

ИСКУССТВЕННЫЙ ИНТЕЛЛЕКТ В ГОСУДАРСТВЕННОМ УПРАВЛЕНИИ



Научная статья

DOI: 10.55959/MSU2073-2643-21-2024-4-151-169

МОДЕЛЬ ДОВЕРИЯ К ПРИКЛАДНЫМ СИСТЕМАМ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА

С.В. Гарбук

Всероссийский институт научной и технической информации Российской академии наук (ВИНИТИ РАН), Москва, Российская Федерация

garbuk@list.ru

Аннотация. Рассматривается проблема оценки и подтверждения доверия к прикладным системам искусственного интеллекта (СИИ). Показано, что, несмотря на разнообразие работ в данной области, на практике отсутствует единый подход к подтверждению доверия. При этом разные исследователи делают акцент на тех или иных аспектах доверия, без попытки охватить проблему в целом. Свойство доверия сводится к эмоциональной характеристике, конструктивный смысл которой в контексте управления жизненным циклом СИИ не определен. В некоторых случаях происходит не вполне оправданное «очеловечивание» понятия доверия, в результате которого характеристики технической системы полностью замещаются субъективным отношением к ее создателям, поставщикам или владельцам.

В работе предпринята попытка обоснования целостной модели доверия к СИИ, основанной на формализации связи между доверием и качеством системы. При этом показано, что характеристика доверия может быть использована для принятия решения эксплуатантом, регуляторами и другими заинтересованными лицами о возможности (или невозможности) применения системы по назначению в определенных условиях эксплуатации. Выявлены основные аспекты доверия, учитывающие как общие вопросы доверия к автоматизированным системам (техническая надежность, информационная безопасность и импортонезависимость), так и специфику систем ИИ на основе алгоритмов машинного обучения (функциональная корректность). Предложен подход к оцениванию характеристик СИИ по этим аспектам.

Полученные результаты могут быть использованы для развития системы оценки соответствия в области технологий ИИ, а также при выборе систем, обеспечивающих решение задач интеллектуальной обработки данных в различных отраслях экономики и социальной сферы.

Ключевые слова: системы искусственного интеллекта, доверие к искусственному интеллекту, качество систем искусственного интеллекта, оценка соответствия, функциональная корректность программных средств, тестирование программного обеспечения.

Для цитирования: Гарбук С.В. Модель доверия к прикладным системам искусственного интеллекта // Вестник Московского университета. Серия 21. Управление (государство и общество). 2024. Т. 21. № 4. С. 151–169.

Дата поступления в редакцию: 18.09.2024

TRUST MODEL FOR ARTIFICIAL INTELLIGENCE APPLICATION SYSTEMS

Garbuk S.V.

All-Russian Institute for Scientific and Technical Information of the Russian Academy of Sciences, Moscow, Russian Federation.

garbuk@list.ru

Abstract. The problem of assessment and trust validation of applied artificial intelligence systems is examined. Despite the diversity of studies in this area, it is shown that in fact there is no unified approach to trust validation. At the same time, different researchers emphasize one or another aspect of trust without attempting to cover the problem as a whole. The trust feature is reduced to an emotional characteristic, the constructive meaning of which in the context of artificial intelligence systems lifecycle management is not defined. In some cases there is an unjustified 'humanization' of the concept of trust, as a result of which the characteristics of a technical system are completely replaced by the subjective treatment of its creators, suppliers or owners.

The paper attempts to substantiate a holistic trust model for artificial intelligence systems based on formalization of the relationship between trust and system quality. In this case the paper shows that the characteristic of trust can be used to make a decision by the operator, regulators and other interested parties about the possibility (or impossibility) to use the system as intended in certain operating conditions. The main aspects of trust are identified, taking into account both the general issues of trust in automated systems (technical reliability, information security and import independence) and the specifics of AI systems based on machine learning algorithms (functional correctness), and

an approach to evaluating the characteristics of the artificial intelligence systems according to these aspects is proposed.

The obtained results can be used for the development of conformity assessment system in the field of AI technologies, as well as for the selection of systems that provide the solution of intelligent data processing tasks in various industries and social sphere.

Key words: artificial intelligence systems, trust in artificial intelligence, quality of artificial intelligence systems, conformity assessment, conformity assessment, functional correctness of software tools, software testing.

For citation: Garbuk S.V. Trust model for artificial intelligence application systems // Lomonosov Public Administration Journal. Series 21. 2024. Vol. 21. № 4. P. 151–169.

Received: 18.09.2024

Введение. Проблема доверия к системам искусственного интеллекта

Термины «доверенные системы» (англ. trusted systems) и «доверенные автоматизированные системы» начали использоваться с середины 80-х гг. Так, еще в 1985 г. министерством обороны США был подготовлен стандарт, устанавливающий требования к критериям оценивания доверенных автоматизированных систем (АС) военного назначения DoD 5200.28-STD¹. Этот документ был полностью посвящен аспектам доверия, связанными с обеспечением информационной безопасности (ИБ) автоматизированных систем обработки данных, применяемых в МО США. В частности, документ позволял:

- пользователям АС определять, какая информация ограниченного доступа может обрабатываться в системе с тем или иным уровнем доверия;
- разработчикам АС учитывать требования в области ИБ при создании новых систем для чувствительных применений;
- заказчикам АС устанавливать требования в области ИБ при приобретении систем.

В документе DoD 5200.28-STD впервые было четко обозначено, что безопасная обработка информации в АС достигается за счет выполнения двух групп требований: а) особых требований к характеристикам безопасности; б) требований доверия к процессу создания и эксплуатации АС. Этот подход в дальнейшем был раз-

¹ DoD 5200.28-STD. Department of Defense Standard. Department of Defense Trusted Computer System Evaluation Criteria. December 1985.

вит в разработанных позднее и широко применяемых и по сей день стандартах серии ISO/IEC 15408-X:2009 («Критерии оценки безопасности информационных технологий», т.н. «Общие критерии»).

Подчеркнем, что как в раннем документе DoD 5200.28-STD, так и в «Общих критериях» речь идет исключительно о доверии к АС в смысле защиты информации, обрабатываемой в системе. Так, в DoD 5200.28-STD указывается, что интерпретация характеристик безопасности АС осуществляется с учетом функциональных характеристик конкретной системы, при этом оценка самих функциональных характеристик, связанных с применением системы по назначению, не предполагается. В «Общих критериях» речь идет о подтверждении характеристик продуктов информационных технологий (ИТ) с функциональными возможностями безопасности (средств защиты информации), а также о подтверждении мер доверия, предпринятых к таким продуктам ИТ. Особо указывается, что результат оценки в соответствии со стандартами серии ISO/IEC 15408, «имеет значимость только в контексте свойств безопасности, которые были оценены, и методов оценки, которые использовались».

Применительно к системам искусственного интеллекта (СИИ) вопросы доверия начали рассматриваться начиная с 2018 г. Наиболее активная деятельность по подготовке международных технических стандартов в области доверия к системам ИИ (СИИ) проводится профильным подкомитетом ISO/IEC SC42 Artificial Intelligence в рамках рабочей группы WG 3 Trustworthiness. В частности, из 33 опубликованных международных стандартов 12 разрабатывались именно в этой группе (всего подкомитет SC 42 насчитывает пять рабочих, две специальные рабочие и шесть объединенных рабочих групп).

В результате этой деятельности термины доверия к СИИ получили несколько более расширенное по сравнению с областью ИБ толкование. В частности, в техническом отчете подкомитета SC 42² были предложены следующие определения:

- доверие — степень уверенности пользователя или другой заинтересованной стороны в том, что продукт или система будут вести себя как задумано;
- обладание доверием (англ. trustworthiness) — свойство объекта являться надежным и достоверным.

² ISO/IEC TR 24028:2020 (E). Information technology — Artificial intelligence — Overview of trustworthiness in artificial intelligence. Technical Report. First Edition, 2020-05.

В соответствии с этими определениями доверие к СИИ предполагает не только достаточно низкий уровень рисков, связанных с возможностью реализации угроз ИБ, но и соответствие ожиданиям потребителя в отношении иных свойств системы. На основе отчета был подготовлен национальный технический стандарт ГОСТ Р 59276-2020³, в котором доверие рассматривается в качестве важнейшего условия, определяющего возможность применения СИИ для решения конкретных задач обработки данных: «Уверенность потребителя, организаций, ответственных за регулирование вопросов создания и применения систем искусственного интеллекта, и иных заинтересованных сторон в том, что система способна выполнять возложенные на нее задачи с требуемым качеством». Стандарт определяет доверие как бинарную величину, определяющую возможность применения конкретной СИИ s для решения прикладной задачи ИИ g в определенных (предусмотренных) условиях эксплуатации:

$$\text{Trust}(s, g, W(E)) \in \{0,1\}, \quad (1)$$

где E — набор существенных факторов эксплуатации СИИ; $W(E)$ — плотность распределения вероятностей этих факторов.

Дальнейшее развитие такой подход получил в международном документе по стандартизации⁴, в соответствии с которым СИИ является доверенной (англ. *trustworthy*) при выполнении следующих требований:

- функциональность соответствует назначению в предусмотренных условиях применения;
- целостность программного кода и аппаратных компонентов поддерживается на протяжении всего жизненного цикла (ЖЦ) (с учетом возможности взлома, производственных ошибок, надлежащих действий по обслуживанию и т.д.);
- все действия/задачи ЖЦ, влияющие на потребителей, выполняются этично и справедливо (продажи и маркетинг до конца срока службы).

В работе Патрика Холла⁵ делается акцент на следующих аспектах доверия к СИИ:

³ ГОСТ Р 59276-2020 Системы искусственного интеллекта. Способы обеспечения доверия. Общие положения. М.: Стандартинформ, 2020. 20 с.

⁴ ISO/IEC JTC 1/SC 42/ANG 7 “JTC1 joint development review”. Convenorship: SCC CEN-CLC-JTC21 N216 ANG7 Proposal for NWI AI trustworthiness characterization. 2022-12-28.

⁵ Hall P. Machine Learning for High-Risk Applications. First Edition. Techniques for Responsible AI. O'Reilly Media, Inc, 2022. 112 p.

- информационные атаки (недекларированные возможности, атаки «отравления» данных, атаки на извлечение моделей машинного обучения (МО) и др.);
- сбои в работе (смещенность обучающих данных, дискриминация, непрозрачность алгоритмов и др.);
- злонамеренное применение (этнический профайлинг, киберпреступления с использованием ИИ, смертоносные автономные СИИ и др.).

Таким образом, объективно присущая алгоритмам МО особенность (неполная интерпретируемость) приравняется к таким факторам качества систем, как статистическая смещенность и конфиденциальность данных. В целом, следует отметить, что в разрабатываемых документах по стандартизации и научных работах затрагиваются лишь отдельные аспекты, влияющие на доверие к СИИ: защита информации, робастность, качество используемых данных, статистическая смещенность формируемых оценок, этические аспекты применения ИИ и некоторые другие.

Ситуация с целостным анализом вопросов доверия к СИИ усложняется еще и тем, что сам термин «доверие» имеет гуманитарное происхождение. Подробно опасность избыточной «гуманитаризации» вопроса доверия к СИИ рассмотрена в работе⁶. Показано, что при использовании устоявшегося в области социальных наук понятия «доверие» применительно к СИИ необходимо учитывать следующие особенности и ограничения:

- понятие «доверие» не определено для изолированной СИИ. Доверие к СИИ фактически означает доверие к сообществу, обеспечивающему разработку, создание и поддержание СИИ в процессе эксплуатации, включая компании разработчики и поставщики систем, государственные регуляторы и иные организации (экосистеме ИИ);
- доверие к СИИ всегда включает ту или иную долю безосновательной веры в то, что цели, ценности и мотивы участников экосистемы ИИ в основном совпадают с соответствующими установками пользователя системы;
- уровень доверия зависит не только от свойств объекта — СИИ, но также и от субъекта, формирующего свои представления об объекте, что в целом ряде случаев является мешающим, искажающим объективно присущие СИИ свойства.

⁶ *Гарбук С.В.* Особенности применения понятия «доверие» в области искусственного интеллекта. Искусственный интеллект и принятие решений. 2020. № 3. С. 15–21.

Иллюстрацией таких рисков является, например, работа⁷, где в качестве критериев доверия к СИИ приводится неоднородный список, включающий как необходимость обеспечения доверия к обучающим наборам данных (НД), так и участие государства и организаций международного сотрудничества в работах по актуализации этих данных. Тем самым акцент в обеспечении доверия к СИИ смещается в организационно-правовую и даже гуманитарную плоскость: понимание свойств технических систем замещается авторитетом и полномочиями организаций, участвующих в создании и применении этих систем.

Таким образом, в области обеспечения и подтверждения доверия к СИИ в настоящее время следует выделить следующие проблемные вопросы, снижающие практическую значимость понятия «доверие»:

- 1) опасность редуцирования понятия «доверие к СИИ» исключительно к требованиям в области защиты информации в этих системах;
- 2) неполнота аспектов доверия, существенных для заинтересованных сторон, принимающих участие в создании, применении и нормативном регулировании технологий ИИ;
- 3) чрезмерное «очеловечивание» понятия доверия, повышающее уровень субъективизма при оценивании СИИ в ущерб проведению объективных измерений значимых свойств этих технических систем.

Доверие и качество систем искусственного интеллекта

Уточнение понятия «доверие к СИИ» должно заключаться в обеспечении полноты охвата аспектов, влияющих на принятие решений о возможности использования СИИ, а также в повышении объективности оценивания характеристик систем по этим аспектам. При этом целесообразно учитывать, что определения доверия к СИИ^{8, 9, 10}

⁷ Artificial Intelligence and National Security. Greg Allen, Taniel Chan. A study on behalf of Dr. Jason Matheny, Director of the U.S. Intelligence Advanced Research Projects Activity (IARPA). Belfer Center Study. July, 2017.

⁸ ISO/IEC TR 24028:2020 (E). Information technology — Artificial intelligence — Overview of trustworthiness in artificial intelligence. Technical Report. First Edition, 2020-05.

⁹ ГОСТ Р 59276-2020 Системы искусственного интеллекта. Способы обеспечения доверия. Общие положения. М.: Стандартинформ, 2020. 20 с.

¹⁰ ISO/IEC JTC 1/SC 42/ANG 7 “JTC1 joint development review”. Convenorship: SCC CEN-CLC-JTC21 N216 ANG7 Proposal for NWI AI trustworthiness characterization. 2022-12-28.

в значительной мере совпадают с определениями качества продукции, под которым принято понимать:

- совокупность свойств продукции (в данном случае — СИИ), обуславливающих ее пригодность удовлетворять определенные потребности в соответствии с ее назначением¹¹;
- совокупность свойств и характеристик продукции или услуги, которые придают им способность удовлетворять обусловленные или предполагаемые потребности потребителя¹²;
- степень соответствия совокупности присущих характеристик объекта требованиям¹³.

Так же, как и в случае с доверием, оценивание качества осуществляется с учетом позиции определенных заинтересованных сторон (англ. stakeholders), причем интересы и приоритеты этих сторон могут быть представлены в виде совокупности требований к СИИ и доверие в этом случае основывается на объективно подтвержденном соответствии характеристик систем установленным требованиям (рис. 1).

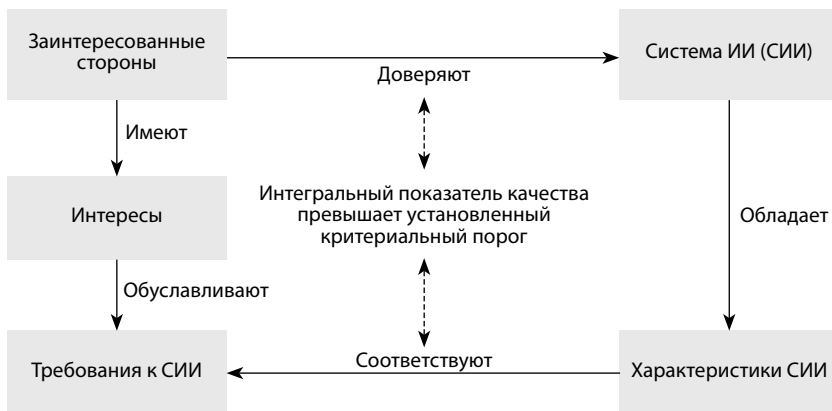


Рис. 1. Требования, характеристики, качество и доверие к СИИ¹⁴

¹¹ ГОСТ 15467-79 (СТ СЭВ 3519-81) Управление качеством продукции. Основные понятия. Термины и определения. Утвержден и введен в действие постановлением Государственного комитета СССР по стандартам от 26 января 1979 г. N 244.

¹² Международные стандарты. «Управление качеством продукции». ИСО 9000-9004, ИСО 8402. М.: Изд-во стандартов, 1988.

¹³ ГОСТ Р ИСО 9000-2015 Системы менеджмента качества. Основные положения и словарь. М.: Стандартиформ, 2015. 77 с.

¹⁴ Составлено автором.

Будем считать, что доверие к СИИ обеспечивается в том и только в том случае, если качество системы, определенное согласно ГОСТ Р ИСО 9000-2015 как степень соответствия характеристик системы установленным требованиям, достигает заданного значения¹⁵:

$$Trust(s, g, W(E)) = \begin{cases} 1, & \text{если } q(s, g, W(E)) \geq q^* \\ 0, & \text{если } q(s, g, W(E)) < q^* \end{cases} \quad (2)$$

где q — интегральный показатель качества СИИ s при решении задачи g в условиях $W(E)$; q^* — критериальное значение показателя качества, при достижении которого система считается доверенной ($Trust(s, g, W(E)) = 1$).

Аспекты доверия к системам искусственного интеллекта

Как уже было отмечено выше, при оценивании доверия к СИИ принципиальным моментом является учет максимального количества аспектов, существенным образом влияющих на возможность применения систем по назначению. Так, например, признавая важность обеспечения доверия на основе тестирования СИИ, Экспертная группа по ИИ высокого уровня Еврокомиссии (High-Level Expert Group on AI, AI HLEG) подготовила документ по самотестированию компаний-разработчиков СИИ, в соответствии с которым разработчики смогли бы оценивать создаваемые ими системы по следующим критериям¹⁶:

- подконтрольность и подотчетность человеку;
- техническая надежность и безопасность;
- конфиденциальность обрабатываемых данных;
- объяснимость используемых алгоритмов;
- отсутствие дискриминации разных групп пользователей и справедливость принимаемых решений;
- поддержка социального и экологического благополучия.

¹⁵ Garbuk S.V. Intellimetry as a way to ensure AI trustworthiness//The Proceedings of the 2018 International Conference on Artificial Intelligence Applications and Innovations (IC-AIAI). Limassol, Cyprus, 6–10.10.2018. С. 27–30. ISBN: 978-1-7281-0412-6. DOI: 10.1109/IC-AIAI.2018.00012.

¹⁶ The Assessment list for trustworthy Artificial Intelligence (ALTAI). For self assessment. European Commission B-1049 Brussels. 17th of July 2020. 34 p.

Данный перечень аспектов доверия дополняет приведенные выше^{17, 18}, однако по-прежнему не претендует на полноту. Предложенное в данной работе определение доверия, основанное на оценке качества СИИ, позволяет выполнить обоснование такого полного перечня с использованием моделей качества систем. В частности, принимая во внимание, что СИИ относятся к промышленной продукции, для описания общих характеристик качества может быть использована универсальная классификация, основанная на характеризующих свойствах¹⁹: характеристики назначения, надежности (безотказности, долговечности, ремонтпригодности, сохраняемости), эргономические, эстетические, технологичности, стандартизации и унификации, транспортабельности, патентно-правовые, экологические, безопасности, экономические.

Кроме того, для более детального описания множества общих характеристик СИИ целесообразно использовать стандарты программной и системной инженерии (SQuaRE), устанавливающие требования к качеству систем и программных средств²⁰. Так, в основополагающем стандарте²¹ для оценивания качества предложено использовать три специализированных модели:

- модель качества при использовании;
- модель качества продукта;
- модель качества данных.

Модели являются взаимодополняющими и рассчитаны на разные категории заинтересованных лиц, включая разработчиков ПО,

¹⁷ ISO/IEC JTC 1/SC 42/AHG 7 “JTC1 joint development review”. Convenorship: SCC CEN-CLC-JTC21 N216 AHG7 Proposal for NWI AI trustworthiness characterization. 2022-12-28.

¹⁸ *Hall P.* Machine Learning for High-Risk Applications. First Edition. Techniques for Responsible AI. O'Reilly Media, Inc, 2022. 112 p.; *Allen G., Chan T.* Artificial Intelligence and National Security. A study on behalf of Dr. Jason Matheny, Director of the U.S. Intelligence Advanced Research Projects Activity (IARPA). Cambridge, Belfer Center Study, 2017.

¹⁹ РД-50-149-79 Методические указания по оценке технического уровня и качества промышленной продукцию. Утвержден постановлением Госстандарта 17.04.79 № 1407.

²⁰ *Гарбук С.В.* Оценка качества систем искусственного интеллекта: особенности проведения и нормативно-техническая база. Контроль качества продукции. 2024, №1. С. 20–26.

²¹ ГОСТ Р ИСО/МЭК 25010-2015. Информационные технологии. Системная и программная инженерия. Требования и оценка качества систем и программного обеспечения (SQuaRE). Модели качества систем и программных продуктов. Дата введения — 2016-06-01.

системных интеграторов, приобретателей, владельцев, специалистов по обслуживанию, подрядчиков, профессионалов обеспечения и управления качеством. Модели различаются также объектами оценивания качества.

На основании универсального документа ГОСТ Р ИСО/МЭК 25010-2015 в 2023 году был подготовлен предварительный национальный стандарт ПНСТ 842-2023²², учитывающий следующие особенности создания и применения СИИ:

- самостоятельно принимать решения;
- основываться на искаженных или неполных данных;
- являться вероятностными;
- адаптироваться в процессе эксплуатации.

В соответствии с ним, а также предварительным национальным стандартом ПНСТ 841-2023²³, также развивающим подходы SQuaRE с учетом особенностей реализации СИИ, модель качества продукта и модель качества при использовании, применимые к любой технической системе, могут применяться и к СИИ. При этом для СИИ было добавлено несколько подхарактеристик, а для некоторых из них изменены значения или контекст. ПНСТ 841-2023 применительно для СИИ развивает также положения по оценке характеристик и подхарактеристик качества, изложенные в стандарте серии SQuaRE²⁴.

Перечень характеристик качества, полученный путем объединения характеристик качества промышленной продукции и общих характеристик СИИ, представлен в табл. 1. Видно, что наибольшую детализацию в документах SQuaRE получили показатели назначения промышленной продукции, в то время как некоторые (характеристики классификационные, конструктивные, состава и структуры, материалоемкости, транспортабельности, патентно-правовые) не нашли отражения в документах ПНСТ 842-2023 и ПНСТ 841-2023 в связи с очевидным отсутствием особенностей для СИИ или в связи с неприменимостью.

²² ПНСТ 842-2023 Программная инженерия. Требования и оценка качества систем и ПО (SQuaRE). Модель качества для систем ИИ (ISO 25059:2023, MOD).

²³ ПНСТ 841-2023 Системная и программная инженерия. Требования и оценка качества систем и программного обеспечения (SQuaRE). Руководство по оценке качества систем ИИ (ISO/IEC DTS 25058, MOD).

²⁴ ГОСТ Р ИСО/МЭК 25040-2014 Информационная технология. Системная и программная инженерия. Требования и оценка качества систем и программного обеспечения (SQuaRE). Процесс оценки (ISO/IEC DTS 25040-2011, IDT).

Характеристики качества и аспекты доверия СИИ²⁵

Промышленная продукция (РД-50-149-79)		СИИ (ПНСТ 842-2023 и ПНСТ 841-2023)		Аспекты доверия
Характеристики	Подхарактеристики	Подхарактеристики	Характеристики	
1. Показатели назначения	Классификационные	–	–	
	Функциональные	Функциональная полнота, ... корректность, ... целесообразность, ... адаптируемость	Функциональная пригодность	Функциональная корректность (ФК), технологическая независимость (ТН)
		Определимость пригодности, изучаемость, доступность, управляемость со стороны пользователя, прозрачность	Удобство использования	Эргономичность (Э), ТН
		Робастность	Надежность	Надежность (Н), ФК
		Полноценность, доверие, удовлетворение, прозрачность	Удовлетворенность	ФК, Э, ТН
		Смягчение отрицательных последствий экономического риска, этического и социального риска	Свобода от риска	ФК, защита информации (ЗИ)
Технической эффективности (эксплуатационные)		Полнота контекста, гибкость	Покрывание контекста	ФК
		Временные характеристики, потенциальные возможности	Уровень производительности	ФК
		Результативность	Эффективность	ФК
		Производительность	Производительность	ФК

²⁵ Составлено автором.

Промышленная продукция (РД-50-149-79)		СИИ (ПНСТ 842-2023 и ПНСТ 841-2023)		Аспекты доверия
Характеристики	Подхарактеристики	Подхарактеристики	Характеристики	
	Конструктивные	-	-	-
	Состава и структуры	-	-	-
2. Показатели надежности	Безотказности, долговечности, ремонтопригодности, сохраняемости	Завершенность, готовность, отказоустойчивость, восстанавливаемость	Надежность	Н, ТН
3. Эргономические показатели	Гигиенические	Комфорт	Удовлетворенность	Э
	Антропометрические			Э
	Физиологические и психофизиологические			Э
	Психологические			Э
4. Эстетические показатели	Информационной выразительности, рациональности формы, целостности композиции, совершенства производственного исполнения и стабильности товарного вида	Удовольствие, прозрачность	Удобство использования	Э, ТН
5. Показатели технологичности	Трудоемкости	Смягчение отрицательных последствий экономического риска		Н, ТН
	Энергоемкости	Характеристики использования ресурсов	Уровень производительности	ФК
	Себестоимости	Смягчение отрицательных последствий экономического риска	Свобода от риска	ФК, ТН

Промышленная продукция (РД-50-149-79)		СИИ (ПНСТ 842-2023 и ПНСТ 841-2023)		Аспекты доверия
Характеристики	Подхарактеристики	Подхарактеристики	Характеристики	
6. Показатели транспортабельности	Материалоемкости	-	-	-
	Транспортабельности	-	-	-
7. Показатели стандартизации и унификации	Стандартизации	Существование, интероперабельность	Совместимость	ТН
	Унификации	Простота внедрения, (взаимо-) заменяемость	Переносимость	ТН
8. Патентно-правовые показатели	Модульность, возможность многократного использования	Сопровождаемость	Сопровождаемость	ТН
	Патентной защиты	-	-	ТН
9. Экологические показатели	Патентной чистоты	-	-	ТН
	Экологичности	Смягчение отрицательных последствий экологического риска	Свобода от риска	ФК
10. Показатели безопасности	Безопасности	Управляемость, защищенность от ошибки пользователя	Удобство использования	Э, ФК
	Безопасности	Конфиденциальность, целостность, неотказуемость, подотчетность, подлинность, возможность вмешательства	Безопасность	ЗИ, ФК
		Смягчение отрицательных последствий риска для здоровья и безопасности	Свобода от риска	ФК, ЗИ

Аспекты доверия к системам искусственного интеллекта²⁶

Аспекты доверия к СИИ	Способы регулирования СИИ
Техническая надежность и эргономичность	Аккредитованные лаборатории по проведению испытаний на устойчивость к климатическим, механическим и иным воздействиям
	КГВС «Мороз-7», «Климат-8» и др.
Защита информации	Система сертификации на соответствие требованиям в области защиты информации
	Национальные стандарты и руководящие документы ФСБ и ФСТЭК России
Технологическая независимость	Реестр отечественного ПО, Минцифры России
	Реестр радиоэлектронной продукции российского происхождения, Минпромторг России
Функциональная корректность	Общие подходы к оценке качества программных систем, не учитывающие особенности систем с ИИ
	Организации по оценке соответствия отсутствуют

Характеристики СИИ в табл. 1 могут быть сгруппированы по следующим аспектам доверия, различающимся по способам регулирования (табл. 2):

- техническая надежность (Н), на обеспечение которой направлены меры по снижению количества отказов аппаратных средств и программного обеспечения (ПО), используемых в СИИ. Для регулирования соответствующих вопросов разработана необходимая нормативная база, созданы органы по оценке соответствия в области надежности;
- эргономичность (Э) СИИ, критическое невыполнение требований к которой также в некоторых случаях может привести к отсутствию доверия к системе. На практике оценка соответствия требованиям в области эргономичности чаще всего осуществляется лабораториями по проведению сертификационных испытаний в области надежности;
- под защитой информации (ЗИ) понимается выполнение требований в области целостности, доступности и конфиденциальности обрабатываемых данных. Разработка нормативных документов и организация работ по оценке соответствия

²⁶ Составлено автором.

в этой области осуществляется уполномоченными государственными регуляторами с привлечением профильных организаций;

- технологическая независимость (ТН) — аспект доверия, особенно актуализировавшийся в современных условиях и характеризующийся уровнем независимости владельцев СИИ при выполнении ими работ по ремонту систем, их тиражированию и модификации. Усилия в этой области носят универсальный характер и сводятся к сокращению зависимости от зарубежных технологий;
- функциональная корректность (ФК) — аспект, в наибольшей степени специфичный для СИИ, реализованных на алгоритмах МО²⁷. Нормативная база и система оценки соответствия в этой области только развиваются, специализированные регуляторы — отсутствуют. Основной способ подтверждения соответствия требованиям в области ФК СИИ на основе алгоритмов МО — репрезентативное тестирование, обеспечение ФК систем заключается в управлении требованиями к целостности и конфиденциальности информационных компонент СИИ²⁸.

Следует отметить различия в характеристиках надежности и ФК систем ИИ. Так, понятие надежности тесно связано с явлением отказа, т.е. события, заключающегося в переходе объекта в полностью или частично неработоспособное состояние²⁹. Надежность программных средств определяется частотой отказов, временем восстановления после отказа, средним временем наработки на отказ и другими показателями³⁰. Некорректная обработка входных данных в СИИ, приводящая к выдаче ошибочных результатов обработки, строго говоря, отказом не является. Отметим также, что надежность объекта снижается в процессе эксплуатации в результате его изменения (износа, истощения ресурсов и т.п.), что в общем

²⁷ Гарбук С.В. Оценка качества систем искусственного интеллекта: особенности проведения и нормативно-техническая база. Контроль качества продукции. 2024. № 1. С. 20–26.

²⁸ Гарбук С.В. Функциональность и безопасность систем искусственного интеллекта: качество данных. Открытые системы. СУБД. 2024. № 1. С. 18–22. [Электронный ресурс]. URL: <https://www.osp.ru/os/2024/01/13058259> (дата обращения: 21.04.2024). DOI: 10.51793/OS.2024.95.90.004

²⁹ ГОСТ Р 27.102-2021 Надежность в технике. Надежность объекта. Термины и определения.

³⁰ Липаев В.В. Сертификация программных средств. М.: СИНТЕГ, 2010. 344 с.

случае неприменимо к ПО³¹. Кроме того, к основным факторам надежности принято относить причины, также не связанные с особенностями используемых в СИИ алгоритмов обработки данных, включая³²:

- конструктивные причины: надежность элементной базы, неправильный выбор элементов, неудачное схемно-компоновочное решение, недостаточная унификация элементов, недостаточная отработка технологий на этапах испытаний;
- производственные причины: нарушение качества материалов, недостаточный контроль входных параметров, недостаточная отработка технологии производства и сборки устройств, общая низкая культура производства;
- эксплуатационные причины: низкая квалификация технического персонала, низкая эффективность контрольно-проверочной аппаратуры, нарушение условий эксплуатации;
- организационные причины: отсутствие требований по поддержанию заданных показателей надежности, несоответствие заводских испытаний реальным условиям эксплуатации, неритмичность эксплуатации.

В результате понятие доверия к СИИ может быть сведено к четырем независимым составляющим, выполнение каждой из которых является обязательным для достижения общего доверия к системе:

$$\begin{aligned} Trust(s, g, W(E)) = \\ = Trust_{HЭ}(s, W_{HЭ}(E)) \cap Trust_{ЗИ}(s, W_{ЗИ}(E)) \cap \\ Trust_{ТН}(s) \cap Trust_{ФК}(s, g, W_{ФК}(E)), \end{aligned} \quad (3)$$

где $Trust_{ХХ}$ — составляющая доверия, связанная с соответствующим аспектом; $W_{HЭ}$ — условия эксплуатации, влияющие на техническую надежность и уровень эргономичности СИИ; $W_{ИБ}$ — условия, влияющие на защищенность информации в системе — модель угроз информационной безопасности; $W_{ФК}$ — условия внешней среды, включая характеристики операционного контекста, влияющие на функциональные возможности СИИ.

Для каждой из компонент доверия справедливо выражение, связывающее уровень доверия со значением интегрального пока-

³¹ ГОСТ Р ИСО/МЭК 25010-2015. Информационные технологии. Системная и программная инженерия. Требования и оценка качества систем и программного обеспечения (SQuARE). Модели качества систем и программных продуктов. Дата введения — 2016-06-01.

³² Андреев А.В., Яковлев В.В., Короткая Т.Ю. Теоретические основы надежности технических систем. СПб.: Изд-во Политехн. ун-та, 2018. 164 с.

зателя качества (2) с учетом выбора релевантных частных показателей качества в соответствии с табл. 1. Наиболее специфичным, учитывающим особенности реализации в СИИ алгоритмов МО, является аспект ФК $Trust_{\Phi K}(s, g, W_{\Phi K}(E))$, при подтверждении которого, в отличие от других аспектов необходимо учитывать вид решаемой задачи по интеллектуальной обработке данных.

Заключение

Таким образом, в работе показано, что понятие «доверие к СИИ» приобретает конструктивный, практически значимый смысл при определении этого понятия как бинарной характеристики, указывающей на возможность применения системы для решения конкретной прикладной задачи ИИ в определенных условиях. Доверие к СИИ возникает в том случае, если ее интегральный показатель качества, определенный как степень соответствия системы совокупности установленных требований, превышает заданный критериальный порог. При этом обоснование целостного перечня аспектов доверия может быть выполнено на основе анализа характеристик качества систем. Такой подход позволяет сформировать модель доверия, включающую четыре основных независимых аспекта доверия: техническая надежность и эргономичность, уровень защиты информации, уровень технологической независимости и функциональная корректность. Доверие к системе в целом наступает лишь в том случае, если качество по каждому из четырех аспектов соответствует установленным требованиям.

Литература

Андреев А.В., Яковлев В.В., Короткая Т.Ю. Теоретические основы надежности технических систем. СПб.: Изд-во Политехн. ун-та, 2018. 164 с.

Гарбук С.В. Особенности применения понятия «доверие» в области искусственного интеллекта // Искусственный интеллект и принятие решений. 2020. № 3. С. 15–21.

Гарбук С.В. Оценка качества систем искусственного интеллекта: особенности проведения и нормативно-техническая база // Контроль качества продукции. 2024. № 1. С. 20–26.

Гарбук С.В. Функциональность и безопасность систем искусственного интеллекта: качество данных // Открытые системы. СУБД. 2024. № 1. С. 18–22. DOI: 10.51793/OS.2024.95.90.004

Липаев В.В. Сертификация программных средств. М.: СИНТЕГ, 2010. 344 с.

Allen G., Chan T. Artificial Intelligence and National Security. A study on behalf of Dr. Jason Matheny, Director of the U.S. Intelligence Advanced Research Projects Activity (IARPA). Cambridge, Belfer Center Study, 2017.

Garbuk S.V. Intellimetry as a way to ensure AI trustworthiness // The Proceedings of the 2018 International Conference on Artificial Intelligence Applications and Innovations (IC-AIAI). Limassol, Cyprus, 6-10.10.2018. C.27-30. DOI: 10.1109/IC-AIAI.2018.00012

Hall P. Machine Learning for High-Risk Applications. First Edition. Techniques for Responsible AI. O'Reilly Media, Inc, 2022. 112 p.

ИНФОРМАЦИЯ ОБ АВТОРЕ:

Гарбук Сергей Владимирович — к.т.н., с.н.с., Всероссийский институт научной и технической информации Российской академии наук (ВИНИТИ РАН), Технический комитет по стандартизации № 164 «Искусственный интеллект» (ТК 164), Москва, Российская Федерация; *e-mail*: garbuk@list.ru

ABOUT THE AUTHOR:

Sergey V. Garbuk — PhD (Engineering), senior researcher, All-Russian Institute for Scientific and Technical Information of the Russian Academy of Sciences, Technical Committee for Standardization No. 164 'Artificial Intelligence' (TC 164), Moscow, Russian Federation; *e-mail*: garbuk@list.ru