

Научная статья

УДК 159.9.072:004.8

DOI: 10.18384/3033-6414-2026-3-58-79

ПСИХОЛОГИЧЕСКИЕ МЕХАНИЗМЫ ФОРМИРОВАНИЯ ДОВЕРИЯ К ИСКУССТВЕННОМУ ИНТЕЛЛЕКТУ: КРИТИЧЕСКИЙ АНАЛИЗ ОПРЕДЕЛЕНИЙ И ИНТЕГРАТИВНАЯ МОДЕЛЬ

Отчерцова В. О.^{1,2*}, Блинова Е. Е.¹

¹Херсонский государственный педагогический университет, г. Херсон, Российская Федерация

²Мелитопольский государственный университет, г. Мелитополь, Российская Федерация

*Корреспондирующий автор, e-mail: votchertsova@gmail.com

Поступила в редакцию 16.12.2025

После доработки 17.12.2025

Принята к публикации 23.12.2025

Аннотация

Цель. Провести критический анализ существующих определений доверия к искусственному интеллекту (ИИ), выявить ключевые противоречия в концептуализации феномена и предложить интегративное определение, учитывающее специфику взаимодействия человека с интеллектуальными агентами.

Процедура и методы. Проведён систематический критический анализ определений доверия к ИИ из источников 2020–2025 гг., включая крупномасштабные эмпирические исследования, систематические обзоры, теоретические работы. Анализ проводился по критериям: структурные компоненты, antecedentes, типология, уровни анализа, культурная специфичность. Метод теоретического исследования включал отбор источников по критериям эмпирической обоснованности, теоретической рефлексивности и актуальности, систематизацию определений, сравнительный анализ по ключевым измерениям, выявление противоречий и синтез интегративного определения.

Результаты. Выявлено отсутствие консенсуса в определениях: доверие к ИИ концептуализируется как психологическая установка, многоуровневый социальный феномен или редуцируется до технических характеристик. Обнаружены десять ключевых парадоксов: прозрачности, опыта, антропоморфизма, предубеждения к автоматизации против неприятия алгоритмов, культурный и другие. Доверие к ИИ отличается от межличностного семью уникальными чертами: непрозрачность, антропоморфизм, объяснимость, эпистемическое по сравнению с социальным доверием, компенсаторное доверие, U-образная динамика опыта.

Теоретическая и/или практическая значимость. Предложено интегративное определение доверия к ИИ, объединяющее три компонента (когнитивный, аффективный, поведенческий), четыре уровня факторов (индивид, система, контекст, общество) и направленное на обоснованное доверие. Выявлены критические пробелы в исследованиях и приоритетные направления. Результаты значимы для разработки надёжных ИИ-систем, эффективных регуляторных рамок и программ повышения ИИ-грамотности.

Ключевые слова: доверие к искусственному интеллекту, обоснованное доверие, антропоморфизм, смещение автоматизации, неприятие алгоритмов, объяснимость ИИ

Для цитирования: Отчерцова В. О., Блинова Е. Е. Психологические механизмы формирования доверия к искусственному интеллекту: критический анализ определений и интегративная модель // Психологические науки. 2026. № 3. С. 58–79. <https://doi.org/10.18384/3033-6414-2026-3-58-79>

Original research article

PSYCHOLOGICAL MECHANISMS OF FORMING TRUST IN GENERATIVE AI: A CRITICAL REVIEW OF DEFINITIONS AND AN INTEGRATIVE MODEL

V. Otchertsova^{1,2*}, E. Blinova¹

¹*Kherson State Pedagogical University, Kherson, Russian Federation*

²*Melitopol State University, Melitopol, Russian Federation*

**Corresponding author, e-mail: votchertsova@gmail.com*

Received by the editorial office 16.12.2025

Revised by the author 17.12.2025

Accepted for publication 23.12.2025

Abstract

Aim. To conduct a critical analysis of existing definitions of trust in generative AI, identify key contradictions in the conceptualization of the phenomenon and propose an integrative definition that considers the specifics of human interaction with intelligent agents.

Methodology. A systematic critical review of trust in AI from sources in 2020–2025 has been conducted including large-scale empirical studies, systematic reviews, theoretical works. The analysis was carried out according to the criteria: structural components, antecedents, typology, levels of analysis, cultural specificity. The method of theoretical research included the selection of sources according to the criteria of empirical validity, theoretical reflexivity and relevance, systematization of definitions, comparative analysis of key dimensions, identification of contradictions and synthesis of integrative definitions.

Results. There is a lack of consensus in definitions: trust in generative AI is conceptualised as a psychological attitude, a multilevel social phenomenon, or reduced to technical characteristics. Ten key paradoxes have been discovered: transparency, experience, anthropomorphism, prejudice towards automation versus rejection of algorithms, cultural and others. Trust in AI differs from interpersonal trust in seven unique ways: opacity, anthropomorphism, explainability, epistemic versus social trust, compensatory trust, and the U-shaped dynamics of experience.

Research implications. An integrative review of trust in AI is proposed. It combines three components (cognitive, affective, behavioural), four levels of factors (individual, system, context, society) and it is aimed at reasonable trust. Critical gaps in research and priority areas have been identified. The results are significant for the development of reliable AI systems, effective regulatory frameworks, and AI literacy programs.

Keywords: trust in artificial intelligence, reasonable trust, anthropomorphism, automation bias, algorithm aversion, explainability of AI

For citation: Otchertsova, V. O. & Blinova, E. E. (2026). Psychological Mechanisms of Forming Trust in Generative AI: A Critical Review of Definitions and an Integrative Model. In: *Psychological Sciences*, 3, 58–79. <https://doi.org/10.18384/3033-6414-2026-3-58-79>

ВВЕДЕНИЕ

Стремительное внедрение систем искусственного интеллекта (ИИ) в различные сферы человеческой деятельности – от здравоохранения и образования до финансов и юстиции – актуализирует вопрос о психологических механизмах формирования доверия к этим технологиям. Согласно глобальному исследованию Н. Гильмпай и соавторов, охватившему 48000 респондентов из 47 стран, 66% людей регулярно используют ИИ, однако лишь 46% готовы ему доверять, причём доверие снизилось по сравнению с 2022 г. [1]. Этот парадокс использования без доверия указывает на сложность и многомерность феномена доверия к ИИ, требующего глубокого научного осмысления.

Проблема доверия к ИИ не является чисто технической: она лежит на пересечении психологии, социологии, информатики и этики. Как справедливо отмечают исследователи, доверие выступает критическим фактором принятия технологий [2], определяет интенсивность использования ИИ-систем [3] и влияет на готовность делегировать им решения в условиях неопределённости [4]. При этом недостаточное доверие может препятствовать внедрению полезных ИИ-решений (неприятие алгоритмов), тогда как чрезмерное доверие создаёт риски некритического принятия ошибочных рекомендаций (смещение автоматизации) [5]. Достижение оптимального, обоснованного доверия (*appropriate reliance*) – ключевая задача как для разработчиков ИИ-систем, так и для психологической науки.

Однако анализ научной литературы выявляет фундаментальную проблему: отсутствие консенсуса в понимании самого феномена доверия к ИИ. Библиометрический обзор М. Бэн и соавторов, охватывающий 1156 основных публикаций за 24 года исследований, констатирует, что «нет единого определения доверия к ИИ; каждое исследование операционализирует по-своему» [6, с. 1654].

Эта концептуальная неопределённость затрудняет сравнение результатов исследований, препятствует построению кумулятивного знания и осложняет разработку практических рекомендаций. Более того, доверие к ИИ часто смешивается со смежными, но не тождественными понятиями: доверием к автоматизации (*trust in automation*), цифровым доверием (*digital trust*), надёжностью ИИ (*AI trustworthiness*) и технологическим доверием (*technology trust*). При этом, как подчёркивают Ю. Ли и соавторы, доверие к ИИ обладает специфическими характеристиками, отличающими его от межличностного доверия и доверия к традиционным автоматизированным системам [7].

Анализ научной литературы по проблеме доверия к ИИ показывает стремительный рост исследовательского интереса к данной тематике. Систематический обзор Ю. Дэнг и Дж. Ли 562 эмпирических исследований выявляет четыре ключевые группы antecedентов доверия: способности ИИ-системы, антропоморфизм, объяснимость и индивидуальные факторы пользователей [8]. Крупномасштабные глобальные исследования демонстрируют значительные кросс-культурные различия в доверии к ИИ: в коллективистских культурах (Китай) доверие составляет 72%, тогда как в индивидуалистских (США) – лишь 32%¹, что указывает на критическую роль культурного контекста в формировании доверия. Исследования динамики доверия выявляют нелинейную U-образную зависимость от опыта взаимодействия, аналогичную эффекту Даннинга-Крюгера: минимальный опыт ассоциируется с неприятием алгоритмов, средний – со смещением автоматизации, и только высокий опыт ведёт к обоснованному доверию [5].

Вместе с тем, несмотря на растущее количество эмпирических исследований, теоретическая концептуализация дове-

¹ Edelman Trust Barometer 2025: Artificial Intelligence. URL: <https://www.edelman.com/trust/2025/trust-barometer> (дата обращения: 28.01.2026).

рия к ИИ остаётся фрагментированной. В научном дискурсе сосуществуют множественные, зачастую противоречивые определения и операционализации данного конструкта. Психологические подходы акцентируют когнитивный, аффективный и поведенческий компоненты доверия как установки¹; социологические – многоуровневую структуру, различающую доверие к технологии, организации, институту и обществу [9]; технические – объективные характеристики надёжности, прозрачности и справедливости ИИ-систем². Междисциплинарные подходы стремятся интегрировать различные аспекты, однако рискуют потерять операциональную ясность [10]. Критически, большинство исследований проводится на WEIRD-выборках (Western – западные, Educated – образованные, Industrialized – индустриализированные, Rich – богатые, Democratic – демократические), что ограничивает генерализуемость результатов [8].

Особую проблему представляют парадоксы доверия к ИИ, которые остаются недостаточно осмысленными в научной литературе. К примеру, парадокс прозрачности: объяснимость признаётся критическим фактором доверия (40% респондентов определяют её как ключевой риск), однако лишь 17% организаций активно работают над её обеспечением³. Парадокс антропоморфизма: приписывание ИИ человеческих качеств усиливает аффективное доверие, но одновременно снижает эпистемическую бдительность

через «эвристику уверенности» [11]. Парадокс компенсаторного доверия: недоверие к людям и социальным институтам может парадоксально усиливать доверие к ИИ, воспринимаемому как свободному от человеческих предвзятостей⁴. Эти и другие парадоксы указывают на то, что доверие к ИИ – это не просто адаптация межличностного доверия к технологическому контексту, а качественно новый феномен, требующий собственной концептуализации.

Таким образом, актуальность настоящего исследования определяется, во-первых, практической значимостью доверия к ИИ для успешного внедрения интеллектуальных систем в различные сферы жизни; во-вторых, отсутствием консенсуса в научном сообществе относительно концептуализации данного феномена; в-третьих, необходимостью выявления специфических характеристик доверия к ИИ, отличающих его от традиционных форм доверия; в-четвёртых, потребностью в интегративном подходе, объединяющем психологические, социологические и технические аспекты.

Цель данного исследования – провести критический анализ существующих определений доверия к искусственному интеллекту, выявить ключевые измерения, противоречия и пробелы в концептуализации феномена, а также разработать интегративное определение, учитывающее специфику взаимодействия человека с интеллектуальными агентами.

Для достижения цели решаются следующие задачи:

– систематизировать определения доверия к ИИ из русскоязычных и англоязычных источников и провести их сравнительный анализ по ключевым измерениям (структура, antecedentes, типология, уровни анализа);

¹ Shang R., Hsieh G., Shah C. Trusting Your AI Agent Emotionally and Cognitively: Development and Validation of a Semantic Differential Scale for Affective and Cognitive Trust // arXiv preprint. URL: <https://arxiv.org/abs/2408.05354> (дата обращения: 28.01.2026).

² NIST. AI Risk Management Framework: Trustworthy AI Characteristics. 2023. URL: <https://airc.nist.gov/airmf-resources/airmf/3-sec-characteristics> (дата обращения: 28.01.2026).

³ McKinsey & Company. Building AI trust: The key role of explainability // QuantumBlack. URL: <https://www.mckinsey.com> (дата обращения: 02.02.2026).

⁴ Galindez-Acosta J. Giraldo Huertas J. Trust in AI emerges from distrust in humans: A machine learning study on decision-making guidance. // arXiv preprint. URL: <https://arxiv.org/abs/2511.16769> (дата обращения: 28.01.2026). DOI: 10.48550/arXiv.2511.16769.

– выявить парадоксы и противоречия в концептуализации доверия к ИИ, препятствующие построению единой теоретической рамки;

– определить специфические черты доверия к ИИ, отличающие его от межличностного доверия и доверия к традиционной автоматизации;

– предложить интегративное определение доверия к ИИ, объединяющее психологические, социологические и технические аспекты;

– обосновать направления будущих исследований, необходимые для преодоления выявленных пробелов.

Методология теоретического исследования. Настоящее исследование выполнено в рамках метода систематического критического анализа научной литературы. Процедура включала пять этапов. На первом этапе осуществлен отбор источников по следующим критериям: эмпирическая обоснованность (наличие данных или систематического обзора данных); теоретическая рефлексивность (явная концептуализация доверия); репрезентативность подходов (психологические, социологические, технические, междисциплинарные); актуальность (публикации последних пяти лет, с акцентом на 2023-2025 гг.). В выборку вошли тридцать определений доверия к ИИ: пятнадцать из англоязычных источников (систематические обзоры, крупномасштабные эмпирические исследования, теоретические работы) и пятнадцать из русскоязычных (работы по цифровому доверию, философско-методологический анализ, экспериментальные исследования). На втором этапе определения систематизированы в аналитические таблицы с указанием ключевых компонентов, antecedентов и консеквенций. На третьем этапе проведён сравнительный анализ определений по следующим измерениям: структурные компоненты доверия (когнитивный, аффективный, поведенческий); antecedенты (предикторы) доверия; типология доверия (начальное (initial), непрерывное

(continuous), эпистемическое, социальное и др.); уровни анализа (технология, организация, институт, общество); парадоксы и противоречия; культурная специфичность концептуализаций. На четвёртом этапе выявлены противоречия и пробелы в существующих подходах. На пятом этапе осуществлен синтез интегративного определения на основе критического анализа.

Далее в первом разделе представлен критический анализ определений доверия к ИИ по основным теоретическим подходам (психологический, социологический, технический, междисциплинарный). Во втором разделе проанализированы ключевые измерения и выявлены десять парадоксов доверия к ИИ. В третьем разделе определены специфические черты, отличающие доверие к ИИ от межличностного доверия. В четвёртом разделе предложено интегративное определение и представлена процессуальная модель формирования доверия к ИИ. В заключении сформулированы основные выводы, выявлены критические пробелы в исследованиях и обоснованы направления будущих работ.

КРИТИЧЕСКИЙ АНАЛИЗ ОПРЕДЕЛЕНИЙ ДОВЕРИЯ К ИСКУССТВЕННОМУ ИНТЕЛЛЕКТУ

Значительная часть исследователей концептуализирует доверие к ИИ как психологическую установку с трёхкомпонентной структурой, восходящей к классической модели установки в социальной психологии и развивающей применительно к доверию модели А. Б. Купрейченко [12]. В рамках этого подхода выделяются когнитивный, аффективный и поведенческий компоненты.

Когнитивный компонент представляет собой оценку способностей, компетентности и надёжности ИИ-системы. Систематический обзор исследований межличностного доверия, доверия к автоматизации и доверия к ИИ, проведённый Ю. Ли и соавторами, выделяет

три измерения факторов, влияющих на формирование доверия: характеристики субъекта доверия (trustor), характеристики объекта доверия (trustee) и контекст их взаимодействия. При этом ключевыми параметрами восприятия ИИ-системы как объекта доверия выступают компетентность (competence) и теплота (warmth), отражающие соответственно оценку способностей системы и её воспринимаемых намерений [7].

В более широком контексте, систематический обзор прогресса и вызовов исследований доверия к ИИ подчёркивает: «Доверие к ИИ – это психологическое состояние, характеризующееся готовностью принять уязвимость на основе позитивных ожиданий от намерений или поведения ИИ-системы» [3], что восходит к классическому определению доверия, предложенному Д. Руссо и соавторами. Центральным в этих определениях является когнитивная оценка надёжности и предсказуемости поведения системы.

Аффективный компонент доверия к ИИ – эмоциональное отношение к системе, чувство комфорта или тревоги при взаимодействии – длительное время недооценивался в исследованиях, фокусировавшихся преимущественно на когнитивных аспектах. Однако экспериментальные исследования демонстрируют, что эмоции играют критическую роль в формировании доверия к цифровым системам, причём эффект асимметричен: негативные эмоции значительно сильнее снижают доверие, чем позитивные эмоции способны его повысить – феномен, известный в психологии как «эффект негативности» [13]. Это указывает на то, что аффективный компонент может оказывать более сильное влияние на доверие, чем рациональная когнитивная оценка способностей системы.

Поведенческий компонент проявляется в готовности полагаться на ИИ и делегировать ему решения. Глобальное исследование Н. Гильмпай и соавторов, охватившее более 48000 респондентов из

47 стран, операционализирует доверие к ИИ через готовность использовать ИИ-системы и полагаться на них в различных контекстах повседневной жизни и работы. Исследование показало, что несмотря на широкое распространение ИИ (66% респондентов регулярно используют ИИ), только 46% готовы доверять ИИ-системам [5].

Экспериментальные исследования делегирования принятия решений ИИ-системам также акцентируют поведенческий аспект доверия. К. Кандриан и А. Шерер продемонстрировали, что люди предпочитают делегировать решения ИИ по сравнению с человеческими агентами, особенно когда решения связаны с потенциальными потерями, что отражает готовность полагаться на ИИ на основе воспринимаемых преимуществ его объективности и рациональности [14].

Критический анализ психологических определений выявляет ряд проблем. Во-первых, не всегда ясно, можно ли приписывать ИИ «намерения» (intentions). В отличие от людей, ИИ-системы не обладают сознанием и намерениями в полном смысле слова, хотя пользователи могут приписывать им таковые в процессе антропоморфизации. Во-вторых, проблематичным остаётся соотношение доверия и полагания: А. Кюпер и Н. Крамер подчёркивают важность разграничения этих конструктов, поскольку доверие предсказывает обоснованное доверие, но не тождественно ему [15]. Человек может полагаться на ИИ без доверия (прагматическое использование) или, напротив, доверять, но не использовать систему в силу внешних ограничений. В-третьих, трёхкомпонентная модель, хорошо описывающая межличностное доверие, требует адаптации к специфике ИИ, поскольку взаимодействие с машиной качественно отличается от взаимодействия с человеком.

Ряд исследователей акцентирует многоуровневую природу доверия к ИИ, различая доверие к конкретной технологии,

к организациям-разработчикам, к институциональным регуляторам и к обществу в целом. Этот подход развивает идею о том, что доверие формируется не в вакууме, а в сложной социальной экосистеме, где технологические, организационные, институциональные и культурные факторы взаимодействуют.

Глобальное исследование доверия к искусственному интеллекту, проведённое KPMG совместно с Университетом Квинсленда в 2023 г. (17000 респондентов из 17 стран), выявило, что формирование доверия к ИИ происходит по четырём основным путям: институциональному (наличие гарантий, регулирования и законов), мотивационному (воспринимаемые преимущества использования ИИ), снижению неопределённости (решение проблем и рисков, связанных с ИИ) и просветительскому (понимание людьми использования ИИ и цифровых технологий)¹. Это отражает понимание того, что доверие к ИИ определяется не только характеристиками самой технологии, но и социальным контекстом её создания и применения – технология не нейтральна, а встроена в социальные отношения.

А. Г. Санина и коллеги развивают эту идею, концептуализируя цифровое доверие (включающее доверие к ИИ) как многокомпонентный феномен, объединяющий доверие к техническим системам, государственным институтам и организациям, а также обобщённое межличностное доверие [9]. Важно, что в этом подходе доверие к технологии неотделимо от институционального доверия: доверие к техническим системам выступает медиатором между доверием к конкретным технологиям (включая ИИ) и доверием к управлению на основе данных в целом [9]. Иными словами, пользователи доверяют ИИ не «вообще», а в той мере, в какой они доверяют организациям, соз-

дающим и управляющим этими системами, а также регуляторам, устанавливающим правила их использования.

Р. Лукьяненко и соавторы предлагают наиболее амбициозную системную концептуализацию, определяя доверие к ИИ как «проблему взаимодействия систем, требующую междисциплинарного подхода на основе системного мышления и общей теории систем» [10, с. 1995]. Этот подход подчёркивает, что доверие возникает не как свойство отдельного элемента (технологии или пользователя), а как эмерджентное свойство системы взаимодействий, включающей технологические артефакты, человеческих акторов, организации, институты и культурные нормы.

Критический анализ социологических определений выявляет их сильные и слабые стороны. С одной стороны, многоуровневые модели, безусловно, отражают сложность феномена и позволяют учитывать макросоциальные факторы, игнорируемые чисто психологическими подходами. С другой стороны, они усложняют операционализацию и измерение: эмпирические исследования часто не разграничивают уровни, измеряя «доверие к ИИ» обобщённо, что затрудняет сравнение результатов. Кроме того, связь между уровнями остаётся недостаточно исследованной: например, влияет ли институциональное доверие к регуляторам на доверие к конкретной ИИ-системе, и если да, то каковы механизмы этого влияния?

В технической литературе по разработке ИИ-систем доверие часто редуцируется к объективным характеристикам системы: надёжности (reliability), прозрачности (transparency), объяснимости (explainability) и справедливости (fairness). Этот подход отражает инженерную парадигму, согласно которой доверие к ИИ может быть «спроектировано» через обеспечение определённых технических свойств системы.

Концепция «доверенного ИИ» интегрирует технические, этические и со-

¹ KPMG & University of Queensland. Trust in artificial intelligence: A global study. URL: <https://kpmg.com/xx/en/our-insights/ai-and-technology/trust-in-artificial-intelligence.html> (дата обращения: 28.01.2026).

циальные требования к ИИ-системам. Согласно руководству Европейской комиссии, доверенный ИИ должен быть законным, этичным и надёжным, что операционализируется через требования прозрачности, справедливости, подотчётности и технической надёжности¹. Согласно Национальной стратегии развития ИИ на период до 2030 г.: «Доверенный ИИ – это технологическое решение, соответствующее стандартам безопасности, разработанное с учётом объективности, непредвзятости, этичности и исключения применения для нанесения угрозы гражданам и государству»². Здесь смешиваются два различных понятия: «заслуживающий доверия, надёжный» и «доверие как психологическое состояние». Надёжная система может не вызывать доверия, если пользователи не осведомлены о её характеристиках; ненадёжная система может пользоваться доверием в силу успешного маркетинга или антропоморфного дизайна.

Исследование компании McKinsey подчёркивает критическую роль объяснимости: «40% респондентов определяют объяснимость как ключевой риск при внедрении генеративного ИИ, но только 17% организаций активно работают над ней»³. Это выявляет парадокс прозрачности, о котором речь пойдёт далее. Исследование факторов восприятия ИИ-систем с применением метода совместного анализа демонстрирует гетероген-

ность предпочтений пользователей: они формируют различные кластеры, где для одних приоритетна безопасность и конфиденциальность, для других – справедливость и объяснимость, для третьих – сертификация и подотчётность. При этом прозрачность выступает наиболее сильным предиктором доверия к алгоритмам [16].

Критический анализ технических определений выявляет их принципиальную ограниченность: редукция доверия к техническим характеристикам игнорирует субъективный психологический компонент. Доверие – это не свойство системы, а отношение пользователя к системе, формирующееся под влиянием множества факторов, лишь часть которых связана с объективными характеристиками ИИ. Как показывают исследования динамики доверия, оно определяется не только надёжностью системы, но и опытом взаимодействия пользователя, образуя U-образную кривую: минимальный опыт ассоциируется с неприятием алгоритмов, средний – со смещением алгоритмизации, высокий – с обоснованным доверием [1]. Более того, прозрачность может иметь обратный эффект, если раскрывает нежелательные аспекты работы алгоритма, противоречащие ценностям пользователей [17].

Наиболее перспективные определения стремятся интегрировать психологические, социологические и технические аспекты, признавая многомерность феномена доверия к ИИ.

Систематический обзор 562 эмпирических исследований, проведённый К. Данг и Г. Ли, концептуализирует доверие к ИИ через antecedentes и consecuencias: ключевыми antecedentes выступают способности ИИ, антропоморфизм, объяснимость и индивидуальные факторы пользователя, а основными consecuencias – поведенческие намерения, отношение к ИИ и его принятие [8]. Это процессуальная модель, где доверие выступает медиатором между характеристиками

¹ European Commission. Ethics Guidelines for Trustworthy AI. 2019. URL: <https://digital-strategy.ec.europa.eu/en/library> (дата обращения: 28.01.2026).

² Указ Президента Российской Федерации от 10.10.2019 № 490 «О развитии искусственного интеллекта в Российской Федерации» (вместе с «Национальной стратегией развития искусственного интеллекта на период до 2030 года») // Собрание законодательства РФ. 14.10.2019. № 41. Ст. 5700. URL: <http://www.kremlin.ru/acts/bank/44731> (дата обращения: 20.01.2026).

³ McKinsey & Company. Building AI trust: The key role of explainability // QuantumBlack. 2024. November 25. URL: <https://www.mckinsey.com/capabilities/quantumblack/our-insights/building-ai-trust-the-key-role-of-explainability> (дата обращения: 28.01.2026).

системы/пользователя и поведенческими исходами. Преимущество такого подхода – операционализуемость: доверие измеряется не напрямую, а через его предикторы и последствия.

Исследование культурных различий в восприятии ИИ вносит важное дополнение: «Доверие к ИИ культурно детерминировано: индивидуалисты воспринимают ИИ как внешнюю угрозу уникальности, коллективисты – как расширение себя» [18]. Это подчёркивает, что концептуализация доверия не может быть универсальной: в различных культурах доверие формируется по различным механизмам и имеет различное содержание. Данные глобального исследования Edelman Trust Barometer 2025 подтверждают это: в Китае 72% респондентов доверяют ИИ, в США – лишь 32%, что невозможно объяснить только различиями в опыте использования технологий¹.

Философско-методологический анализ Т. Г. Лешкевич выявляет парадоксальную природу доверия к ИИ: оно основано на противоречии между необходимостью полагаться на технически непрозрачные системы и потребностью в понимании механизмов их работы [19]. Существование этого парадокса сопровождается «малопонятным компромиссом, суть которого в игнорировании всей полноты рисков в пользу удобства пользования быстрыми интеллектуальными системами» [19, с. 42]. Это определение концептуализирует уникальную дилемму ИИ: в отличие от традиционных инструментов, многие современные ИИ-системы (особенно глубокого обучения) являются эпистемологически непрозрачными – даже разработчики не могут полностью объяснить, как система пришла к решению. Таким образом, доверие к ИИ требует принятия фундаментальной неопределённости, что качественно отличает

его от межличностного доверия, где всегда существует возможность запросить объяснение у другого человека.

Критический анализ междисциплинарных определений показывает, что, хотя они и более полны, однако рискуют стать слишком широкими и потерять операциональную ясность. Необходим баланс между теоретической сложностью, отражающей многомерность феномена, и эмпирической измеримостью, позволяющей проводить систематические исследования.

Таким образом, проведённый анализ определений доверия к ИИ выявил четыре основных теоретических подхода (психологический, социологический, технический, междисциплинарный), каждый из которых акцентирует различные аспекты феномена. Психологические определения фокусируются на трёхкомпонентной структуре доверия как установки, но сталкиваются с проблемой адаптации концепции намерений к нечеловеческим агентам. Социологические определения подчёркивают многоуровневость доверия, но затрудняют операционализацию. Технические определения редуцируют доверие к объективным характеристикам, игнорируя субъективный психологический компонент. Междисциплинарные определения стремятся к интеграции, но рискуют потерять ясность. Критически ни один из подходов не является исчерпывающим, что указывает на необходимость синтеза, учитывающего сильные стороны каждого и преодолевающего их ограничения.

СПЕЦИФИКА ДОВЕРИЯ К ИИ

Анализ определений и эмпирических исследований позволяет выделить семь специфических черт, отличающих доверие к ИИ от межличностного доверия и доверия к традиционной автоматизации.

Эпистемологическая непрозрачность (black box problem). В отличие от традиционных инструментов, работу которых можно понять и объяснить, и от людей, у

¹ Edelman Trust Barometer 2025: Artificial Intelligence. Edelman. URL: <https://www.edelman.com/trust/2025/trust-barometer> (дата обращения: 28.01.2026).

которых всегда можно запросить объяснение решения, многие современные ИИ-системы (особенно глубокого обучения) являются эпистемологически непрозрачными: даже разработчики не могут полностью объяснить, как система пришла к конкретному решению.

Исследования компании McKinsey¹ и фармацевтической компании Sanofi² описывают фундаментальный компромисс между производительностью и объяснимостью: организации всё чаще сталкиваются с дилеммой выбора между более точными, но сложными моделями (глубокие нейронные сети, большие языковые модели) и менее точными, но прозрачными моделями (деревья принятия решений, линейная регрессия). McKinsey отмечает, что «организации всё чаще сталкиваются с компромиссом между точностью модели и интерпретируемостью: более сложные модели часто достигают более высокой производительности, но за счёт снижения прозрачности»³. Sanofi подчёркивает: «чем более сложной является ИИ-система, тем сложнее объяснить, как она работает»⁴.

Критически, в межличностном доверии такая дилемма не возникает: мы можем попросить человека объяснить его решение, и даже если объяснение неполно или рационализировано *post hoc*, сам факт возможности диалога снижает

когнитивную неопределённость. В доверии к ИИ такой возможности часто нет, что требует принятия фундаментальной неопределённости и качественно меняет природу доверия.

Антропоморфизм – приписывание ИИ человеческих качеств (намерений, эмоций, агентности, моральной ответственности) – феномен, уникальный для взаимодействия с интеллектуальными системами. В межличностном доверии антропоморфизация не требуется, поскольку объект доверия уже является человеком. В доверии к традиционной автоматизации (например, автопилоту самолета) антропоморфизация минимальна, поскольку система не демонстрирует поведения, имитирующего интеллект.

Исследование доверия к генеративным ИИ-тьюторам показывает, что антропоморфизм усиливает аффективное доверие (эмпатию, ощущение социального присутствия), но одновременно снижает эпистемическую бдительность через «эвристику уверенности» [11, с. 7]. Пользователи, взаимодействующие с антропоморфным ИИ (использующим местоимение «я», синтез речи, эмодзи), менее критично оценивают его выводы, допуская больше ошибок в задачах, требующих проверки информации. Это создаёт риск чрезмерного доверия и смещения автоматизации.

Исследование влияния антропоморфизма на доверие выявляет нелинейную зависимость: умеренная антропоморфизация (человекоподобное имя, местоимение «я», дружелюбный тон) повышает доверие; избыточная (гиперреалистичная аватарка, имитация эмоций) может вызывать эффект «зловещей долины» (*uncanny valley*) и снижение доверия [11]. Более того, эффект модулируется установками: люди с негативными установками к ИИ реагируют на антропоморфизм снижением доверия, воспринимая его как манипуляцию.

Таким образом, антропоморфизм – двойственный фактор доверия к ИИ, не

¹ Building AI trust: The key role of explainability. McKinsey & Company. QuantumBlack. 2024. URL: <https://www.mckinsey.com/capabilities/quantum-black/our-insights/building-ai-trust-the-key-role-of-explainability> (дата обращения: 28.01.2026).

² All in on AI, Transparent & Explainable. Sanofi Magazine. URL: <https://www.sanofi.com/en/magazine/our-science/ai-transparent-and-explainable> (дата обращения: 28.01.2026).

³ Building AI trust: The key role of explainability. McKinsey & Company. URL: <https://www.mckinsey.com/capabilities/quantumblack/our-insights/building-ai-trust-the-key-role-of-explainability> (дата обращения: 10.10.2025).

⁴ All in on AI, Transparent & Explainable. Sanofi Magazine. URL: <https://www.sanofi.com/en/magazine/ai-in-healthcare/ai-transparent-and-explainable> (дата обращения: 10.10.2025).

имеющий аналога в межличностном доверии: он может как усиливать, так и снижать доверие в зависимости от степени, контекста и характеристик пользователя.

Смещение автоматизации и неприятие алгоритмов как уникальные феномены. Смещение автоматизации (тенденция чрезмерно полагаться на алгоритмические рекомендации, даже когда они ошибочны), и неприятие алгоритмов (отказ от алгоритма после первой ошибки, даже если он в среднем точнее человека) – феномены, специфичные для доверия к алгоритмам и ИИ [11]. В межличностном доверии аналоги отсутствуют или слабо выражены.

Исследование Б. Хосе и А. Томас показывают, что люди применяют асимметричные стандарты к ошибкам ИИ и людей: от алгоритмов ожидают совершенства, от людей допускают ошибки [11]. Одна ошибка ИИ может разрушить доверие, построенное долгим позитивным опытом, тогда как к ошибкам человека-эксперта относятся более терпимо, объясняя их объективной сложностью задачи или случайностью. Это указывает на то, что доверие к ИИ более хрупкое и менее устойчивое к нарушениям, чем межличностное доверие.

Контролируемость ИИ выступает ключевым модератором: если пользователь может вмешаться и скорректировать решение системы (human-in-the-loop), неприятие алгоритмов снижается [11]. Это имеет важные следствия для дизайна ИИ-систем: полная автоматизация, лишаящая пользователя контроля, может парадоксально снижать доверие и принятие системы.

Объяснимость (ХАИ) как уникальное требование. Требование объяснимости (explainability) специфично для ИИ. В межличностном доверии мы редко требуем полного объяснения того, как человек пришёл к решению: достаточно общего понимания его компетентности и намерений. В доверии к традиционной автоматизации работа системы обычно

прозрачна (if-then правила, явные алгоритмы). ИИ находится между этими полюсами: достаточно сложен, чтобы требовать объяснения, но достаточно непрозрачен, чтобы его давать было сложно или невозможно.

Исследования показывают, что объяснимость критична для калибровки доверия: без объяснений люди либо чрезмерно доверяют, либо не доверяют [9; 15]. Однако исследование ключевых факторов восприятия ИИ выявляет парадокс: объяснимость не всегда усиливает доверие – зависит от того, насколько объяснение соответствует ожиданиям пользователя [11]. Если объяснение раскрывает использование переменных, которые пользователь считает неэтичными, или логику, противоречащую его ценностям, объяснимость может снизить доверие.

Более того, эффективность объяснений зависит от ИИ-грамотности пользователей: технические объяснения (веса признаков, значения SHAP) понятны экспертам, но бесполезны для обычных пользователей. Необходимы адаптивные объяснения, учитывающие уровень компетентности и цели пользователя.

Исследование формирования доверия у детей и взрослых вводит важное различие, специфичное для ИИ: *эпистемическое доверие (к знаниям и информации ИИ) в сравнении с социальным доверием (к ИИ как агенту взаимодействия)* [20]. В межличностном доверии эти аспекты обычно не разделяются: доверяя человеку как партнёру (социально), мы склонны доверять и его информации (эпистемически), и наоборот.

Для ИИ это различие критично. Дети больше доверяют ИИ эпистемически (верят, что ИИ даст правильный ответ на фактуальный вопрос), чем социально: они предпочитают играть с людьми и обращаться к людям за личной информацией, в то время как для фактуальной информации они всё чаще выбирают технологические источники. Взрослые демонстрируют ещё более выраженное

разделение: они социально предпочитают людей и сопротивляются социальному давлению со стороны роботов, но при этом показывают тенденцию к эпистемическому чрезмерному доверию машинам, предпочитая технологические источники информации человеческим экспертам. Это объясняет парадокс: человек может полагаться на ИИ для получения информации (эпистемическое доверие), но не воспринимать его как партнёра или друга (отсутствие социального доверия).

Исследование В. Го и соавторов о сравнении доверия к врачу-человеку и ИИ-диагносту подтверждает это: в высокочувствительных задачах (медицинская диагностика) люди всё ещё предпочитают доверять людям, несмотря на осведомлённость о статистической точности ИИ¹. Возможное объяснение: медицина – это не только эпистемическая (точность диагноза), но и социальная практика (эмпатия, эмоциональная поддержка, холистическое понимание пациента как личности). ИИ может превзойти людей эпистемически, но не социально.

Уникальный механизм, не наблюдаемый в традиционном межличностном доверии: *доверие к ИИ может возникать из недоверия к людям*. Недавнее исследование Х. С. Галиндес-Акоста и Х. Хиральдо Уэртас вводит концепцию «отложенного доверия» (deferred trust) – компенсаторный когнитивный механизм, при котором недоверие к человеческим агентам, вызванное воспринимаемой предвзятостью, ненадёжностью или контекстуальными неудачами, перенаправляет эпистемическое полагание на ИИ-системы, воспринимаемые как более нейтральные или компетентные². Эмпирические данные

показывают, что более низкое предварительное доверие к людям (священникам, сверстникам, взрослым) систематически предсказывало более высокий выбор ИИ.

Люди с низким институциональным доверием (к врачам, юристам, чиновникам, СМИ) демонстрируют более высокое доверие к ИИ в соответствующих доменах. ИИ воспринимается как «объективный», свободный от человеческой предвзятости, эмоциональности, коррупции и личных интересов. Этот феномен особенно выражен в фактуальных контекстах, где ИИ доминирует над человеческими информантами, в то время как в социальных и моральных сценариях люди продолжают предпочитать человеческих агентов³.

Это создаёт *парадокс компенсаторного доверия*: в обществах с низким социальным капиталом и институциональным доверием (например, в постсоветских странах) доверие к ИИ может быть парадоксально высоким, потому что ИИ рассматривается как альтернатива ненадёжным людям и институтам. Однако это доверие может быть иллюзорным, поскольку ИИ-системы разрабатываются людьми, обучаются на данных, отражающих человеческие предвзятости, и управляются организациями, которым люди не доверяют. Таким образом, компенсаторное доверие к ИИ может быть формой эскапизма, избегания проблемы низкого социального доверия, а не её решения.

Исследование М. С. Горовиц и соавторов показывает, что 9000 респондентов из девяти стран выявляет уникальную для ИИ *динамику доверия в зависимости от опыта, аналогичную эффекту Даннинга-Крюгера* [5]. Минимальный опыт ассоциируется с неприятием алгоритмов – недоверием из-за страха неизвестного. Средний опыт парадоксально ассоциируется с максимальным чрезмерным доверием (смещением автоматизации) – иллюзией понимания. Только

¹ Konya-Baumbach E., Krummrey G., Abdullahi S. Diagnosis by AI: Why trust is the key to acceptance // SocietyByte. 2025. 12 June. URL: <https://clck.ru/3Uvtn3> (дата обращения: 02.02.2026).

² Galindez-Acosta J., Giraldo Huertas J. Trust in AI emerges from distrust in humans: A machine learning study on decision-making guidance. URL: <https://arxiv.org/abs> (дата обращения: 28.01.2026).

³ Там же.

высокий опыт, сопровождающийся столкновением с ошибками ИИ и рефлексией ограничений технологии, ведёт к обоснованному доверию – реалистичной оценке способностей системы.

В межличностном доверии такая U-образная динамика не описана: обычно доверие либо растёт с опытом позитивного взаимодействия, либо снижается при негативном. Эффект Даннинга-Крюгера в доверии к ИИ указывает на то, что небольшое знание об ИИ опаснее полного незнания, поскольку создаёт иллюзию компетентности и ведёт к некритичному принятию рекомендаций системы.

Доверие к ИИ отличается от межличностного доверия семью уникальными чертами: эпистемологическая непрозрачность, антропоморфизм как двойственный фактор, смещение автоматизации и неприятие алгоритмов, объяснимость как уникальное требование, различие эпистемического и социального доверия, компенсаторное доверие (от недоверия к людям к доверию к ИИ), U-образная динамика опыта (эффект Даннинга-Крюгера). Эти черты указывают на то, что доверие к ИИ – это не просто адаптация межличностного доверия к технологическому контексту, а качественно новый феномен, требующий собственной концептуализации, теории и методологии исследования.

ИНТЕГРАТИВНОЕ ОПРЕДЕЛЕНИЕ И ПРОЦЕССУАЛЬНАЯ МОДЕЛЬ ДОВЕРИЯ К ИИ

На основе критического анализа определений, выявления парадоксов и определения специфических черт мы предлагаем следующее интегративное определение доверия к искусственному интеллекту.

Доверие к искусственному интеллекту – это динамическое, многоуровневое психологическое отношение (установка), формирующееся на пересечении: индивидуальных характеристик (склонность к доверию, ИИ-грамотность, опыт взаимодействия, личностные черты, демогра-

фические переменные); характеристик ИИ-системы (объективные: способности, надёжность, прозрачность; воспринимаемые: антропоморфизм, объяснимость, справедливость); контекста применения (домен, чувствительность задачи, воспринимаемый риск, социальные нормы); социокультурных факторов (культурные ценности, институциональное доверие, регуляторная среда).

Доверие к ИИ включает в себя три компонента:

- когнитивный компонент (оценка способностей, надёжности и прозрачности ИИ-системы в конкретном контексте применения);

- аффективный компонент (эмоциональное отношение к ИИ, включающее чувство комфорта/дискомфорта, эмпатию при антропоморфизации и тревогу/энтузиазм относительно взаимодействия);

- поведенческий компонент (готовность полагаться на ИИ и делегировать ему решения в условиях неопределённости и принятия уязвимости).

Специфические черты доверия к ИИ, отличающие его от межличностного доверия: эпистемологическая непрозрачность, антропоморфизм как двойственный фактор, смещение автоматизации и неприятие алгоритмов, объяснимость как уникальное требование, различие эпистемического и социального доверия, компенсаторное доверие (через недоверие к людям), U-образная динамика опыта (эффект Даннинга-Крюгера).

Предложенное интегративное определение обладает следующими достоинствами. Во-первых, оно интегрирует психологические (трёхкомпонентная структура установки), социологические (многоуровневость факторов) и технические (характеристики системы) аспекты, преодолевая односторонность существующих подходов. Во-вторых, оно признаёт динамичность доверия, изменяющегося во времени и в зависимости от опыта взаимодействия, в отличие от статиче-

ских концептуализаций. В-третьих, оно включает культурное измерение, критичное для кросс-культурных исследований и глобального внедрения ИИ. В-четвёртых, оно операционализируемо: каждый компонент может быть измерен существующими или разрабатываемыми инструментами. В-пятых, оно учитывает специфику ИИ, не редуцируя доверие к ИИ к межличностному или технологическому доверию.

На основе интегративного определения и систематического обзора 562 эмпирических исследований [8] мы предлагаем процессуальную модель, представленную на рисунке 1. Модель показывает, что доверие формируется на пересечении четырёх групп antecedентов: характеристик ИИ-системы (способности, надёжность, прозрачность, объяснимость, справедливость, антропоморфизм), индивидуальных факторов (склонность к доверию, ИИ-грамотность, опыт, личность, установки, демография), контекстуальных факторов (домен, чувствительность задачи, риск, социальные нормы, регуляторная среда, доверие к разработчику) и культурных ценностей (индивидуализм-коллективизм), выступающих модератором всей модели.

Сформированное доверие включает три компонента (когнитивный, аффективный, поведенческий) и влияет на три типа консеквенций: поведенческие (принятие ИИ, интенсивность использования, делегирование решений, следование рекомендациям, обоснованное доверие), психологические (удовлетворённость, воспринимаемая полезность, снижение тревоги, позитивные установки, эмоциональная привязанность, лояльность к бренду) и социальные (распространение ИИ, социальное принятие, требования регуляции, общественные дебаты, формирование норм, институционализация ИИ). Критически модель включает обратную связь: опыт использования (поведенческие консеквенции) формирует новый опыт взаимодействия (индивидуальный

фактор), влияя на последующее доверие и обеспечивая динамику калибровки.

На рисунке 2 представлена многоуровневая модель, интегрирующая четыре уровня анализа: технологический (конкретная ИИ-система с тремя компонентами доверия), организационный (работчики, провайдеры, корпоративная этика), институциональный (регуляторы, стандарты, сертификация, законодательство) и социальный (культурные ценности, социальные нормы, общественное мнение). Модель показывает, что доверие к ИИ не формируется изолированно на уровне технологии, но встроено в социальную экосистему, где технологические, организационные, институциональные и культурные факторы взаимодействуют.

ЗАКЛЮЧЕНИЕ

Настоящее исследование было направлено на критический анализ существующих определений доверия к искусственному интеллекту, на выявление ключевых противоречий и разработку интегративного определения. Систематический анализ определений из источников 2020–2025 гг. выявил четыре основных теоретических подхода – психологический, социологический, технический, междисциплинарный, – каждый из которых акцентирует различные аспекты феномена, но ни один не является исчерпывающим.

Проведённый анализ продемонстрировал отсутствие консенсуса в концептуализации: доверие к ИИ операционализируется по-разному в различных исследованиях, что затрудняет сравнение результатов и построение кумулятивного знания. Выявлено десять ключевых парадоксов доверия к ИИ: прозрачности (объяснимость признаётся критичной, но мало кто работает над ней), компетентности (надёжность не всегда главный предиктор), опыта (U-образная кривая Даннинга-Крюгера), антропоморфизма (усиливает доверие, снижает бдительность), чрезмерное доверие против недоверия при ошибке, культурный,

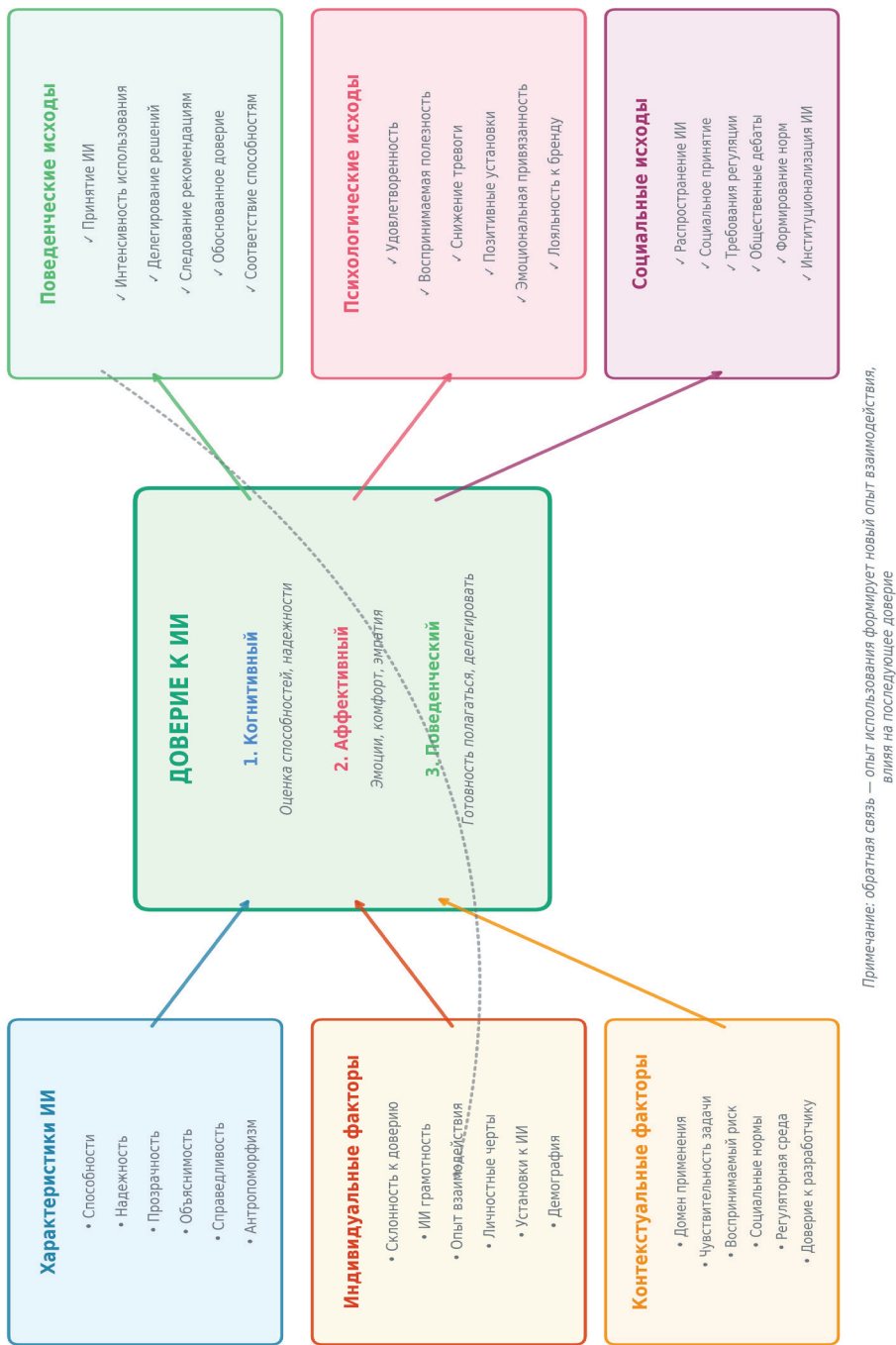


Рис. 1 / Fig. 1. Интегративная модель antecedents and consequences of trust in generative AI

Источник: данные авторов.

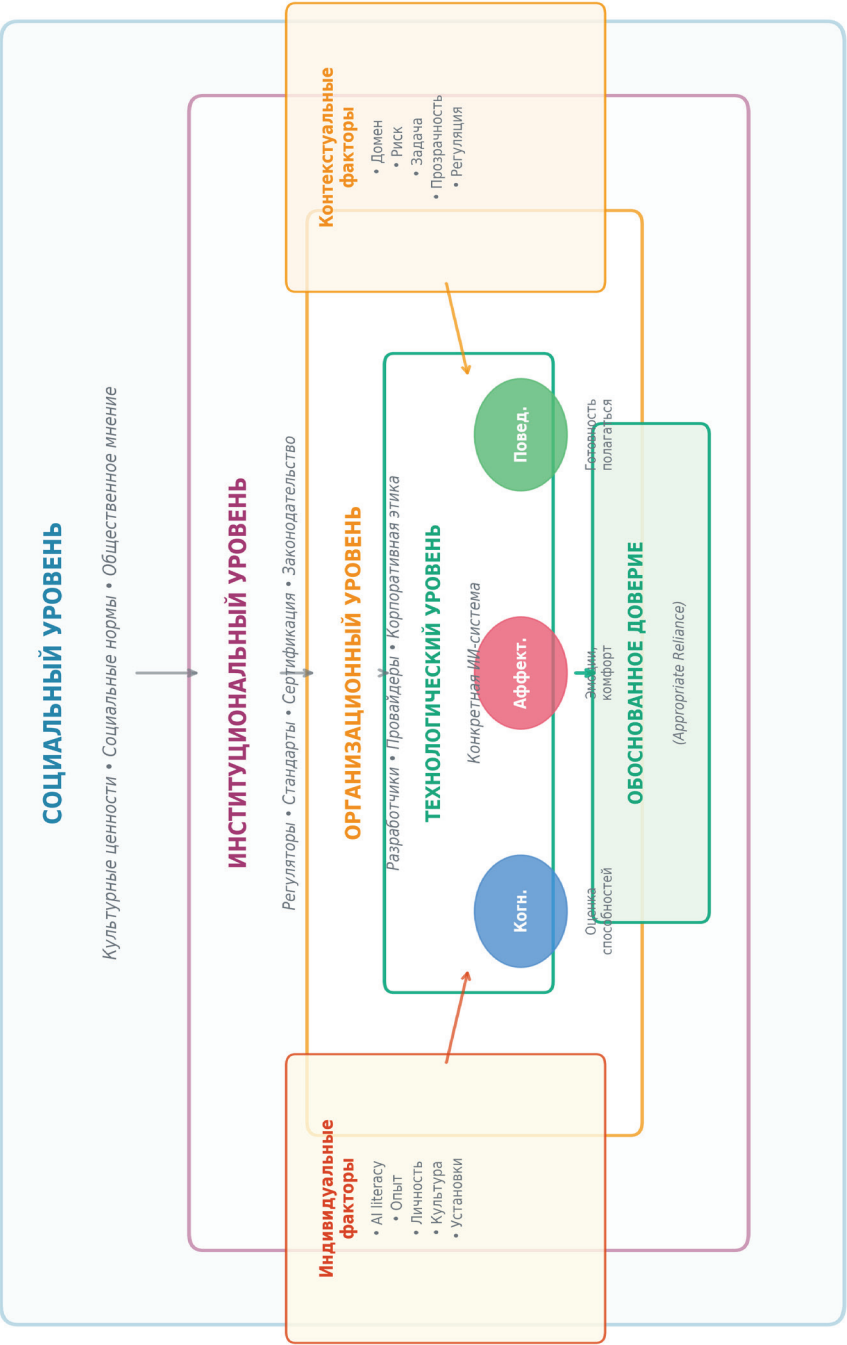


Рис. 2 / Fig. 2. Многоуровневая модель доверия к искусственному интеллекту / Multilevel model of trust in artificial intelligence

Источник: данные авторов.

эмоциональной асимметрии (негативные эмоции сильнее позитивных), компенсаторного доверия (недоверие к людям ведёт к доверию к ИИ), независимости (доверие к ИИ и к людям коррелируют слабо), сосуществования (доверие и недоверие как независимые конструкты).

Определены семь специфических черт, отличающих доверие к ИИ от межличностного доверия: эпистемологическая непрозрачность (*black box problem*), антропоморфизм как двойственный фактор, смещение автоматизации и неприятие алгоритмов как уникальные феномены, объяснимость как уникальное требование, различие эпистемического и социального доверия, компенсаторное доверие через недоверие к людям, U-образная динамика опыта (эффект Даннинга-Крюгера в контексте ИИ). Эти черты указывают на то, что доверие к ИИ – это не просто адаптация межличностного доверия к технологическому контексту, а качественно новый феномен.

На основе проведённого анализа предложено интегративное определение доверия к ИИ, объединяющее три компонента (когнитивный, аффективный, поведенческий), четыре уровня факторов (индивид, система, контекст, общество) и направленное на обоснованное доверие. Разработана процессуальная модель, показывающая формирование доверия на пересечении antecedентов, его проявление в трёх компонентах и влияние на три типа консеквенций (поведенческие, психологические, социальные) с обратной связью.

Систематический анализ выявил восемь критических пробелов, препятствующих развитию теории и практики доверия к ИИ.

Первый пробел – отсутствие консенсуса в определениях. Как показывает наш анализ, существующие определения операционализируют доверие к ИИ по-разному, что затрудняет сравнение результатов. Необходима разработка консенсусного определения через междуна-

родные консорциумы по аналогии с DSM в психиатрии. Наше интегративное определение может послужить отправной точкой для такой работы.

Второй пробел – смешение уровней анализа. Многие эмпирические исследования не разграничивают доверие к технологии, организации, институту и обществу, измеряя «доверие к ИИ» обобщённо. Необходима разработка многоуровневых инструментов измерения, явно операционализирующих эти уровни и учающих связи между ними.

Третий пробел – игнорирование недоверия как независимого конструкта. Большинство западных исследований трактуют доверие дихотомически, игнорируя российский подход А. Б. Купрейченко, где доверие и недоверие – независимые измерения. Применительно к ИИ это особенно важно, поскольку здоровое недоверие – адаптивная стратегия в условиях непрозрачности и несовершенства систем. Необходимы двухшкальные инструменты.

Четвёртый пробел – недостаток лонгитюдных исследований. Подавляющее большинство исследований – кросс-секционные, измеряющие доверие в один момент времени. Однако доверие к ИИ динамично: первоначальное доверие трансформируется в непрерывное доверие, происходит калибровка после ошибок, изменение с опытом. Необходимы лонгитюдные дизайны, отслеживающие доверие на протяжении опыта взаимодействия.

Пятый пробел – культурная предвзятость (*WEIRD* выборки). Большинство исследований проводится на западных, образованных, индустриализированных, богатых, демократических выборках, но доверие к ИИ принципиально различается между культурами. Необходимы кросс-культурные валидации моделей, разработка культурно-специфичных операционализаций, вовлечение исследователей из незападных стран.

Шестой пробел – недооценка эмоционального компонента. Большинство исследований фокусируются на когнитивной оценке, недооценивая роль эмоций, хотя эмпирические данные показывают, что эмоции оказывают более сильное влияние. Необходимы исследования эмоциональной регуляции доверия, включая нейрофизиологические методы.

Седьмой пробел – недостаток исследований парадоксов. Лишь немногие работы фокусируются на парадоксах доверия к ИИ, хотя именно парадоксы – ключ к пониманию феномена. Необходим философско-методологический анализ противоречий как исследовательская программа.

Восьмой пробел – разрыв между теорией и практикой. Теоретические модели часто не валидируются эмпирически; эмпирические исследования не опираются на прочную теоретическую базу. Необходима интеграция теории и эмпирики как обязательный стандарт.

На основе выявленных пробелов мы выделяем девять приоритетных направлений будущих исследований доверия к ИИ.

1. Разработка и валидация многомерных инструментов измерения доверия к ИИ, включающих все компоненты (когнитивный, аффективный, поведенческий), различающих уровни (технология, организация, институт), типы (начальное, непрерывное, эпистемическое, социальное), измеряющих как доверие, так и недоверие, кросс-культурно валидированных.

2. Лонгитюдные исследования динамики доверия, отслеживающие изменение от начального к непрерывному доверию, калибровку после ошибок ИИ, изменение с опытом взаимодействия. Особый интерес представляют критические инциденты (этические скандалы, крупные ошибки ИИ) и их влияние на общественное доверие.

3. Глубокое изучение парадоксов доверия к ИИ: почему объяснимость иногда снижает доверие? При каких условиях

доминирует смещение автоматизации, а при каких – неприятие алгоритмов? Каков оптимальный уровень антропоморфизации? Когда недоверие к людям усиливает доверие к ИИ?

4. Систематические кросс-культурные исследования, выходящие за рамки измерения «индивидуализм–коллективизм» и включающие другие измерения Г. Хофстеде (дистанция власти, избегание неопределённости, маскулинность–феминность), изучение влияния институционального доверия на доверие к ИИ в разных странах, адаптация моделей для незападных культур.

5. Исследования обоснованного доверия и калибровки доверия: как достичь оптимального соответствия между доверием и реальными способностями ИИ? Какие интервенции эффективны для калибровки? Роль объяснимости, обратной связи, обучения пользователей.

6. Предметно-ориентированные модели доверия к ИИ. Большинство исследований изучают «доверие к ИИ» вообще, но различные типы ИИ (генеративный ИИ, экспертные системы, роботы, чатботы) и домены применения (здравоохранение, финансы, образование, юстиция) могут требовать различных механизмов доверия.

7. Этические аспекты доверия к ИИ: когда доверие этически оправдано, а когда – нет? Проблема манипуляции доверием через дизайн (дарк-паттерны, эксплуатация антропоморфизма). Разработка этических гайдлайнов для дизайна ИИ, формирующего доверие.

8. Доверие в контексте ошибок и вреда: как восстанавливается доверие после ошибок ИИ? Роль извинений, компенсаций, объяснений. Ассиметричное воздействие: одна ошибка может разрушить доверие, построенное долгим позитивным опытом (неприятие алгоритмов).

9. Интеграция доверия к ИИ в более широкие социальные процессы: как доверие к ИИ связано с социальным капиталом, институциональным доверием,

доверием к науке? Может ли низкое социальное доверие компенсироваться высоким доверием к ИИ, или это маркер дисфункционального общества?

Теоретическая значимость настоящего исследования заключается в следующем: во-первых, предложено интегративное определение доверия к ИИ, объединяющее психологические, социологические и технические аспекты и учитывающее специфику взаимодействия человека с интеллектуальными агентами. Это определение может служить основой для построения единой теоретической рамки, преодолевающей фрагментарность существующих подходов. Во-вторых, систематизированы десять ключевых парадоксов доверия к ИИ, что открывает новую исследовательскую программу, фокусирующуюся на внутренних противоречиях феномена. В-третьих, выявлены семь специфических черт, отличающих доверие к ИИ от межличностного доверия, что обосновывает необходимость разработки собственной концептуализации, не редуцирующей доверие к ИИ к традиционным формам доверия. В-четвёртых, обоснованы десять приоритетных направлений будущих исследований, что может способствовать более систематическому и координированному развитию научного знания в данной области.

Практическая значимость исследования определяется следующим: во-первых, результаты значимы для дизайна надёжных и вызывающих доверие ИИ-систем: понимание механизмов формирования доверия, парадоксов и специфических черт позволяет разработчикам проектировать системы, избегающие смещения автоматизации и неприятия алгоритмов и способствующие обоснованному доверию. Во-вторых, выявленные культурные различия критичны для глобального внедрения ИИ: универсальные подходы, игнорирующие культурный контекст, обречены на низкую эффективность. Необходима адаптация ИИ-систем и

коммуникационных стратегий к культурным особенностям различных регионов. В-третьих, результаты могут быть использованы для разработки эффективной регуляторной политики в области ИИ: понимание того, что формирует доверие общества к ИИ, позволяет создавать законодательство и стандарты, способствующие ответственным инновациям. В-четвёртых, выявленная U-образная динамика опыта имеет важные следствия для программ ИИ-грамотности: необходимо не просто повышать осведомлённость об ИИ, но и развивать критическое мышление, предотвращая формирование смещения автоматизации у пользователей со средним уровнем знаний.

Настоящее исследование имеет ограничения, связанные с методом теоретического анализа. Отбор источников, хотя и осуществлён по явным критериям, может не охватывать всё многообразие подходов к концептуализации доверия к ИИ. Интегративное определение, предложенное на основе критического анализа, требует эмпирической валидации. Необходимы исследования, операционализирующие все компоненты определения и проверяющие его предсказательную валидность. Выявленные парадоксы и специфические черты основаны на анализе литературы и требуют дальнейшего эмпирического подтверждения в различных культурных контекстах и доменах применения ИИ.

Доверие к искусственному интеллекту – это не просто адаптация межличностного доверия к технологическому контексту, а новый феномен, требующий собственной концептуализации, теории и методологии исследования. Понимание психологических механизмов формирования доверия к ИИ критично не только для науки, но и для практики: для дизайна надёжных и этичных ИИ-систем, разработки эффективной регуляторной политики и, в конечном счёте, для построения общества, где ИИ служит благу человека, а не заменяет или подрывает

человеческую автономию и достоинство. Выявленные парадоксы указывают на сложность и многогранность феномена, требующего детального подхода как в исследованиях, так и в практике. Направления будущих исследований,

обоснованные в настоящей работе, могут способствовать преодолению выявленных пробелов и построению единой теоретической и эмпирической базы для понимания доверия к ИИ.

ЛИТЕРАТУРА

1. Trust, attitudes and use of artificial intelligence: A global study 2025 / N. Gillespie, S. Lockey, T. Ward, A. Macdade, G. Hassed // The University of Melbourne and KPMG. 2025. URL: <https://assets.kpmg.com> (дата обращения: 10.10.2025). DOI: 10.26188/28822919.
2. Stevens A. F., Stetson P. Theory of trust and acceptance of artificial intelligence technology (TrAAIT): An instrument to assess clinician trust and acceptance of artificial intelligence // Journal of Biomed Information. 2023. № 148. URL: <https://pubmed.ncbi.nlm.nih.gov> (дата обращения: 10.10.2025). DOI: 10.1016/j.jbi.2023.104550.
3. Measuring trust in artificial intelligence: validation of an adapted Trust in Automation and AI Scale (TIAS) / M. J. McGrath, O. Lack, J. Tisch, A. Duenser // Frontiers in Artificial Intelligence. 2025. № 8. URL: <https://www.frontiersin.org/journals/artificial-intelligence> (дата обращения: 10.10.2025). DOI: 10.3389/frai.2025.1582880.
4. Trust in AI: progress, challenges, and future directions / S. Afroogh, A. Akbari, E. Malone, M. Kargar, H. Alambeigi // Humanities and Social Sciences Communications. 2024. № 11. URL: <https://arxiv.org/abs/2403.14680> (дата обращения: 10.10.2025). DOI: 10.1057/s41599-024-04044-8.
5. Horowitz M. C., Kahn L. Bending the Automation Bias Curve: A Study of Human and AI-Based Decision Making in National Security Contexts // International Studies Quarterly. 2024. № 68. Iss. 2. URL: <https://arxiv.org/abs/2306.16507> (дата обращения: 10.10.2025). DOI: 10.1093/isq/sqae020.
6. Twenty-four years of empirical research on trust in AI: a bibliometric review of trends, overlooked issues, and future directions / M. Benk, S. Kerstan, F. v. Wangenheim, A. Ferrario // AI & Society. 2025. № 40. P. 2083–2106. DOI: 10.1007/s11301-023-00401-5.
7. Developing trustworthy artificial intelligence: insights from research on interpersonal, human-automation, and human-AI trust / Y. Li, B. Wu, Y. Huang, S. Luan // Frontiers in Psychology. 2024. № 15. URL: <https://www.frontiersin.org/journals/psychology> (дата обращения: 10.10.2025). DOI: 10.3389/fpsyg.2024.1382693.
8. Dang Q., Li G. Unveiling trust in AI: the interplay of antecedents, consequences, and cultural dynamics // AI & Society. 2025. P. 669–692. DOI: 10.1007/s00146-025-02477-6.
9. Цифровое доверие как основа датацентричного управления: результаты опросного эксперимента / А. Т. Санина, А. М. Семенова, Е. М. Стырин, А. Т. Атаева // Мониторинг общественного мнения: экономические и социальные перемены. 2025. № 1. С. 168–194. DOI: 10.14515/monitoring.2025.1.2582.
10. Lukyanenko R., Maass W., Storey V. C. Trust in artificial intelligence: From a Foundational Trust Framework to emerging research opportunities // Electronic Markets. 2022. Vol. 32, № 11. P. 1993–2020. DOI: 10.1007/s12525-022-00605-4.
11. Jose B., Thomas A. Digital anthropomorphism and the psychology of trust in generative AI tutors: an opinion-based thematic synthesis // Frontiers in Computer Science. 2025. № 7. URL: <https://www.frontiersin.org/journals/computer-science> (дата обращения: 10.10.2025). DOI: 10.3389/fcomp.2025.1638657.
12. Купрейченко А. Б. Психология доверия и недоверия. М.: Институт психологии РАН, 2008. 564 с.
13. Vaish A., Grossmann T., Woodward A. Not all emotions are created equal: the negativity bias in social-emotional development // Psychological Bulletin. 2008. № 134 (3). P. 383–403. DOI: 10.1037/0033-2909.134.3.383.
14. Candrian C., Scherer A. Rise of the machines: Delegating decisions to autonomous AI // Computers in Human Behavior. 2022. № 134. URL: <https://www.sciencedirect.com/science> (дата обращения: 10.10.2025). DOI: 10.1016/j.chb.2022.107308.
15. Küper A., Krämer N. Psychological Traits and Appropriate Reliance: Factors Shaping Trust in AI //

- International Journal of Human–Computer Interaction. 2025. № 41 (7). URL: <https://www.sciencedirect.com/science/article> (дата обращения: 10.10.2025). DOI: 10.1080/10447318.2024.2348216.
16. Kieslich K., Keller B., Starke C. Artificial intelligence ethics by design. Evaluating public perception on the importance of ethical design principles of AI // *Big Data & Society*. 2022. Vol. 9, № 1. URL: <https://journals.sagepub.com> (дата обращения: 10.10.2025). DOI: 10.1177/20539517221092956.
 17. Schilke O., Reimann M. The transparency dilemma: How AI disclosure erodes trust // *Organizational Behavior and Human Decision Processes*. 2025. № 188. URL: <https://www.sciencedirect.com/science/article> (дата обращения 10.10.2025). DOI: 10.1016/j.obhdp.2025.104405.
 18. Barnes A. J., Zhang Y., Valenzuela A. AI and culture: Culturally dependent responses to AI systems // *Current Opinion in Psychology*. 2024. № 58. URL: <https://www.sciencedirect.com/science/article> (дата обращения: 10.10.2025). DOI: 10.1016/j.copsyc.2024.101939.
 19. Лешкевич Т. Г. Парадокс доверия к искусственному интеллекту и его обоснование // *Философия науки и техники*. 2023. Т. 28. № 1. С. 34–47. DOI: 10.21146/2413-9084-2023-28-1-34-47.
 20. Hoehl S., Krenn B., Vincze M. Honest machines? A cross-disciplinary perspective on trustworthy technology for children // *Frontiers in Developmental Psychology*. 2024. № 2. URL: <https://www.frontiersin.org/journals/developmental-psychology/articles> (дата обращения: 10.10.2025). DOI: 10.3389/fdpsy.2024.1308881.

REFERENCES

1. Gillespie, N., Lockey, S., Ward, T., Macdade, A. & Hassed, G. (2025). Trust, Attitudes and Use of Artificial Intelligence: A Global Study 2025. In: *The University of Melbourne and KPMG*. URL: <https://assets.kpmg.com> (accessed: 10.10.2025). DOI: 10.26188/28822919.
2. Stevens, A. F. & Stetson, P. (2023). Theory of Trust and Acceptance of Artificial Intelligence Technology (TrAAIT): An Instrument to Assess Clinician Trust and Acceptance of Artificial Intelligence. In: *Journal of Biomed Information*, 148. URL: <https://pubmed.ncbi.nlm.nih.gov> (accessed: 10.10.2025). DOI: 10.1016/j.jbi.2023.104550.
3. McGrath, M. J., Lack, O., Tisch, J. & Duenser, A. (2025). Measuring Trust in Artificial Intelligence: Validation of an Adapted Trust in Automation and AI Scale (TIAS). In: *Frontiers in Artificial Intelligence*, 8. URL: <https://www.frontiersin.org/journals/artificial-intelligence> (accessed: 10.10.2025). DOI: 10.3389/frai.2025.1582880.
4. Afroogh, S., Akbari, A., Malone, E., Kargar, M. & Alambeigi, H. (2024). Trust in AI: Progress, Challenges, and Future Directions. In: *Humanities and Social Sciences Communications*, 11. URL: <https://arxiv.org/abs/2403.14680> (accessed: 10.10.2025). DOI: 10.1057/s41599-024-04044-8.
5. Horowitz, M. C. & Kahn, L. (2024). Bending the Automation Bias Curve: A Study of Human and AI-Based Decision Making in National Security Contexts. In: *International Studies Quarterly*, 68, 2. URL: <https://arxiv.org/abs/2306.16507> (accessed: 10.10.2025). DOI: 10.1093/isq/sqae020.
6. Benk, M., Kerstan, S., Wangenheim v. F. & Ferrario, A. (2025). Twenty-Four Years of Empirical Research on Trust in AI: A Bibliometric Review of Trends, Overlooked Issues, and Future Directions. In: *AI & Society*, 40, 2083–2106. DOI: 10.1007/s11301-023-00401-5.
7. Li, Y., Wu, B., Huang, Y. & Luan, S. (2024). Developing Trustworthy Artificial Intelligence: Insights from Research on Interpersonal, Human-Automation, and Human-AI Trust. In: *Frontiers in Psychology*, 15. URL: <https://www.frontiersin.org/journals/psychology> (accessed: 10.10.2025). DOI: 10.3389/fpsyg.2024.1382693.
8. Dang, Q. & Li, G. (2025). Unveiling Trust in AI: The Interplay of Antecedents, Consequences, and Cultural Dynamics. In: *AI & Society*, 669–692. DOI: 10.1007/s00146-025-02477-6.
9. Sanina, A. G., Semenova, A. M., Styryn, E. M. & Ataeva, A. G. (2025). Digital Trust as the Basis for Data-Centric Governance: Results of a Survey Experiment. In: *Monitoring Public Opinion: Economic and Social Changes*, 1, 168–194. DOI: 10.14515/monitoring.2025.1.2582.
10. Lukyanenko, R., Maass, W. & Storey, V. C. (2022). Trust in Artificial Intelligence: From a Fundamental Trust Framework to Emerging Research Opportunities. In: *Electronic Markets*, 32, 11, 1993–2020. DOI: 10.1007/s12525-022-00605-4
11. Jose, B. & Thomas, A. (2025). Digital Anthropomorphism and the Psychology of Trust in Generative AI Tutors: An Opinion-Based Thematic Synthesis. In: *Frontiers in Computer Science*, 7. URL: <https://www.frontiersin.org/journals/computer-science> (accessed: 10.10.2025). DOI: 10.3389/fcomp.2025.1638657.

12. Kupreychenko, A. B. (2008). *Psychology of Trust and Distrust*. Moscow: Institute of Psychology, Russian Academy of Sciences publ. (in Russ.).
13. Vaish, A., Grossmann, T. & Woodward, A. (2008). Not All Emotions are Created Equal: The Negativity Bias in Social-Emotional Development. In: *Psychological Bulletin*, 134 (3), 383–403. DOI: 10.1037/0033-2909.134.3.383.
14. Candrian, C., Scherer, A. (2022). Rise of the Machines: Delegating Decisions to Autonomous AI. In: *Computers in Human Behavior*, 134. URL: <https://www.sciencedirect.com/science> (accessed: 10.10.2025). DOI: 10.1016/j.chb.2022.107308.
15. Küper, A. & Krämer, N. (2025). Psychological Traits and Appropriate Reliance: Factors Shaping Trust in AI. In: *International Journal of Human–Computer Interaction*, 41 (7). URL: <https://www.sciencedirect.com/science/article> (accessed 10.10.2025). DOI: 10.1080/10447318.2024.2348216.
16. Kieslich, K., Keller, B. & Starke, C. (2022). Artificial Intelligence Ethics by Design. Evaluating Public Perception on the Importance of Ethical Design Principles of AI. In: *Big Data & Society*, 9 (1). URL: <https://journals.sagepub.com> (accessed: 10.10.2025). DOI: 10.1177/20539517221092956.
17. Schilke, O. & Reimann, M. (2025). The Transparency Dilemma: How AI Disclosure Erodes Trust. In: *Organizational Behavior and Human Decision Processes*, 188. URL: <https://www.sciencedirect.com/science/article> (accessed: 10.10.2025). DOI: 10.1016/j.obhdp.2025.104405.
18. Barnes, A. J., Zhang, Y. & Valenzuela, A. (2024). AI and culture: Culturally Dependent Responses to AI systems. In: *Current Opinion in Psychology*, 58. URL: <https://www.sciencedirect.com/science/article> (accessed: 10.10.2025). DOI: 10.1016/j.copsyc.2024.101939.
19. Leshkevich, T. G. (2023). The Paradox of Trust in Artificial Intelligence and Its Justification. In: *Philosophy of Science and Technology*, 28 (1), 34–47. DOI: 10.21146/2413-9084-2023-28-1-34-47.
20. Hoehl, S., Krenn, B. & Vincze, M. (2024). Honest Machines? A Cross-Disciplinary Perspective on Trustworthy Technology for Children. In: *Frontiers in Developmental Psychology*, 2. URL: <https://www.frontiersin.org/journals/developmental-psychology/articles> (accessed: 10.10.2025). DOI: 10.3389/fd-pys.2024.1308881.

ИНФОРМАЦИЯ ОБ АВТОРАХ

Отчерцова Вероника Олеговна (г. Херсон) – аспирант кафедры прикладной психологии и развития личности Института психологии Херсонского государственного педагогического университета; ассистент кафедры психологии социально-гуманитарного факультета Мелитопольского государственного университета;

ORCID: 0009-0003-6721-1566; e-mail: votchertsova@gmail.com

Блинова Елена Евгеньевна (г. Херсон) – доктор психологических наук, профессор, заведующий кафедрой прикладной психологии и развития личности Института психологии Херсонского государственного педагогического университета;

ORCID: 0000-0003-3011-6082; e-mail: elena.blynova@yandex.ru

INFORMATION ABOUT THE AUTHORS

Veronika O. Otchertsova (Kherson) – Postgraduate Student, Department of Applied Psychology and Personality Development, Institute of Psychology, Kherson State Pedagogical University; Assistant Lecturer of the Department of Psychology, Faculty of Social Sciences and Humanities, Melitopol State University; ORCID: 0009-0003-6721-1566; e-mail: votchertsova@gmail.com

Elena E. Blinova (Kherson) – Dr. Sci. (Psychology), Prof., Head of the Department, Department of Applied Psychology and Personality Development, Institute of Psychology, Kherson State Pedagogical University; ORCID: 0000-0003-3011-6082; e-mail: elena.blynova@yandex.ru