не будет знать, что эти сбои заранее предусмотрены экспериментатором). Если во втором случае выявится более низкий уровень доверия ИИ, это послужит обоснованием гипотезы о том, что регулярный негативный опыт такого рода может вызывать не только кратковременное, но и устойчивое недоверие ИИ.

Представляется весьма вероятной связь уровня доверия ИИ с характерным для данного оператора локусом контроля. Согласно Дж. Роттеру, внешний локус контроля характеризуется атрибуцией своих успехов и неудач внешним факторам, к каковым относятся, в частности, рекомендации ИИ. При внутреннем локусе контроля субъект полагается, в первую очередь, на свои собственные представления (Rotter, 1990). Можно ожидать, что человек- оператор с внешним локусом контроля в случае расхождения его представлений с рекомендациями ИИ охотнее примет эти рекомендации, чем при внутреннем локусе контроля. Дополнительные варианты атрибуции выделил М. Селигман. В его концепции рассматриваются оптимистический и пессимистический атрибутивные стили (Селигман, 2017). При первом из них субъект приписывает успехи себе, а неудачи – внешним факторам, при втором – наоборот. Можно ожидать, что оператор с пессимистическим атрибутивном стилем с большей вероятностью последует рекомендациям ИИ, чем оператор с оптимистическим стилем.

Степень доверия ИИ будет зависеть также от прогрессистских или, напротив, консервативных идеологических установок субъекта: ориентация на прогресс будет повышать доверие, консервативная ориентация – понижать. Так, Н. Постман (N.

Postman) выступает с консервативных позиций против любых высоких технологий, одним из высших достижений которых является ИИ. Он утверждает, что высокие технологии приводят к отчуждению их создателей («элиты») от рядовых пользователей, порождая у последних ощущение своего бессилия (Postman, 1993).

Поскольку внедрение ИИ осуществляется представителями руководства, а не рядовыми операторами, степень доверия ему будет определяться также отношением человека-оператора ко всем «вышестоящим» инстанциям – непосредственному начальству, социальной элите, науке и т.п. Особенно резко доверие ИИ будет снижаться при наличии конспирологических представлений о тайных группах («мировом правительстве»), действующих против интересов рядовых граждан. На доверии ИИ будут негативно сказываться также предположения пользователя о том, что при создании и обучении алгоритмов не была учтена возможность каких-либо необычных, но представляющихся ему вполне вероятными ситуаций.

Имеющийся у субъекта уровень доверия сложной автоматике влияет на его индивидуальное отношение к ситуациям, связанным с риском неприемлемого ущерба (например, катастрофической аварии или собственной гибели). В экономических моделях такие ситуации соответствуют риску полного разорения (Фалин, 1994). В модели, разработанной Дж. Нейманом и О. Моргенштерном (Нейман и Моргенштерн, 1970), решение принимается, исходя из значения функции полезности $U = \rho \cdot u^{+} - (1-\rho) \cdot u^{-}$, где ρ – вероятность возможного «выигрыша» при положительном решении, u^{+} – его величина; $(1-\rho)$ – вероятность возможного «проигрыша» при том же решении, u^{-} – его величина. Но если вероятность неприемлемого ущерба превышает некоторую заранее заданную константу ε, функция полезности неприменима. Сколь бы высок ни был возможный «выигрыш», в этом случае положительное решение не может быть принято.

В предложенной нами модели максимально допустимая вероятность неприемлемого ущерба ε рассматривается как субъективно допустимый уровень риска (Венгер, 2013, 2018). При высоких значениях этого параметра человек готов

участвовать в ситуациях с относительно высокой вероятностью неприемлемого ущерба – например, может присоединиться в заведомо опасной экспедиции. Эту стратегию можно охарактеризовать как *стратегию риска*. Напротив, низкие значения ε соответствуют *стратегии избегания риска*, установке на безопасность (разумеется, при такой установке становятся невозможными и особо крупные «выигрыши»).

В реальной жизни никогда не известна точная вероятность катастрофического исхода. Имеется лишь ее субъективная оценка (субъективная вероятность). Точность этой оценки характеризуется величиной ошибки, причем сама эта величина также представляет собой субъективную оценку. Пусть некоторое решение d с положительной функцией полезности может привести к катастрофе с субъективной вероятностью q. Тогда это решение может быть принято лишь в том случае, если не только $q < \varepsilon$, но и субъективная оценка ошибки не слишком высока, т. е. неравенство $q < \varepsilon$ должно выполняться с достаточно высокой надежностью. Надежность оценки тем выше, чем больше информации о возможных последствиях решения d удалось получить. Степень требуемой надежности η – еще один индивидуальный параметр, характеризующий индивидуальную стратегию принятия решений.

Рассмотрим пример предиктивной аналитики оборудования, когда ИИ подает сигнал опасности, требующий остановки сложного производственного процесса. Предположим, что цена такой остановки очень высока, но авария, о возможности которой предупреждает ИИ, ведет к неприемлемому ущербу (скажем, это катастрофическая авария на АЭС). Оператор может отклонить совет ИИ и принять соответствующий риск (решение d^r) или принять совет и остановить процесс (решение d), но в последнем случае, если тревога ложная, есть риск значительных потерь из-за напрасной остановки.

У человека-оператора имеется некоторое исходное представление о вероятности этой катастрофы – априорная субъективная вероятность q_0. При этом возможны две ситуации. Если оператор к рассматриваемому моменту не заметил никаких тревожных

признаков в работе оборудования (совет ИИ оказался для него сюрпризом), априорная субъективная вероятность $q_0 << \varepsilon$. Если же такие признаки появились, но человек-оператор не принял самостоятельного решения остановить процесс, то можно допустить, что субъективная априорная вероятность аварии повысилась, но все же не превысила допустимую: $q_0 < \varepsilon$. После получения от ИИ сигнала опасности получаем апостериорную вероятность $q=(q_0|S,\zeta)$, т.е. субъективную вероятность аварии при условии, что от ИИ получен сигнал S. Значение q зависит от множества факторов: от априорной вероятности q_0 и от набора условий, обозначенных параметром ζ, среди которых степень доверия ИИ, определяемая всеми описанными выше соображениями; вероятность (частота) ложных аварийных сигналов («ложных тревог»), подаваемых ИИ, пр. Заметим, что при сверхвысокой цене аварии неизбежно будет относительно большое количество «ложных тревог».

В зависимости от соотношений между значениями описанных параметров возможны следующие решения:

1. Если $q \geq \varepsilon$, то оператор остановит процесс (решение d'): апостериорная субъективная вероятность аварии превышает субъективно допустимую вероятность неприемлемого ущерба.

2. Если $q < \varepsilon$, то решение будет зависеть от того, достаточно ли высока надежность оценки σ, т.е. достаточно ли надежно выполнено это соотношение. А именно:

2.1. в случае достаточно высокой надежности ($\sigma > \eta$) сигнал опасности будет проигнорирован (решение d'');

2.2. при $q < \varepsilon$; $\sigma \leq \eta$ в свою очередь возможны два разных решения:

2.2a d' – остановка оборудования в соответствии с рекомендацией ИИ;

2.2b d' – попытка собрать дополнительную информацию для повышения надежности оценки истинного значения вероятности q.

Вариант 2.2b возможен лишь при наличии доступных источников такой информации и достаточного времени для ее получения.

Разные варианты решения будут приводить к разным эмоциональным состояниям. Так, вариант 1 будет сопровождаться уверенностью в правильности своего решения, но в сочетании с беспокойством, тревогой вызванными высокой ценой остановки процесса (поскольку оператор осознает высокую вероятность «ложной тревоги»).

Вариант 2.1 также будет сопровождаться чувством уверенности, причем без каких-либо опасений: оператору достаточно информации для принятия однозначного решения, предотвращающего потери, связанные с остановкой процесса. Однако объективно при этом варианте повышен риск допустить аварию.

Вариант 2.2a будет сопровождаться неуверенностью в правильности решения, особо выраженной тревогой как по поводу высокой цены остановки процесса, так и по поводу возможной ошибочности своего решения.

Вариант 2.2b при наличии источников информации и достаточного времени для ее получения будет в меньшей степени сопровождаться тревогой, но он в любом случае является промежуточным и в конченом итоге, в зависимости от полученных данных, сведется к одному из предыдущих.

Неравенство $q \geq \varepsilon$ достигается тем легче, чем ниже значение ε, т.е. чем более оператор осторожен, склонен к избеганию риска (подчеркнем, что риск в рассматриваемом случае состоит в наступлении аварии из-за отклонения оператором совета ИИ.) Неравенство $\sigma \leq \eta$ достигается тем легче, чем выше значение η, т.е. чем более надежная дополнительная информация необходима оператору для отклонения совета ИИ.

Таким образом, оба постулированных в модели параметра – и ε, и η – характеризуют личностную тревожность субъекта. Выраженность тревожности повышается с падением параметра ε и с ростом параметра η. Однако психологическая

природа тревожности в этих случаях различна. Низкие значения ε приводят к беспокойству по поводу внешних обстоятельств, тревога будет прямо пропорциональна цене неоправданной остановки процесса, т.е. следования ошибочной рекомендации ИИ. Высокие значения η будут приводить к неуверенности в себе, почти независимо от этой цены. Для человека с потребностью в высоко надежной информации будет особенно затруднено принятие решений в условиях ограниченности времени.

Различные сочетания этих двух параметров порождают четыре крайних варианта стратегии (табл.).

Таблица

Стратегии принятия решений при риске неприемлемого ущерба

Значение		**Стратегия**
ε	**η**	
Высокое	Низкое	Экстремальный риск
	Высокое	Разумный риск
Низкое	Низкое	Избирательное избегание риска
	Высокое	Стабильное избегание риска

Стратегия экстремального риска обеспечивает высокую эффективность в непредсказуемых ситуациях, требующих немедленных действий, при выраженном дефиците времени – особенно, в условиях уже развивающейся аварии[6]. Соответствует наиболее низкому уровню личностной тревожности.

Стратегия разумного риска наиболее выгодна в непредсказуемых ситуациях, но с достаточным запасом времени на принятие решения и при наличии дополнительных источников информации.

Стратегия избирательного избегания риска уместна в стабильных, достаточно предсказуемых ситуациях, но при выраженном дефиците времени и/или отсутствии дополнительных источников информации.

[6] Разумеется, нет гарантии, что действия, предпринятые в таких условиях, будут оптимальными. Однако речь идет о ситуации, в которой нерешительность, отсутствие действий заведомо пагубны.

Стратегия стабильного избегания риска наиболее благоприятна в стабильных, предсказуемых ситуациях, при достаточном запасе времени на принятие решения и наличии дополнительных источников информации. Соответствует наиболее высокому уровню личностной тревожности.

Отметим, что в рассмотренной задаче предиктивной аналитики состояния оборудования имеется и другой выбор, когда оператор склоняется к необходимости остановить процесс (субъективная вероятность аварии приближается к порогу ε), а ИИ не подает сигнала к отключению. Ситуация как бы переворачивается: теперь основной риск – поверить ошибочному молчанию ИИ и тем самым допустить аварию. При этом ИИ либо не видит опасности (его модель не точна), либо видит какие-то аномальные признаки, но не считает их критическими. Последний случай приводит к важному выводу: ИИ должен не только рекомендовать останов, но и информировать об аномальном поведении оборудования, даже если оно не «дотягивает» до аварии. (Кстати, такое сообщение также можно интерпретировать как совет ИИ.) Представляется, что описанные выше стратегии принятия решений применимы и в этом случае.

ОБОСНОВАННЫЕ РЕГЛАМЕНТЫ ВЗАИМОДЕЙСТВИЯ ЧЕЛОВЕКА-ОПЕРАТОРА И ИИ

Резюмируем основные требования к регламентированию взаимодействия человека-оператора и ИИ в ситуации выбора «принять или отклонить совет», позволяющие смягчить проанализированные выше опасения человека-оператора и повысить уровень его доверия автоматике.

– человек-оператор должен понимать, что совет ИИ принципиально неидеален. В ходе подготовки/ обучения оператору должны быть даны представления о механизмах выработки и границах точности советов ИИ, а также информация о последствиях их принятия или отклонения;

– даже если совет правилен, человеку-оператору может недоставать информации для его принятия, а форма совета может затруднять выбор. Необходимо обеспечить

прозрачность ИИ: обоснование совета, раскрытие (ограниченное рамками безопасности производства и защиты ноу-хау) информации о данных, на основе которых сформирован совет;

– должна быть гарантирована объективность исходных данных, используемых ИИ, следует минимизировать негативные факторы, приводящие к любой дискриминации человека-оператора и обеспечить его контроль за работой системы ИИ;

– необходим реально действующий механизм рассмотрения претензий человека-оператора, включая его обязательное к исполнению право на справедливое разбирательство возможных инцидентов при использовании ИИ, учет коллективных интересов (работников, производственного подразделения, компании в целом, общества);

– следует предусмотреть процедуру расследования случаев существенного влияния на выбор человека-оператора внешних факторов, в т.ч. несовершенного интерфейса, кибератак, информационной перегрузки, давления административных и социальных факторов и пр.;

– там, где это обосновано, должно быть предусмотрено возмещение материальных и моральных убытков, понесенных человеком-оператором;

– с учетом постоянных изменений в столь прорывном направлении промышленной автоматизации, как ИИ, и накапливаемого на предприятии опыта его использования, необходимы регулярные курсы переобучения персонала (собственно операторов, обслуживающего и сопровождающего персонала).

Разумеется, указанные регламенты взаимодействия должны быть предельно четкими и однозначными, их необходимо регулярно обновлять по мере изменений на производстве и в алгоритмах ИИ.

ЗАДАЧИ БУДУЩИХ ИССЛЕДОВАНИЙ

Искусственному интеллекту нет альтернативы во все большем числе задач, включая управление сложной техникой и промышленную автоматизацию. Имеет место опережающее внедрение инструментов ИИ, затрагивающих безопасность людей, производственных активов и инфраструктуры (беспилотные средства передвижения, защита критически важного оборудования и пр.). Надо готовиться к взаимодействию со все более мощным ИИ, хотя пока критический уровень «сильного искусственного интеллекта» в прикладных задачах далеко не достигнут.

Высокое доверие человека-оператора искусственному интеллекту – бесспорная предпосылка эффективного производства. Уровень доверия системам ИИ определяется целым рядом объективных факторов, традиционно находящихся в поле интересов широкого круга исследователей, включая, не в последнюю очередь инженерных психологов. Значительное влияние на уровень доверия оказывают также субъективные индивидуально-психологические характеристики операторов. Предпринятый в настоящей работе анализ некоторых из них должен быть подкреплен в психологическом эксперименте, что потребует разработки специальных опросников и критериев оценки уровня доверия/недоверия.

Несомненно важно учитывать эргономические факторы взаимодействия системы «человек-оператор – ИИ». Чем больше у человека-оператора уверенности в намеренности когнитивного агента, тем больший ресурс внимания ему необходим при коммуникации. Изменения в человеко-машинной коммуникации, вызванные использованием систем ИИ, могут также повлиять на уровень доверия человека-оператора ИИ. Известно, что люди склонны больше доверять когнитивным агентам, подобным машинам, а не себе подобным[7]. И наоборот, в случае подозрения в

[7] Например, тяжело больные и немощные пациенты легче контактируют с ассистентами-роботами, чем с живыми медсёстрами. Можно предположить, что операторам в опасных и уязвимых ситуациях

злонамеренности большее недоверие вызывают когнитивные агенты, похожие на людей, а не на машины (Zak et al., 2005).

Помимо вышеуказанных факторов, повышение доверия и снижение недоверия ИИ могут быть достигнуты путем создания обоснованных регламентов взаимодействия с ним человека-оператора, что позволит гармонизировать кооперацию элементов человеко-машинной системы, использующей алгоритмы ИИ. Важнейшая предпосылка поддержания и укрепления доверия операторов – обучение, регулярное переобучение и специализированный тренинг персонала, использующего и сопровождающего системы искусственного интеллекта на предприятиях. Последнее требование может стать предметом отдельного обсуждения.

ЛИТЕРАТУРА

Акимова А.Ю. Доверие и недоверие человека технике: Социально-психологический подход / Под ред. А.А. Обознова. М.: Изд-во «Институт психологии РАН», 2020.

Акимова А.Ю., Обознов А.А. Доверие и недоверие человека технике // Психологический журнал. 2016. Т. 37. № 6. С. 56-69.

Акимова А.Ю., Обознов А.А. Факторы повышения доверия и снижения недоверия человека технике // Психологические исследования. 2017. Т. 10. №53. С. 8. DOI: 10.54359/ps.v10i53.369.

Большие системы: моделирование организационных механизмов / В.Н. Бурков, Б. Данев, А.К. Еналеев, В.В. Кондратьев, Т.Б. Нанева, А.В. Щепкин. М.: Наука, 1989.

Венгер А.Л. Математическое моделирование эмоциональных процессов // Автоматизация в промышленности. 2013. № 7. С. 59–63.

Венгер А.Л. Математическая модель принятия решений в экстремальной ситуации // Автоматизация в промышленности. 2018. № 6. С. 32–36.

Голиков Ю. Я. Методология психологических проблем проектирования техники. М.: Пер СЭ, 2003.

предпочтительней коммуницировать с когнитивными агентами без выраженных антропоморфных признаков.

Дозорцев В.М., Венгер А.Л. О проблеме доверия человека-оператора искусственному интеллекту // Автоматизация в промышленности. 2022. № 2. С.10–17. DOI: 10.25728/avtprom.2022.02.02.

Костин А.Н. Избранные труды: психологические проблемы автоматизации высоких технологий. М.: Изд-во «Институт психологии РАН», 2021. Сер. Психологическое наследие. DOI: 10.38098/ptrmn.2021.003.

Купрейченко А. Б. Психология доверия и недоверия. М.: Изд-во «Институт психологии РАН», 2008.

Ломов Б. Ф. Человек и техника. М.: Советское радио, 1966.

Мельникова Ю. ИИ в промышленности почувствовал почву // ComNews, 30.09.2021. URL: https://www.comnews.ru/content/216670/2021-09-30/2021-w39/ii-promyshlennosti-pochuvstvoval-pochvu (дата обращения: 25.06.2022).

Минский М. Персептроны / М. Минский, С. Пейперт; пер. с англ. Г.Л. Гимельфарба и В.М. Шарыпанова. М.: Мир, 1971.

Нейман Дж. Теория игр и экономическое поведение / Дж. Нейман, О. Моргенштерн; пер. с англ. Н.Н. Воробьёва. М.: Наука, 1970.

Развитие отдельных высокотехнологичных направлений (Белая книга). М.: НИУ ВШЭ, 2022. URL: https://issek.hse.ru/mirror/pubs/share/565446894.pdf (дата обращения: 25.06.2022).

Селигман М. Как научиться оптимизму. Измените взгляд на мир и свою жизнь / Пер. с англ. Альпина Паблишер. М.: Изд-во «Альпина Паблишер», 2017.

Фалин Г.И. Математический анализ рисков в страховании. М.: Российский юридический издательский дом, 1994.

Хорошилов Б.М. Семантизация машины как "существа" во взаимодействии человека и сложной техники в профессиональной деятельности // Вестник Новосибирского государственного университета. Серия: Психология. 2009. Т. 3. № 1. С. 24-34.

Этичное применение искусственного интеллекта / П.М. Готовцев, Р.В.Душкин, О.П. Кузнецов, В.Э. Мильке, А.В. Незнамов, Е.Г. Потапова. // Этические проблемы цифровых технологий. Аналитический доклад, 2020. URL: https://ethics.cdto.ranepa.ru/3_3 (дата обращения: 25.06.2022).

Dzindolet M.T. The Perceived Utility of Human and Automated Aids in a Visual Detection Task / M.T. Dzindolet, L.G. Pierce, H.P. Beck, L.A. Dawe. // Human Factors. 2002. V. 44. № 1. P. 79-94. DOI: 10.1518/0018720024494856.

Hart G., Goldwater B. Recent False Alerts from the Nation's Missile Attack Warning System. Washington: U.S. Government Printing Office, 1980. URL:

https://babel.hathitrust.org/cgi/pt?id=uc1.31210005931942&view=1up&seq=1&skin=2021 (дата обращения: 25.06.2022).

Jones S.E. Against technology: from the Luddites to Neo-Luddism. N.Y.: Taylor & Francis, 2006.

Lee J., See K. Trust in technology: Designing for appropriate reliance // Human Factors. 2004. V. 46. № 1. P. 50–58. DOI: 10.1518/hfes.46.1.50_30392.

Lewicki R.J. Trust and distrust: New relationships and realities / R.J. Lewicki, D.J. McAllister, R.J. Bies // The Academy of Management Review. 1998. Vol. 23. № 3. Pp. 438–458. DOI: 10.2307/259288.

McCarthy J. What is artificial intelligence. Stanford University, 1998. URL: http://www-formal.stanford.edu/jmc/whatisai/whatisai.html (дата обращения: 25.06.2022).

Postman N. Technopoly: The Surrender of Culture to Technology. N.Y.: Vintage Books, 1993.

Rotter J.B. Internal versus external control of reinforcement: A case history of a variable // American Psychologist. 1990. V. 45. №. 4. P. 489-493. DOI: 10.1037/0003-066X.45.4.489.

Zak P.J. Oxytocin is associated with human trustworthiness / P.J. Zak, R. Kurzban, W.T. Matzner. // Hormones and Behavior. 2005. V. 48. № 5. P. 522-527. DOI: 10.1016/j.yhbeh.2005.07.009.

Статья поступила в редакцию: 07.06.2022. Статья опубликована: 10.07.2022.

ON THE PROBLEM OF HUMAN OPERATORS' TRUST IN ARTIFICIAL INTELLIGENCE

© 2022 Victor M. Dozortsev*, Alexander L. Venger**

** Doctor of Technical Sciences, Development Director, MIPT Center for Digital Technologies LLC*
E-mail: victor.dozortsev@mipt-cdt.ru

*** Doctor of Psychological Sciences, Professor of Psychology Department of Dubna State University*
E-mail: alvenger@gmail.com

Artificial intelligence (AI) is rapidly penetrating industrial automation, particularly as part of recommendation systems and intelligent decision-making systems that give advice to human operators on managing complex technological systems. This qualitatively strengthens human-machine control systems, but also generates new challenges to ensure the safety of workers and production assets, at the center of which is the problem of the human operator' trust/distrust in algorithms based on AI. This problem is considered in the paper in the general context of trust/distrust in technics against the background of its progressive complication and intellectualization. Using practical examples, the specifics of the use of AI algorithms in the control of complex technological systems are discussed. The psychological (including individual psychological) components of the problem are analyzed. A model of operator decision-making in potentially dangerous situations is proposed and studied. A meaningful interpretation of the model is given in terms of a subjectively acceptable level of risk and personal anxiety on the example of a predictive analytics of technological equipment condition, when ignoring the correct advice to stop the equipment received by the operator from AI algorithms is fraught with catastrophic losses, and taking the wrong advice is an unjustified expenditure of material resources and time. Measures to increase trust and reduce distrust of AI algorithms are discussed, including maintaining permanent operator activity in the "human-operator—AI" control system, considering the motivational component of operator decision-making in the presence of AI advice, ensuring transparency and interpretability of advice, developing reasonable regulations for interaction between the operator and AI within the framework of intelligent control systems.

Keywords: human operator, trust/distrust in artificial intelligence, recommendation systems, intelligent decision-making systems, subjectively acceptable level of risk, predictive analytics of equipment condition.

REFERENCES

Akimova, A.Yu. (2020). Doverie i nedoverie cheloveka tekhnike: Sotsial'no-psikhologicheskij podkhod [Human trust and distrust in technics: A Socio-psychological approach]. A.A. Oboznov (Ed.). Moscow: Institute of Psychology RAS Publ. (in Russian).

Akimova, A.Yu., & Oboznov, A.A. (2016). Doverie i nedoverie cheloveka tekhnike [Human trust and distrust of technics]. *Psikhologicheskij zhurnal* [*Psychological Journal*], Vol. 37, 2, 56-69 (in Russian).

Akimova, A.Yu., & Oboznov, A.A. (2017). Faktory povysheniya doveriya i snizheniya nedoveriya cheloveka tekhnike [Factors of increasing trust and reducing human distrust in technics]. *Psikhologicheskie issledovaniya* [*Psychological research*], Vol. 10, 53, 8 (in Russian). DOI: 10.54359/ps.v10i53.369.

Burkov, V.N., Danev, B., Enaleev, A.K., Kondrat'ev, V.V., Naneva T.B., & Shchepkin, A.V. (1989). Bol'shie Sistemy: Modelirovanie Organizatsionnykh Mekhanizmov [Large systems: modeling of organizational mechanisms]. Moscow: Nauka (in Russian).

Venger, A.L. (2013). Matematicheskoe modelirovanie emotsional'nykh protsessov [Mathematical modeling of emotional processes]. *Avtomatizatsiya v promyshlennosti* [*Automation in Industry*], Vol. 7, 59-63 (in Russian).

Venger, A.L. (2018). Matematicheskaya model' prinyatiya reshenij v ekstremal'noj situatsii [Mathematical model of decision-making in an extreme situation]. *Avtomatizatsiya v promyshlennosti* [*Automation in Industry*], Vol. 6, 32-36 (in Russian).

Golikov, Yu.Ya. (2003). Metodologiya Psikhologicheskikh Problem Proektirovaniya Tekhniki [Methodology of psychological problems of technics design]. Moscow: Per SE (in Russian).

Dozortsev, V.M., & Venger, A.L. (2022). O probleme doveriya cheloveka-operatora iskusstvennomu intellektu [On the problem of human operator's trust in artificial intelligence]. *Avtomatizatsiya v promyshlennosti* [*Automation in Industry*], Vol. 2, 10-17 (in Russian). DOI: 10.25728/avtprom.2022.02.02.

Kostin, A.N. (2021). Izbrannye Trudy: Psikhologicheskie problemy avtomatizatsii vysokikh tekhnologij – Ser. Psikhologicheskoe nasledie [Selected works: Psychological problems of automation of high technologies – Ser. Psychological heritage]. Moscow: Institute of Psychology RAS Publ. (in Russian). DOI: 10.38098/ptrmn.2021.003.

Kuprejchenko, A.B. (2008). Psikhologiya Doveriya i Nedoveriya [Psychology of trust and distrust]. Moscow: Institute of Psychology RAS Publ. (in Russian).

Lomov, B.F. (1966). Chelovek i tekhnika [Man and technics]. Moscow: Sovetskoe radio (in Russian).

Mel'nikova, Yu. (2021). II v promyshlennosti pochuvstvoval pochvu [AI in industry has felt the ground]. Comnews, 30.09.2021. – URL: https://www.comnews.ru/content/216670/2021-09-30/2021-w39/ii-promyshlennosti-pochuvstvoval-pochvu (Accessed: 25.06.2022) (in Russian).

Minskij, M., & Pejpert, S. (1971). Perseptrony [Perceptrons]. (G.L. Gimel'farb, & V.M. Sharypanov, Trans). Moscow: Mir (in Russian).

Nejman, Dzh., & Morgenshtern, O. (1970). Teoriya igr i Ekonomicheskoe Povedenie [Game theory and economic behavior]. (N.N. Vorob'yov, Trans). Moscow: Nauka (in Russian).

Razvitie otdel'nykh vysokotekhnologichnykh napravlenij (Belaya kniga) (2022). [Development of certain high-tech areas (White Paper)]. Moscow: NIU VSHE. –

URL: https://issek.hse.ru/mirror/pubs/share/565446894.pdf (Accessed: 25.06.2022) (in Russian).

Seligman, M. (2017). Kak Nauchit'sya Optimizmu. Izmenite Vzglyad na Mir i Svoyu Zhizn' [How to learn optimism. Change the way you look at the world and your life]. (Al'pina Publisher, Trans). Moscow: Al'pina Publisher Publ. (in Russian).

Falin, G.I. (1994). Matematicheskij Analiz Riskov v Strakhovanii [Mathematical analysis of risks in insurance]. Moscow: Rossijskij yuridicheskij izdatel'skij dom (in Russian).

Khoroshilov, B.M. (2009). Semantizatsiya mashiny kak "sushchestva" vo vzaimodejstvii cheloveka i slozhnoj tekhniki v professional'noj deyatel'nosti [Semantization of the machine as a "creature" in the interaction of man and complex technics in professional activity]. *Vestnik Novosibirskogo gosudarstvennogo universiteta. Ser.: Psikhologiya* [*Bulletin of Novosibirsk State University. Ser.: Psychology*], Vol. 3, 1, 1-24 (in Russian).

Gotovtsev, P.M., Dushkin, R.V., Kyznetsov, O.P., Mil'ke, V.E., Neznamov, A.V., & Potapova, E.G. (2020). Etichnoe primenenie iskusstvennogo intellekta. [Ethical application of artificial intelligence]. *Eticheskie problemy tsifrovykh tekhnologij. Analiticheskij doklad* [*Ethical problems of digital technologies. Analytical report*] — URL: https://ethics.cdto.ranepa.ru/3_3 (Accessed: 25.06.2022) (in Russian).

Dzindolet, M.T., Pierce, L.G., Beck, H.P., & Dawe, L.A. (2002). The Perceived Utility of Human and Automated Aids in a Visual Detection Task. *Human Factors*, 44, 1, 79-94. DOI: 10.1518/0018720024494856.

Hart, G., Goldwater, B. Recent False Alerts from the Nation's Missile Attack Warning System. Washington: U.S. Government Printing Office, 1980. — URL: https://babel.hathitrust.org/cgi/pt?id=uc1.31210005931942&view=1up&seq=1&skin=2021 (Access: 25.06.2022).

Jones, S.E. (2006). Against technology: from the Luddites to Neo-Luddism. N.-Y.: Taylor & Francis.

Lee, J., & See, K. (2004). Trust in technology: Designing for appropriate reliance. *Human Factors*, 46, 1, 50—58. DOI: 10.1518/hfes.46.1.50_30392.

Lewicki, R.J., McAllister, D.J., & Bies, R.J. (1998). Trust and distrust: New relationships and realities. *The Academy of Management Review*, *23*, *3*, 438—458.

McCarthy, J. (1998). What is artificial intelligence. Stanford University. — URL: http://www-formal.stanford.edu/jmc/whatisai/whatisai.html (Accessed: 25.06.2022).

Postman, N. (1993). Technopoly: The Surrender of Culture to Technology. N.-Y.: Vintage Books.

Rotter, J.B. (1990). Internal versus external control of reinforcement: A case history of a variable. *American Psychologist*, 45, 4, 489-493. DOI: 10.1037/0003-066X.45.4.489.

Zak P.J., Kurzban, R., & Matzner, W.T. (2005). Oxytocin is associated with human trustworthiness. *Hormones and Behavior*, 48, 5, 522-527. DOI: 10.1016/j.yhbeh.2005.07.009.

The article was received: 07.06.2022. Published online: 10.07.2022

Библиографическая ссылка на статью:

Дозорцев В.М., Венгер А.Л. Взаимодействие человека-оператора с искусственным интеллектом: проблема доверия // Институт психологии Российской академии наук. Организационная психология и психология труда. 2022. Т. 7. № 2. С. 204 - 232. DOI: 10.38098/ipran.opwp_2022_23_2_009

Dozortsev, V. M., Venger, A. L. (2022). Vzaimodejstvie cheloveka-operatora s iskusstvennym intellektom: problema doverija [On the Problem of Human Operators' Trust in Artificial Intelligence]. *Institut Psikhologii Rossiyskoy Akademii Nauk. Organizatsionnaya Psikhologiya i Psikhologiya truda [Institute of Psychology of the Russian Academy of Sciences. Organizational Psychology and Psychology of Labor]*, 7 (2), 204 - 232. DOI: 10.38098/ipran.opwp_2022_23_2_009

Адрес статьи: http://work-org-psychology.ru/engine/documents/ document795.pdf