ИНЖЕНЕРНАЯ ПСИХОЛОГИЯ И ЭРГОНОМИКА

УДК 159.9

ГРНТИ 15.81.29

ВЗАИМОДЕЙСТВИЕ ЧЕЛОВЕКА-ОПЕРАТОРА С ИСКУССТВЕННЫМ ИНТЕЛЛЕКТОМ: ПРОБЛЕМА КАЛИБРОВКИ ДОВЕРИЯ

© 2023 г. Е.Н. Гардеева

Аспирант, Институт психологии Российской академии наук;
г. Москва, Россия
e-mail: gardeevaen@my.msu.ru,

В последнее время все большее внимание уделяется развитию систем с элементами искусственного интеллекта, что связано с их многочисленными потенциальными преимуществами, такими как безопасность и повышение эффективности труда. Одним из ключевых факторов, влияющих на внедрение технологий искусственного интеллекта является уровень доверия пользователей к ним. Развитие доверия - это динамический процесс, и для безопасного применения ИИ-систем уровень доверия пользователей должен соответствовать реальным возможностям системы. В противном случае, возможно возникновения отношения «сверхдоверия» и «свернедоверия». В первом случае пользователи переоценивают возможности системы, что может приводить к ошибкам в деятельности и в том числе к несчастным случаям. Сверхнедоверие, напротив, связано с отказом от исправной автоматики, что негативно сказывается на эффективности труда и повышает нагрузку на человека-оператора. Существование феноменов «сверхдоверия» и «сверхнедоверия» ставит перед разработчиками ИИ-систем вопрос о поддержке оптимального уровня доверия пользователя к системе во время взаимодействия с ней. Процесс обеспечения баланса между надежностью системы и доверием пользователя к ней называется «калибровкой» доверия, а результатом этого процесса является отношение «откалиброванного» доверия. Данная статья является

литературным обзором современных зарубежных исследований проблемы калибровки доверия пользователей ИИ-системам за 5 лет (в период с 2018 по 2023 гг.). Таким образом, в работе анализируется проблема «калибровки» доверия, дается определение процесса «калибровки» доверия доверия пользователей ИИ-системам. Также приводятся основные исследовательские модели калибровки доверия ИИ-системам (в том числе модель К. Хоффа и М. Башир и модель Й.Крауса). Рассматриваются практические рекомендации разработчикам ИИ-систем, способствующие калибровке доверия пользователя. Рекомендации направлены на повышение прозрачности системы (прозрачность ограничений, прозрачность намерений). Обсуждаются направления перспективных исследований.

Ключевые слова: человек-оператор, искусственный интеллект, калибровка доверия, сверхдоверие и сверхнедоверие, человеко-машинное взаимодействие, доверие человека технике.

АКТУАЛЬНОСТЬ

Технологии искусственного интеллекта становятся все более распространенными во всех сферах нашей жизни. Примерами применения систем с элементами искусственного интеллекта (ИИ-систем) являются автономные транспортные средства, медицинские услуги, виртуальные агенты, различные веб-сервисы и др. Являясь следствием поступательной «интеллектуализации» техники, применение ИИ-систем следует рассматривать не только в контексте человеко-машинного взаимодействия как проблемы психологии труда и когнитивной эргономики (Дозорцев и Венгер, 2022; Величковский, 2018), но и организационной психологии, поскольку эти системы могут существенно повлиять на принятие решений и управление знаниями в организациях (Danaher, 2017; Brynjolfsson, Mitchell & Rock, 2014; Huang & Rust, 2018).

ВВЕДЕНИЕ

Развитие, усложнение, совершенствование и расширение возможностей ИИ-систем, которые становятся все более «интеллектуальными» и саморегулирующимися, приводят к существенным изменениям традиционного человеко-машинного взаимодействия — оно становится все более интерактивным и приобретает признаки общения между людьми. В результате пользователь начинает относиться к ИИ-системе,

не только как к техническому объекту, но и помощнику (Акимова, 2020). Помимо этого высокая «интеллектуальность» ИИ-систем в той или иной степени превращает их в «когнитивные», т.е. способные к сбору и переработке информации, принятию рациональных решений. Речь идет уже не о технической системе, которой управляет оператор, а о взаимодействии двух «когнитивных» систем (Величковский, 2018). Некоторые исследователи даже отмечают смену парадигмы в данной сфере исследований. Вместо человеко-машинного *взаимодействия* (human machine interaction), все чаще речь идет о человеко-машинном *сотрудничестве* (human machine cooperation) (Matsas, Vosniakos, 2018; Schaefer, Hill & Jentsch, 2019; Jarrahi, 2018). Такой подход предполагает отношение пользователя к ИИ-системе не как инструменту, а как партнеру. Тем самым, ставятся новые задачи перед разработчиками человеко-машинных интерфейсов. Одна из таких задач заключается в том, что человеко-машинный интерфейс должен не просто обеспечивать эффективную коммуникацию пользователя с ИИ-системой, но и создавать у пользователя образ доверия к «электронному партнеру» (Сергеев, 2014). В этой связи возникает необходимость изучения факторов, влияющих на доверие пользователей к ИИ-системам (Asan & Choudhury, 2021).

Вопрос об уместности или неуместности применения информационных технологий в деятельности человека-оператора ставился ещё в 1990-х (Parasuraman & Riley, 1997). В этом исследовании выделяются три негативных сценария взаимодействия с технологиями: «неиспользование технологий» (disuse – отказ), «чрезмерное использование технологий» (misuse – неуместное использование) и «злоупотребление технологиями» (abuse – злонамеренное применение технологий для достижения личной выгоды). Во всех случаях возникает реальная опасность снижения эффективности и надежности системы «человек-машина». Такие «сценарии» взаимодействия принято относить к ошибкам доверия. Многие современные авторы разделяют необходимость подобной классификации и предлагают свои толкования данных феноменов (Дозорцев и Венгер, 2022; Величковский, 2018; Hoff & Bashir, 2015; Lee, See, 2004).

Неуместное применение технологий искусственного интеллекта, проявляющееся в отсутствии должного контроля за ИИ-системой, принято называть «сверхдоверием» (Обознов, Акимова & Рунец, 2021), или в трактовке других авторов «избыточным доверием» (Величковский, 2018). Считается, что сверхдоверие является одной из причин несчастных случаев при взаимодействии человека-оператора с ИИ-системами (Robinette et al., 2016; Salomons et al., 2018), в т.ч., смертельных дорожно-транспортных происшествий (ДТП) с участием робомобилей (Inagaki & Itoh, 2013; Faas, Kao & Baumann, 2020; Faas, Mathis & Baumann, 2020). С другой стороны, чрезмерное недоверие ИИ-системам заставляет пользователей быть излишне скептичными и не использовать полезные функции этих систем (De Visser et al., 2020). Последствием чрезмерного недоверия является неоправданно повышенный контроль ИИ-систем, что приводит к когнитивной перегрузке (Величковский, 2018; Сергеев, 2020). Такая ошибка доверия называется «сверхнедоверием» (Обознов, Акимова & Рунец, 2021) или «избыточным недоверием» (Величковский, 2018). На сегодняшний день специалистами в области когнитивной эргономики и человеко-машинного взаимодействия ведутся активные работы по изучению доверия пользователей «прорывным» технологиям, в т.ч. включая ИИ-системы (Дозорцев и Венгер, 2022; Обознов, Акимова & Рунец, 2021; Сергеев, 2020). Однако, на наш взгляд в них не уделяется должного внимания ключевому вопросу – проблеме *калибровки доверия* пользователей ИИ-системам.

Целью статьи является анализ современных зарубежных исследований проблемы калибровки доверия пользователей ИИ-системам.

Анализируются три аспекта исследований этой проблемы:

- что понимается под «калибровкой» доверия пользователей ИИ-системам;
- исследовательские модели калибровки доверия ИИ-системам;
- содержание практических рекомендаций (если таковые имеются) разработчикам ИИ-систем.

Методология проведения заявленного анализа. Для поиска релевантной литературы использовалась платформа Google Scholar. Запрос осуществлялся по следующим ключевым словам: калибровка доверия, искусственный интеллект (ИИ), взаимодействие агента и человека, взаимодействие робота и человека, адаптивная автоматизация, адаптивные технологии, доверие к роботу, доверие к технологиям. Для анализа литературы также применялась техника перекрестных ссылок.

Мы ограничили поиск статьями, опубликованными за последние 5 лет (с 2018 по 2023 г.г.), чтобы рассмотреть эмпирические работы, сопутствующие современному технологическому развитию ИИ. В результате поиска было обнаружено 237 работ. Их дальнейших анализ и исключение повторяющихся публикаций выявил 79 рецензируемых англоязычных статей и сборников материалов конференций из таких областей, как эргономика, взаимодействие человека и компьютера, взаимодействие человека и робота, информационные системы, информационные технологии и машиностроение. Из них 28 статей представляют эмпирические исследования, прямо или косвенно затрагивающие вопросы калибровки доверия человека к ИИ-системе.

ТЕОРЕТИЧЕСКИЙ ОБЗОР

Доверие признано важным фактором межличностных отношений между людьми, а в перспективе - между людьми и машинами. Исследователи из самых разных областей изучали роль доверия в отношениях между людьми, между людьми и организациями, а также между организациями (Lee, See, 2004). Однако важная роль доверия не ограничивается отношениями между людьми, оно также подчеркивается как ключевой фактор взаимодействий между людьми и компьютерами (Madsen & Gregor, 2000; Alarcon et al., 2017), людьми и роботами (Chi et al, 2021; Hancock et al., 2011; Natarajan & Gombolay, 2020; Sanders et al., 2011), людьми и автоматикой (Hoff & Bashir, 2015; Lee, See, 2004; Muir, 1994).

В работе Э. Гликсон и А. Вулли (Glikson & Woolley, 2020) проводится обстоятельный обзор исследований доверия пользователей ИИ-системам. На основе

анализа исследований в данной области за 20 лет (в период с 1999 по 2019 гг.) показано, как различные характеристики и особенности ИИ-систем влияют на формирование у пользователей доверия к ним на когнитивном и эмоциональном уровне. Обзор ключевых трудностей, с которыми сталкиваются разработчики и пользователи ИИ-систем, рассматривается в статье Локки и соавторов (Lockey et al., 2021). Авторы данной статьи предлагают "мультистейкхолдерный" подход, позволяющий учитывать интересы всех заинтересованных, в создании ИИ-системы, сторон (Lockey et al., 2021). Современное состояние исследований в области доверия во взаимодействии пользователей с системами искусственного интеллекта рассматривается в обзорной в статье Т. Уэно и коллег (Ueno et al., 2022).

Определение ИИ-систем. Системы с элементами искусственного интеллекта - это «системы с функциями, использующими возможности ИИ, которые непосредственно доступны конечному пользователю». Возможности ИИ-систем заключаются в том, чтобы делать прогнозы, давать рекомендации, а также принимать решения, влияющие на реальную или виртуальную среду посредством обучения, рассуждений и самокоррекции (Gillath et al., 2021). Понятие ИИ-систем может быть применено к целому ряду технологий, включая роботов всех видов, виртуальных агентов, агентов с голосом или «воплощенных» агентов (embodied agents), а также алгоритмам опосредствованного взаимодействия людей с помощью интерфейса и т.д.

Определение калибровки доверия. Многие авторы отмечают необходимость соблюдения баланса между возможностями ИИ-систем и доверием пользователя человека к ним, общепринятый термин для описания указанного баланса отсутствует. Одни авторы называет его «уместностью доверия» (trust appropriateness; Jensen, Khan & Albayram, 2020), другие «осведомленной безопасностью» (informed safety; Khastgir et al., 2018), третьи «перцептивной точностью» (perceptual accuracy; (Merritt et al., 2015). Однако, большинство авторов оперирует понятием «откалиброванное доверие» (calibrated trust; Muir, 1987). Результаты проведенного анализа позволяют установить

два определения калибровки доверия, на которые ссылаются чаще всего. Одно определение предложено Б. Мьюир (Muir B.) и формулируется следующим образом: «Процесс приведения доверия в соответствие с объективной мерой надежности называется калибровкой доверия» (Muir, 1994, с. 1918). Другое определение предлагают Д. Ли и К. Си (Lee J. D., See K. A): «Под калибровкой понимается соответствие между доверием человека к автоматике и ее возможностями» (Lee, See, 2004, с. 55). Надо отметить, что актуальность проблемы калибровки доверия пользователя (хотя термин «калибровка» при этом не использовался) ранее подчеркивалась и в работах отечественных психологов. Так в исследованиях А.Н. Костина показано, что излишне высокое доверие космонавта автоматическим системам управления космическим кораблем могло приводить к игнорированию информации о неисправности автоматики, а необоснованное недоверие – отключению исправной автоматики (Костин, 2011). Не вызывает сомнения, что откалиброванный уровень доверия, точно отражающий реальные возможности ИИ-систем (Muir, 1987), является важным условием их безопасного и эффективного использования. Соответственно, знание психологических процессов формирования откалиброванного доверия пользователей ИИ-системам имеет решающее значение для их разработки и эффективного применения.

Исследовательские модели калибровки. Существует целый ряд моделей доверия пользователей ИИ-системам (Madsen & Gregor, 2000; Keller, 2019; Meske & Bunde, 2020; Jacovi A. et al., 2021; Alarcon et al., 2017). Частыми являются ссылки на пятифакторную модель Б. Мьюир (Muir, 1994) и модель Д. Ли и К. Си (Lee, See, 2004), однако согласно проведенному ранее мета-анализу (Ueno et al., 2022) самой популярной является «трехслойная» модель доверия (Trust Layers) К. Хоффа и М. Башир (Hoff & Bashir, 2015). Согласно данной модели, доверие пользователя ИИ-системе можно разделить на три составляющие: диспозиционное доверие (dispositional trust), ситуационное доверие (situational trust) и выученное (learned trust) доверие.

Диспозиционное доверие трактуется как долгосрочная тенденция, обусловленная культурой, возрастом, полом и личностью пользователя и не зависит от контекста его взаимодействия с ИИ-системой или характеристик конкретной ИИ-системы. Ситуационное доверие зависит от контекста взаимодействия пользователя с ИИ-системой – например, сложности задач пользователя, а также его уверенности в себе, способности к концентрации внимания и т.п. Выученное доверие основывается на прошлом опыте пользователя взаимодействия с подобными системами, а также личном опыте текущего взаимодействия с данной конкретной ИИ-системой. Кроме того, в модели К. Хоффа и М. Башир доверие рассматривается на двух этапах: до взаимодействия с ИИ-системой и во время взаимодействия с ней. К первому этапу относятся процессы связанные с формированием диспозиционного, ситуационного и первоначально выученного доверия. Второй этап связан с «динамически выученным доверием» (dynamic learned trust) и охватывает процессы, влияющие на динамику доверия в конкретном моменте взаимодействия пользователя с ИИ-системой. Позднее в своей диссертации Й. Краус (J. Kraus) выделил данный этап в отдельную стадию, назвав её «Калибровка доверия» и адаптировав таким образом модель К. Хоффа и М. Башир (Kraus, 2020). На сегодняшний день модель Й. Крайса является единственной, в которой учитывается процесс калибровки доверия пользователя ИИ-системе.

Методы калибровки. Б. Мьюир (Muir, 1987) описала 4 способа улучшения калибровки доверия пользователя ИИ-системе, которые могут быть сведены к двум: (1) улучшение у пользователя способности оценивать надежность и функционал ИИ-системы и (2) выявление ранее плохо откалиброванных «зон» доверия и их повторная калибровка. Эти «зоны» П. Макдермот и Р. Бринк (McDermott P. L., Brink R. N.) впоследствии назвали «Точками калибровки» (Calibration Points) и разработали одноименный подход, позволяющий инженерам-разработчикам выявлять уязвимости человеко-машинного интерфейса для достижения откалиброванного доверия пользователей ИИ-системам. С положениями и примерами применения данного подхода

можно ознакомиться в их работе (McDermott & Brink, 2019). Обобщая принципы проектирования человеко-машинного интерфейса, можно сказать, что основная задача ИИ-системы состоит в предоставлении «правильной» информации, в «правильное» время и «правильным» способом (Fischer, 2001). Важно отметить, что в данной работе проектирование ИИ-системы понимается как частный случай создания человеко-машинного интерфейса, в связи с чем данные понятия иногда используются как синонимичные.

Практические рекомендации разработчикам ИИ-систем. В данной статье представлены итоги систематизации практических советов по проектированию человеко-машинных интерфейсов, обеспечивающих поддержание откалиброванного доверия пользователя ИИ-системам. Большинство рекомендаций создавались для автомобилей с режимом автоматического вождения, однако эти рекомендации имеют общий характер, благодаря чему применимы и для других ИИ-систем. Для удобства они сгруппированы в соответствии с двумя этапами становления доверия пользователя к ИИ-системе, описанными выше в модели К. Хоффа и М. Башир: до и во время взаимодействия с системой.

До взаимодействия с ИИ-системой. Экспериментально показано, что реалистичная информация об ограничениях системы важна для предотвращения чрезмерного доверия или недоверия (Kraus, 2020; Forster, 2018; Khastgir et al., 2018). В своих исследованиях Й. Краус и коллеги показали, что уровень доверия водителей, знавших заранее о недостатках автомобиля, управляемых ИИ-системами, не снижался при наступлении сбоев этих систем (Kraus, 2020). Это соотносится с данными полученными при исследовании систем поддержки принятия решения (Atoyan, Duquet & Robert, 2006). В связи с этим, С. Хастгир (Khastgir S.) и коллеги вводят специальное понятие «осведомленной безопасности» (informed safety), которое предполагает информирование водителей о безопасных пределах работы ИИ-системы, управляющей автомобилем (робомобилем). Тем самым, водители могут с необходимой точностью

откалибровать уровень своего доверия к ИИ-системе робомобиля на соответствующем уровне (Khastgir et al., 2018). Ю. Чжан (Zhang Y.) и коллеги предположили, что в ситуациях, когда ИИ-системы и пользователь имеют взаимодополняющие границы ошибок, калибровка уровня доверия водителя может быть более эффективной. Данное предположение требует дальнейшей проверки (Zhang, Liao & Bellamy, 2020).

Во время взаимодействия с ИИ-системой. В своей работе М. Фаас (Faas et al., 2021) предлагает помимо передачи информации об автоматическом режиме вождения, использовать дополнительное сообщение о намерении робомобиля уступить пешеходу дорогу и называет это статусно-интентным интерфейсом. Ц. Ши (J. She) в своих работах (She, 2020; She, Neuhoff & Yuan, 2021) развивает эту идею, предлагая применять в проектировании интерфейса стратегию адаптивной коммуникации. Данная стратегия заключается в следующем: робомобиль, завидев пешехода начинает тормозить и сообщает ему об этом с помощью звукового, вербального или, чаще всего, визуального сигнала (например, надписи «я торможу» на дисплее, обращенном к пешеходу). Такое сообщение отражает намерение робомобиля уступить дорогу. Однако, если автоматика зафиксирует, что пешеход сомневается в том, видит ли его автомобиль и насколько безопасно переходить дорогу, то ИИ-система может продемонстрировать для него второе сообщение, подтверждающее намерение уступить дорогу. Согласно предположению авторов исследования (She, 2020; She, Neuhoff & Yuan, 2021) таким образом система будет имитировать поведение водителя, когда тот машет рукой пешеходу, давая ему тем самым понять, что он его «видит» и уступает дорогу. Кроме того, благодаря внедрению данной стратегии пешеход может убедиться в том, что автоматика на робомобиле работают правильно. Данные выводы согласуются с результатами, полученными в исследованиях адаптивной автоматики, применяемой в других областях (Wilson & Russell, 2007; Kaber & Endsley, 2004).

Неоднозначны данные о количестве обратной связи, необходимой пользователю для комфортной и продуктивной эксплуатации ИИ-системы. В исследовании Д. Шин и

коллег (Shin, 2021) была найдена положительная связь между подробностью объяснений, предлагаемых ИИ-системой и удовлетворенностью пользователей. Однако известно, что чрезмерный объем обратной связи приводит к когнитивной перегрузке пользователя (Величковский, 2018). Пока остается неясно, каким образом можно определить оптимальный объем обратной связи для конкретного пользователя в конкретной ситуации. В этой связи, примечательны результаты, полученные Ф. Винтерсбергером и коллегами (Wintersberger et al., 2020). Результаты их исследования не только показали, что обратная связь (об объектах в среде вождения) важна для потенциальных пользователей, но и то, какое количество желаемой обратной связи коррелирует с доверием водителя ИИ-системе в конкретной ситуации вождения. Однако на данный момент неизвестно, каким образом можно диагностировать уровень ситуационного доверия непосредственно в процессе взаимодействия водителя с ИИ-системой, при этом не отвлекая его от контроля за дорожной обстановкой. Попытка решения данного вопроса была предпринята исследовательской группой Ч. Ху (Hu & Wang, 2022). Исследователи представили количественную модель динамического доверия (quantitative trust dynamic model) водителя к системе круиз-контроля, которая учитывая физиологические показатели водителя, предсказывает уровень его доверия ИИ-системе.

Направления будущих исследований. Команды «Человек – ИИ-система» больше не являются чем-то из области фантастики. Для создания эффективных команд необходимо создать структуру, позволяющую людям и ИИ-системам работать вместе как единое целое. Краеугольным камнем этой структуры должна стать способность ИИ-системы сообщать о своих возможностях и информировать пользователя о возможных несоответствиях уровня его доверия ИИ-системе её фактическим возможностям. Для такой калибровки используются два метода в зависимости от текущего уровня доверия пользователя ИИ-системе – в случае сверхдоверия ИИ-система будет пытаться уменьшить доверие, а в случае сверхнедоверия, наоборот, его повысить («восстановить»

как описывается в одной из работ; De Visser et al., 2020). Эти методы выступают как подсказки пользователю для калибровки его доверия (trust cognitive clues - TCC). Подсказки могут быть вербальными, визуальными, звуковыми, физическими или их комбинацией. Существует ряд исследований, посвященных эффективности различных типов подсказок в разных ситуациях. Окумура с коллегами (Okamura & Yamada, 2020) продемонстрировали, что вербальные сигналы более эффективны, чем звуковые и визуальные. Было экспериментально показано, что подсказки для калибровки доверия могут как понижать (Okamura & Yamada, 2020; Perkins, Khavas & Robinette, 2021), так и повышать доверие пользователя (Perkins, Khavas & Robinette, 2021). Примечательно, что данный эффект проявляется вне зависимости от реальной производительности ИИ-системы: подсказки снижали доверие к высоко надежной системе и повышали к системе с низкой надежностью. Кроме того, ставится вопрос об «уважении доверия» пользователя – способности ИИ-системы распознать внешнюю по отношению к ней причину недоверия пользователя (например, плохое настроение пользователя никак не связано с надежностью системы) и просто принять его состояние, не пытаясь повысить или снизить доверие. Таким образом, ставится задача на будущее - проектирование ИИ-системы таким образом, чтобы она могла точно определить, когда необходимо калибровать уровень доверия пользователя, а когда «уважать доверие» пользователя (Perkins, Khavas & Robinette, 2021).

ЗАКЛЮЧЕНИЕ

В статье были проанализированы англоязычные исследования, опубликованные за последние 5 лет и посвященные проблеме калибровки доверия пользователей ИИ-системам в процессе их взаимодействия. В общем смысле под такой калибровкой понимается достижение баланса между уровнем доверия пользователя к ИИ-системе и надежностью этой системы. Наиболее распространенной является «трёхслойная» модель доверия пользователя ИИ-системе, предложенная в работах К. Хоффа и М.Башир. Однако, на наш взгляд, трехэтапная модель Й. Крауса более полно раскрывает процесс

формирования доверия пользователя ИИ-системе за счет включения специального этапа калибровки доверия. Кроме того, перспективными являются исследования, посвященные диагностики уровня ситуационного доверия пользователя ИИ-системе, а также исследования «уважения» ИИ-системой доверия к ней с стороны пользователя.

ЛИТЕРАТУРА

Акимова А.Ю. Доверие и недоверие человека технике: Социально–психологический подход / Под ред. А.А. Обознова. М.: Изд–во «Институт психологии РАН», 2020. 287 с.

Величковский Б.Б. Психологические проблемы когнитивной эргономики // Мир психологии. 2018. Т. 96. №. 4. С. 102–115.

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

Костин А.Н. Автоматизация в пилотируемой космонавтике: проблемы и социально–психологические детерминанты // Национальный психологический журнал. 2011. №. 1. С. 85–89.

Обознов А.А., Акимова А.Ю., Рунец О.В. Феномены сверхдоверия и сверхнедоверия оператора к интерфейсу «Человек–Искусственный интеллект» // Институт психологии Российской академии наук. Организационная психология и психология труда. 2021. Т. 6. №. 2. С. 4–20. DOI: 10.38098/ipran.opwp_2021_19_2_001

Сергеев С.Ф. Методологические проблемы человеко-машинного интерфейса // XII всероссийское совещание по проблемам управления ВСПУ-2014, Москва, 16–19 июля 2014 года. Москва: Институт проблем управления им. В.А. Трапезникова РАН, 2014. С. 6414-6421

Сергеев С.Ф. Психологические аспекты проблемы искусственного интеллекта // Институт психологии Российской академии наук. Организационная психология и психология труда. 2020. Т. 5. №. 4. С. 33–53. DOI: 10.38098/ipran.opwp.2020.17.4.002

Alarcon G.M., Militello L.G., Ryan P., Jessup S.A., Calhoun C.S., Lyons J.B. A descriptive model of computer code trustworthiness // Journal of Cognitive Engineering and Decision Making. 2017. V. 11. №. 2. P. 107–121. DOI: 10.1177/1555343416657236

Asan O., Choudhury A. Research trends in artificial intelligence applications in human factors health care: mapping review // JMIR human factors. 2021. V. 8. №. 2. P. e28236. DOI:10.2196/28236

Atoyan H., Duquet J.R., Robert J.M. Trust in new decision aid systems // Proceedings of the 18th Conference on l'Interaction Homme–Machine (Montreal Canada, April 18 - 21 2006). New York: Association for Computing Machinery. 2006. pp. 115–122. DOI:10.1145/1132736.1132751

Brynjolfsson E., Mitchell T., Rock D. What can machines learn and what does it mean for occupations and the economy? // AEA papers and proceedings. 2014 Broadway, Suite 305, Nashville, TN 37203: American Economic Association, 2018. V. 108. P. 43–47. DOI:10.1257/pandp.20181019

Chi O.H., Jia S., Li Y., Gursoy D. Developing a formative scale to measure consumers' trust toward interaction with artificially intelligent (AI) social robots in service delivery // Computers in Human Behavior. 2021. V. 118. P. 106700. DOI: 10.1016/j.chb.2021.106700

Danaher J. Will life be worth living in a world without work? Technological unemployment and the meaning of life // Science and engineering ethics. 2017. V. 23. №. 1. P. 41–64. DOI: 10.1007/s11948-016-9770-5

De Visser E.J. Towards a theory of longitudinal trust calibration in human–robot teams / E.J. de Visser, M.M.M. Peeters, M.F. Jung, S. Kohn, T.H. Shaw, R. Pak, M.A. Neerincx // International journal of social robotics. 2020. V. 12. №. 2. P. 459–478. DOI: 10.1007/s12369-019-00596-x

Faas S.M., Kraus J., Schoenhals A., Baumann M. Calibrating pedestrians' trust in automated vehicles: does an intent display in an external HMI support trust calibration and safe crossing behavior? // Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems. (Yokohama Japan, May 8-13 2021). New York: Association for Computing Machinery. 2021. P. 1–17. DOI:10.1145/3411764.3445738

Faas S.M., Kao A.C., Baumann M. A longitudinal video study on communicating status and intent for self–driving vehicle–pedestrian interaction // Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI '20) (Honolulu HI, USA, April 25 - 30, 2020). New York: Association for Computing Machinery, 2020. pp. 1–14. DOI:10.1145/3313831.3376484

Faas S.M., Mathis L.A., Baumann M. External HMI for self–driving vehicles: Which information shall be displayed? // Transportation research part F: traffic psychology and behaviour. 2020. V. 68. P. 171–186. DOI:10.1016/j.trf.2019.12.009

Fischer G. User modeling in human–computer interaction // User modeling and user–adapted interaction. 2001. V. 11. P. 65–86. DOI:10.1023/A:1011145532042

Forster Y., Kraus J., Feinauer S., Baumann M. Calibration of trust expectancies in conditionally automated driving by brand, reliability information and introductionary videos: An online study // Proceedings of the 10th International Conference on Automotive User Interfaces and Interactive Vehicular Applications. (Toronto, Canada, September 23 - 25 2018). New York: Association for Computing Machinery, 2018. pp 118–128. DOI: 10.1145/3239060.3239070

Gillath O. Attachment and trust in artificial intelligence / O. Gillath, T. Ai, M.S. Branicky, S. Keshmiri, R.B. Davison, R. Spaulding // Computers in Human Behavior. 2021. V. 115. P. 106607. DOI:10.1016/j.chb.2020.106607

Glikson E., Woolley A.W. Human trust in artificial intelligence: Review of empirical research //Academy of Management Annals. 2020. V. 14. №. 2. P. 627–660. DOI: 10.5465/annals.2018.0057

Hancock P.A. A meta–analysis of factors affecting trust in human–robot interaction / P.A. Hancock, D.R. Billings, K.E. Schaefer, J.Y. Chen, E.J. De Visser, R. Parasuraman // Human factors. 2011. V. 53. №. 5. P. 517–527. DOI:10.1177/0018720811417254

Hoff K.A., Bashir M. Trust in automation: Integrating empirical evidence on factors that influence trust // Human factors. 2015. V. 57. №. 3. P. 407–434. DOI: 10.1177/0018720814547570

Huang M. H., Rust R. T. Artificial intelligence in service // Journal of service research. 2018. V. 21. №. 2. P. 155–172. DOI:10.1177/1094670517752459

Hu C., Wang J. Trust–based and individualizable adaptive cruise control using control barrier function approach with prescribed performance // IEEE Transactions on Intelligent Transportation Systems. 2022. V. 23. №. 7. P. 6974–6984. DOI: 10.1109/TITS.2021.3066154

Inagaki T., Itoh M. Human's overtrust in and overreliance on advanced driver assistance systems: a theoretical framework //International journal of vehicular technology. 2013. DOI: 10.1155/2013/951762

Jacovi A., Marasović A., Miller T., Goldberg Y. Formalizing trust in artificial intelligence: Prerequisites, causes and goals of human trust in AI // Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21). Association for Computing Machinery. (Virtual Event Canada, March 03 - 10 2021). New York: Association for Computing Machinery, 2021. pp. 624–635. DOI:10.1145/3442188.3445923

Jarrahi M.H. Artificial intelligence and the future of work: Human–AI symbiosis in organizational decision making // Business horizons. 2018. V. 61. №. 4. P. 577–586. DOI: 10.1016/j.bushor.2018.03.007

Jensen T., Khan M.M.H., Albayram Y. The role of behavioral anthropomorphism in human–automation trust calibration // Proceedings Artificial Intelligence in HCI: First International Conference, AI-HCI 2020, Held as Part of the 22nd HCI International Conference, HCII 2020 (Copenhagen, Denmark, July 19–24). Berlin, Heidelberg: Springer-Verlag, 2020. 33–53. DOI:10.1007/978-3-030-50334-5_3

Kaber D.B., Endsley M.R. The effects of level of automation and adaptive automation on human performance, situation awareness and workload in a dynamic control task // Theoretical issues in ergonomics science. 2004. V. 5. №. 2. P. 113–153. DOI: 10.1080/1463922021000054335

Keller R.P. Observational oversight for understanding trust in interactive human and AI systems: Doctor's thesis. Monterey, CA; Naval Postgraduate School, 2019.

Khastgir S., Birrell, S., Dhadyalla, G., Jennings, P. Calibrating trust through knowledge: Introducing the concept of informed safety for automation in vehicles // Transportation research part C: emerging technologies. 2018. V. 96. P. 290–303. DOI:10.1016/j.trc.2018.07.001

Kraus J.M. Psychological processes in the formation and calibration of trust in automation : Doctor's thesis. Universität Ulm, 2020.

Lee J.D., See K.A. Trust in automation: Designing for appropriate reliance //Human factors. 2004. V. 46. №. 1. P. 50–80. DOI:10.1518/hfes.46.1.50_30392

Lockey S., Gillespie N., Holm D., Someh I.A. A review of trust in artificial intelligence: Challenges, vulnerabilities and future directions // Proceedings of the 54th Hawaii International Conference on System Sciences (Honolulu, HI, United States, 4-8 January 2021). P5463 URl: http://hdl.handle.net/10125/71284 (Accessed 18.09.2023). DOI: 10.24251/hicss.2021.664

Madsen M., Gregor S. Measuring human–computer trust // Proceedings of the 11th Australasian Conference on Information Systems. (Brisbane, Australia, December 6-8, 2000) / G. Gable, M. Vitale (Eds.). Brisbane, Australia: Information Systems Management Research Centre, 2000 V. 53. pp. 6–8.

McDermott P.L., Brink R.N. Practical guidance for evaluating calibrated trust // Proceedings of the Human Factors and Ergonomics Society Annual Meeting (Seattle, WA, October 28 - November 1 2019). Los Angeles, CA: SAGE Publications, 2019. V. 63. №. 1. P. 362–366. DOI:10.1177/1071181319631379

Matsas E., Vosniakos G.C. Design of a virtual reality training system for human–robot collaboration in manufacturing tasks // International Journal on Interactive Design and Manufacturing (IJIDeM). 2017. V. 11. P. 139–153. DOI: 10.1007/s12008-015-0259-2

Merritt S.M., Lee, D., Unnerstall, J. L., Huber, K. Are well–calibrated users effective users? Associations between calibration of trust and performance on an automation–aided task // Human Factors. 2015. V. 57. №. 1. P. 34–47. DOI:10.1177/0018720814561675

Meske C., Bunde E. Transparency and trust in human–AI–interaction: The role of model–agnostic explanations in computer vision–based decision support // Artificial Intelligence in HCI: First International Conference, AI–HCI 2020, Held as Part of the 22nd HCI International Conference, HCII 2020, Copenhagen, Denmark, July 19–24, 2020, Proceedings 22. Springer International Publishing, 2020. P. 54–69. DOI:10.1007/978-3-030-50334-5_4

Muir B.M. Trust between humans and machines, and the design of decision aids // International journal of man–machine studies. 1987. V. 27. №. 5–6. P. 527–539. DOI:10.1016/S0020-7373(87)80013-5

Muir B.M. Trust in automation: Part I. Theoretical issues in the study of trust and human intervention in automated systems // Ergonomics. 1994. V. 37. №. 11. P. 1905–1922. DOI:10.1080/00140139408964957

Shin D. The effects of explainability and causability on perception, trust, and acceptance: Implications for explainable AI // International Journal of Human-Computer Studies. 2021. V. 146. P. 102551. DOI: 10.1016/j.ijhcs.2020.102551

Okamura K., Yamada S. Empirical evaluations of framework for adaptive trust calibration in human–AI cooperation // IEEE Access. 2020. V. 8. P. 220335–220351. DOI:10.1109/ACCESS.2020.3042556

Parasuraman R., Miller C. A. Trust and etiquette in high–criticality automated systems // Communications of the ACM. 2004. V. 47. №. 4. P. 51–55. DOI:10.1145/975817.975844

Parasuraman R., Riley V. Humans and automation: Use, misuse, disuse, abuse // Human factors. 1997. V. 39. №. 2. P. 230–253. DOI:10.1518/001872097778543886

Perkins R., Khavas Z.R., Robinette P. Trust calibration and trust respect: A method for building team cohesion in human robot teams // Proceedings of the Artificial Intelligence for Human-Robot Interaction (AI-HRI) symposium as part of AAAI-FSS 2021(Washington, DC, USA, November 4-6, 2021). arXiv preprint arXiv:2110.06809. 2021. DOI: 10.48550/arXiv.2110.06809

Salomons N., van der Linden M., Strohkorb Sebo S., Scassellat B. Humans conform to robots: Disambiguating trust, truth, and conformity // Proceedings of the HRI '18: ACM/IEEE International Conference on Human-Robot Interaction. (Chicago, USA, March 5-8 2018). New York: Association for Computing Machinery, 2018. pp. 187–195. DOI:10.1145/3171221.3171282

Schaefer K. E., Hill S. G., Jentsch F. G. Trust in human–autonomy teaming: A review of trust research from the us army research laboratory robotics collaborative technology alliance // Advances in Human Factors in Robots and Unmanned Systems: Proceedings of the AHFE 2018 International Conference on Human Factors in Robots and Unmanned Systems, July 21–25, 2018, Loews Sapphire Falls Resort at Universal Studios, Orlando, Florida, USA 9. Springer International Publishing, 2019. P. 102–114. DOI: 10.1007/978-3-319-94346-6_10

She J. Advisory and adaptive communication improves trust in autonomous vehicle and pedestrian interaction // International Design Engineering Technical Conferences and Computers and Information in Engineering Conference (Virtual Conference, August 17 – 19 2020). NY: American Society of Mechanical Engineers, 2020. V. 83976. P. V008T08A037. DOI: 10.1115/DETC2020-22692

She J., Neuhoff J., Yuan Q. Shaping pedestrians' trust in autonomous vehicles: an effect of communication style, speed information, and adaptive strategy // Journal of Mechanical Design. 2021. V. 143. №. 9. P. 091401. DOI:10.1115/1.4049866

Ueno T. Trust in human–AI interaction: Scoping out models, measures, and methods / T. Ueno, Y. Sawa, Y. Kim, J. Urakami, H. Oura, K. Seaborn // CHI Conference on Human Factors in Computing Systems Extended Abstracts (New Orleans, USA, 29 April 5 May 2022). NY.: Association for Computing Machinery, 2022. P. 1–7. DOI:10.1145/3491101.3519772

Wilson G.F., Russell C.A. Performance enhancement in an uninhabited air vehicle task using psychophysiologically determined adaptive aiding // Human factors. 2007. V. 49. №. 6. P. 1005–1018. DOI: 10.1518/001872007x249875

Wintersberger P. Explainable automation: Personalized and adaptive UIs to foster trust and understanding of driving automation systems / P. Wintersberger, H. Nicklas, T. Martlbauer, S. Hammer, A. Riener // 12th International Conference on Automotive User Interfaces and Interactive Vehicular Applications (Virtual Event DC USA, September 21 - 22, 2020). NY.: Association for Computing Machinery, 2020. pp. 252–261. DOI: 10.1145/3409120.3410659

Zhang Y., Liao Q.V., Bellamy R.K.E. Effect of confidence and explanation on accuracy and trust calibration in AI–assisted decision making // Proceedings of the 2020 conference on fairness, accountability, and transparency (Barcelona Spain, January 27 - 30 2020).

NY.: Association for Computing Machinery, 2020. pp. 295–305. DOI: 10.1145/3351095.3372852

Статья поступила в редакцию: 10.09.2023. Статья опубликована: 20.10.2023.

THE PROBLEM OF CALIBRATING USER TRUST IN ARTIFICIAL INTELLIGENCE IN FOREIGN STUDIES

© 2023 E.N. Gardeeva

Postgraduate student, Institute of Psychology, Russian Academy of Sciences;
r. Moscow, Russia
e-mail: gardeevaen@my.msu.ru

Recently, the development of systems with artificial intelligence elements (AI-systems) has received increasing attention due to their numerous potential benefits such as safety and improved work efficiency. One of the key factors influencing the acceptance of artificial intelligence technologies is the level of user trust in them. The development of trust is a dynamic process, and for safe application of AI-systems, the level of user trust must match the real capabilities of the system. Otherwise, the relations of «overtrust» and «undertrust» may arise. In the first case, users overestimate the capabilities of the system, which can lead to errors in activities, including accidents. Undertrust, on the contrary, is associated with the rejection of well-functioning automation, which negatively affects work efficiency and increases the load on the user. The existence of the phenomena of «overtrust» and «undertrust» raises the question for the developers of AI-systems of maintaining an optimal level of user trust in the system during interaction with it. The process of balancing the reliability of the system and the user's trust in it is called trust "calibration", and the result of this process is a relation of «calibrated» trust. This paper is a literature review of recent foreign research on the problem of calibrating user trust in AI-systems over a 5-year period (between 2018 and 2023). Thus, the paper analyses the problem of trust «calibration», defines the process of "calibrating" user trust in AI-systems. The main research models of trust calibration of AI-systems (including the model of K. Hoffa and M. Bashir and the model of J. Kraus) are also given. Practical recommendations for AI-system developers to help calibrate user trust are discussed. The recommendations are aimed at increasing the transparency of the system (transparency of limitation, transparency of intentions). Directions for prospective research are discussed.

Key words: human-operator, artificial intelligence, trust calibration, undertrust and overtrust, human-machine interaction.

REFERENCES

Akimova, A.Yu. (2020). *Doveriye i nedoveriye cheloveka tekhnike: Sotsial'no–psikhologicheskiy podkhod* [*Doveriye i nedoveriye cheloveka tekhnike: A socio-psychological approach*]. A.A. Oboznov (Ed.). Moscow: «Institut psihologii RAN»Publ. (in Russian).

Velichkovsky, B.B. (2018). Psikhologicheskiye problemy kognitivnoy ergonomiki [Psychological problems of cognitive ergonomics]. *Mir psihologii* [*World of Psychology*], 96 (4), 102–115. (in Russian).

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. (in Russian). DOI: 10.38098/ipran.opwp_2022_23_2_009

Kostin A.N. (2011). Avtomatizatsiya v pilotiruyemoy kosmonavtike: problemy i sotsial'no-psikhologicheskiye determinanty [Automation in manned astronautics: problems and socio-psychological determinants]. *Natsional'nyy psikhologicheskiy zhurnal* [*National psychological journal*]. 1, 85–89. (in Russian).

Oboznov, A.A., Akimova, A.Yu. & Runets, O.V. (2021). Fenomeny sverhdoverija i sverhnedoverija operatora k interfejsu «chelovek - iskusstvennyj intellekt» [The phenomena of operator over-trust and overmistrust to the interface "human - 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 work*]. 6 (2), 4 – 20. (in Russian). DOI: 10.38098/ipran.opwp_2021_19_2_001

Sergeev, S.F. (2014). Metodologicheskiye problemy cheloveko-mashinnogo interfeysa [Methodological problems of human-machine interface]. *XII vserossiyskoye soveshchaniye po problemam upravleniya VSPU-2014 (16-19 iyunya 2014 goda).* [*XII All-Russian meeting on problems of management VSPU-2014 (July 16–19, 2014)*]. (pp. 6414–6421). Moscow, Institut problem upravleniya Rossiyskoy Akademii Nauk Publ. (In Russian).

Sergeev, S.F. (2020). Psihologicheskie aspekty problemy iskusstvennogo intelekta [Psychological aspects of the problem of 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*], *5* (4), 33–53. (In Russian). DOI: https://doi.org/10.38098/ipran.opwp.2020.17.4.002

Alarcon, G.M., Militello, L.G., Ryan, P., Jessup, S.A., Calhoun, C.S., & Lyons, J.B. (2017). A descriptive model of computer code trustworthiness. *Journal of Cognitive Engineering and Decision Making*, *11*(2), 107-121. DOI: 10.1177/1555343416657236

Asan, O., & Choudhury, A. (2021). Research trends in artificial intelligence applications in human factors health care: mapping review. *JMIR human factors*, *8*(2), e28236. DOI:10.2196/28236

Atoyan, H., Duquet, Jean-Rémi, & Robert, J-M. (2006). Trust in new decision aid systems. In Proceedings of the *18th Conference on l'Interaction Homme-Machine* (*IHM '06*). (*Montreal Canada, April 18 - 21 2006*). (pp. 115–122). New York: Association for Computing Machinery. DOI:10.1145/1132736.1132751

Brynjolfsson, E., Mitchell, T., & Rock, D. (2018, May). What can machines learn and what does it mean for occupations and the economy? In *AEA papers and proceedings* (Vol. 108, pp. 43-47). 2014 Broadway, Suite 305, Nashville, TN 37203: American Economic Association. DOI:10.1257/pandp.20181019

Chi, O.H., Jia, S., Li, Y., & Gursoy, D. (2021). Developing a formative scale to measure consumers' trust toward interaction with artificially intelligent (AI) social robots in service delivery. *Computers in Human Behavior*, *118*, 106700. DOI: 10.1016/j.chb.2021.106700

Danaher, J. (2017). Will life be worth living in a world without work? Technological unemployment and the meaning of life. *Science and engineering ethics*, *23*(1), 41-64. DOI:10.1007/s11948-016-9770-5

De Visser, E.J., Peeters, M.M., Jung, M.F., Kohn, S., Shaw, T.H., Pak, R., & Neerincx, M.A. (2020). Towards a theory of longitudinal trust calibration in human–robot teams. *International journal of social robotics*, *12*(2), 459-478. DOI: 10.1007/s12369-019-00596-x

Faas, S.M., Kraus, J., Schoenhals, A., & Baumann, M. (2021). Calibrating Pedestrians' Trust in Automated Vehicles: Does an Intent Display in an External HMI Support Trust Calibration and Safe Crossing Behavior? In Proceedings of the *2021 CHI Conference on Human Factors in Computing Systems (CHI '21)* (*Yokohama Japan, May 8-13 2021*). (Article 157, pp. 1–17.). New York: Association for Computing Machinery. DOI: https://doi.org/10.1145/3411764.3445738

Faas, S.M., Kao, A.C., & Baumann, M. (2020, April). A longitudinal video study on communicating status and intent for self-driving vehicle–pedestrian interaction. In

Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI '20). (Honolulu HI, USA, April 25 - 30, 2020). (pp. 1-14). *New York: Association for Computing Machinery*. DOI:10.1145/3313831.3376484

Faas, S.M., Mathis, L.A., & Baumann, M. (2020). External HMI for self-driving vehicles: Which information shall be displayed?. *Transportation research part F: traffic psychology and behaviour*, *68*, 171-186. DOI: 10.1016/j.trf.2019.12.009

Fischer, G. (2001). User modeling in human—computer interaction. *User modeling and user-adapted interaction*, *11*, 65-86. DOI: 10.1023/A:1011145532042

Forster, Y., Kraus, J., Feinauer, S., & Baumann, M. (2018, September). Calibration of trust expectancies in conditionally automated driving by brand, reliability information and introductionary videos: An online study. In *Proceedings of the 10th International Conference on Automotive User Interfaces and Interactive Vehicular Applications. (Toronto, Canada, September 23 - 25 2018). (pp. 118-128).* New York: Association for Computing Machinery. DOI: 10.1145/3239060.3239070

Gillath, O., Ai, T., Branicky, M.S., Keshmiri, S., Davison, R.B., & Spaulding, R. (2021). Attachment and trust in artificial intelligence. *Computers in Human Behavior*, *115*, 106607. DOI: 10.1016/j.chb.2020.106607

Glikson, E., & Woolley, A.W. (2020). Human trust in artificial intelligence: Review of empirical research. *Academy of Management Annals*, *14*(2), 627-660. DOI: 10.5465/annals.2018.0057

Hancock, P.A., Billings, D.R., Schaefer, K.E., Chen, J.Y., De Visser, E.J., & Parasuraman, R. (2011). A meta-analysis of factors affecting trust in human-robot interaction. *Human factors*, *53*(5), 517-527. DOI: 10.1177/0018720811417254

Hoff, K.A., & Bashir, M. (2015). Trust in automation: Integrating empirical evidence on factors that influence trust. *Human factors*, *57*(3), 407-434. DOI: 10.1177/0018720814547570

Huang, M.H., & Rust, R.T. (2018). Artificial intelligence in service. *Journal of service research*, *21*(2), 155-172. DOI: 10.1177/1094670517752459

Hu, C., & Wang, J. (2022). Trust-based and individualizable adaptive cruise control using control barrier function approach with prescribed performance. *IEEE Transactions on Intelligent Transportation Systems*, *23*(7), 6974-6984. DOI: 10.1109/TITS.2021.3066154

Inagaki, T., & Itoh, M. (2013). Human's overtrust in and overreliance on advanced driver assistance systems: a theoretical framework. *International journal of vehicular technology*, *2013*. DOI: 10.1155/2013/951762

Jacovi, A., Marasović, A., Miller, T., & Goldberg, Y. (2021, March). Formalizing trust in artificial intelligence: Prerequisites, causes and goals of human trust in AI. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21). Association for Computing Machinery. (Virtual Event Canada, March 03 - 10 2021). (pp. 624-635).* New York: Association for Computing Machinery. DOI: 10.1145/3442188.3445923

Jarrahi, M.H. (2018). Artificial intelligence and the future of work: Human-AI symbiosis in organizational decision making. *Business horizons*, *61*(4), 577-586. DOI: 10.1016/j.bushor.2018.03.007

Jensen, T., Khan, M.M.H., & Albayram, Y. (2020, July). The role of behavioral anthropomorphism in human-automation trust calibration. In *Proceedings Artificial Intelligence in HCI: First International Conference, AI-HCI 2020, Held as Part of the 22nd HCI International Conference, HCII 2020 (Copenhagen, Denmark, July 19–24). (pp. 33-53).* Berlin, Heidelberg: Springer-Verlag Publ. DOI: 10.1007/978-3-030-50334-5_3

Kaber, D.B., & Endsley, M.R. (2004). The effects of level of automation and adaptive automation on human performance, situation awareness and workload in a dynamic control task. *Theoretical issues in ergonomics science*, *5*(2), 113-153. DOI: 10.1080/1463922021000054335

Keller, R.P. (2019). Observational oversight for understanding trust in interactive human and AI systems. *Doctor's thesis*. Monterey, CA; Naval Postgraduate School.

Khastgir, S., Birrell, S., Dhadyalla, G., & Jennings, P. (2018). Calibrating trust through knowledge: Introducing the concept of informed safety for automation in vehicles. *Transportation research part C: emerging technologies*, *96*, 290-303. DOI: 10.1016/j.trc.2018.07.001

Kraus, J.M. (2020). Psychological processes in the formation and calibration of trust in automation. *Doctor's thesis*. Universität Ulm.

Lee, J.D., & See, K.A. (2004). Trust in automation: Designing for appropriate reliance. *Human factors*, *46*(1), 50-80. DOI:10.1518/hfes.46.1.50_30392

Lockey, S., Gillespie, N., Holm, D., & Someh, I.A. (2021). A review of trust in artificial intelligence: challenges, vulnerabilities and future directions. In *Proceedings of the 54th Hawaii International Conference on System Sciences* (*Honolulu, HI, United States, 4-8 January 2021*). P5463. URl: http://hdl.handle.net/10125/71284 (Accessed 18.09.2023). DOI: 10.24251/hicss.2021.664

Madsen, M., & Gregor, S. (2000). Measuring human-computer trust. In G. Gable & M. Vitale (Eds.) Proceedings of the 11th Australasian Conference on Information Systems

(Vol. 53)(pp. 6-8). Brisbane, Australia: Information Systems Management Research Centre.

McDermott, P.L., & Brink, R.N.T. (2019, November). Practical guidance for evaluating calibrated trust. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (*Seattle, WA, October 28 - November 1 2019*) (Vol. 63, No. 1, pp. 362-366). Los Angeles, CA: SAGE Publications. DOI:10.1177/1071181319631379

Matsas, E., & Vosniakos, G.C. (2017). Design of a virtual reality training system for human–robot collaboration in manufacturing tasks. *International Journal on Interactive Design and Manufacturing (IJIDeM)*, *11*, 139-153. DOI:10.1007/s12008-015-0259-2

Merritt, S.M., Lee, D., Unnerstall, J.L., & Huber, K. (2015). Are well-calibrated users effective users? Associations between calibration of trust and performance on an automation-aided task. *Human Factors*, *57*(1), 34-47. DOI:10.1177/0018720814561675

Meske, C., & Bunde, E. (2020). Transparency and trust in human-AI-interaction: The role of model-agnostic explanations in computer vision-based decision support. In *Artificial Intelligence in HCI: First International Conference, AI-HCI 2020, Held as Part of the 22nd HCI International Conference, HCII 2020, Copenhagen, Denmark, July 19–24, 2020, Proceedings 22* (pp. 54-69). Springer International Publishing. DOI:10.1007/978-3-030-50334-5_4

Muir, B.M. (1987). Trust between humans and machines, and the design of decision aids. *International journal of man-machine studies*, *27*(5-6), 527-539. DOI: 10.1016/S0020-7373(87)80013-5

Muir, B.M. (1994). Trust in automation: Part I. Theoretical issues in the study of trust and human intervention in automated systems. *Ergonomics*, *37*(11), 1905-1922. DOI:10.1080/00140139408964957

Shin, D. (2021). The effects of explainability and causability on perception, trust, and acceptance: Implications for explainable AI. *International Journal of Human-Computer Studies*. 146. 102551. DOI: 10.1016/j.ijhcs.2020.102551

Okamura, K., & Yamada, S. (2020). Empirical evaluations of framework for adaptive trust calibration in human-AI cooperation. *IEEE Access*, *8*, 220335-220351. DOI: 10.1109/ACCESS.2020.3042556

Parasuraman, R., & Miller, C.A. (2004). Trust and etiquette in high-criticality automated systems. *Communications of the ACM*, 47(4), 51-55. DOI: 10.1145/975817.975844

Parasuraman, R., & Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. *Human factors*, *39*(2), 230-253. DOI: 10.1518/001872097778543886

Perkins, R., Khavas, Z.R., & Robinette, P. (2021). Trust calibration and trust respect: A method for building team cohesion in human robot teams. In *Proceedings of the Artificial Intelligence for Human-Robot Interaction (AI-HRI) symposium as part of AAAI-FSS 2021 (Washington, DC, USA, November 4-6, 2021).* arXiv preprint arXiv:2110.06809. DOI: 10.48550/arXiv.2110.06809

Salomons, N., Van Der Linden, M., Strohkorb Sebo, S., & Scassellati, B. (2018, February). Humans conform to robots: Disambiguating trust, truth, and conformity. In *Proceedings of the HRI '18: ACM/IEEE International Conference on Human-Robot Interaction. (Chicago, USA, March 5-8 2018)* (pp. 187-195). New York: Association for Computing Machinery. DOI:10.1145/3171221.3171282

Schaefer, K.E., Hill, S.G., & Jentsch, F.G. (2019). Trust in human-autonomy teaming: A review of trust research from the us army research laboratory robotics collaborative technology alliance. In *Advances in Human Factors in Robots and Unmanned Systems: Proceedings of the AHFE 2018 International Conference on Human Factors in Robots and Unmanned Systems, July 21-25, 2018, Loews Sapphire Falls Resort at Universal Studios, Orlando, Florida, USA 9* (pp. 102-114). Springer International Publishing. DOI: 10.1007/978-3-319-94346-6_10

She, J. (2020, August). Advisory and adaptive communication improves trust in autonomous vehicle and pedestrian interaction. In *International Design Engineering Technical Conferences and Computers and Information in Engineering Conference (Virtual Conference, August 17 – 19 2020).* (Vol. 83976, p. V008T08A037). NY: American Society of Mechanical Engineers. DOI: 10.1115/DETC2020-22692

She, J., Neuhoff, J., & Yuan, Q. (2021). Shaping pedestrians' trust in autonomous vehicles: an effect of communication style, speed information, and adaptive strategy. *Journal of Mechanical Design*, *143*(9), 091401. DOI:10.1115/1.4049866

Ueno, T., Sawa, Y., Kim, Y., Urakami, J., Oura, H., & Seaborn, K. (2022, April). Trust in human-AI interaction: Scoping out models, measures, and methods. In *CHI Conference on Human Factors in Computing Systems Extended Abstracts (New Orleans, USA, 29 April 5 May 2022)* (pp. 1-7). NY.: Association for Computing Machinery. DOI: 10.1145/3491101.3519772

Wilson, G.F., & Russell, C.A. (2007). Performance enhancement in an uninhabited air vehicle task using psychophysiologically determined adaptive aiding. *Human factors*, *49*(6), 1005-1018. DOI: 10.1518/001872007x249875

Wintersberger, P., Nicklas, H., Martlbauer, T., Hammer, S., & Riener, A. (2020, September). Explainable automation: Personalized and adaptive UIs to foster trust and understanding of driving automation systems. In *12th International Conference on Automotive User Interfaces and Interactive Vehicular Applications (Virtual Event DC*

USA, September 21 - 22, 2020) (pp. 252-261). NY.: Association for Computing Machinery. DOI: 10.1145/3409120.3410659

Zhang, Y., Liao, Q.V., & Bellamy, R.K. (2020, January). Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. In Proceedings of the 2020 conference on fairness, accountability, and transparency (Barcelona Spain, January 27-30 2020) (pp. 295-305). NY.: Association for Computing Machinery. DOI: 10.1145/3351095.3372852

The article was received: 10.09.2023. Published online: 20.10.2023

Библиографическая ссылка на статью:

Гардеева Е.Н. Взаимодействие человека-оператора с искусственным интеллектом: проблема калибровки доверия // Институт психологии Российской академии наук. Организационная психология и психология труда, 2023. Т. 8, № 3, С. 121 - 146. DOI: 10.38098/ipran.opwp_2023_28_3_006

Gardeeva, E.N. (2023). Vzaimodejstvie cheloveka-operatora s iskusstvennym intellektom: problema kalibrovki doverija [The problem of calibrating user trust in artificial intelligence in foreign studies]. *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].* 8(3), 121 - 146. DOI: 10.38098/ipran.opwp_2023_28_3_006

Адрес статьи: http://work-org-psychology.ru/engine/documents/document931.pdf