ФЕДЕРАЛЬНОЕ АГЕНТСТВО

ПО ТЕХНИЧЕСКОМУ РЕГУЛИРОВАНИЮ И МЕТРОЛОГИИ

НАЦИОНАЛЬНЫЙ СТАНДАРТ РОССИЙСКОЙ ФЕДЕРАЦИИ

ГОСТ Р 59276— 2020

Системы искусственного интеллекта

СПОСОБЫ ОБЕСПЕЧЕНИЯ ДОВЕРИЯ

Общие положения

Издание официальное

Москва
Стандартинформ
2021

Предисловие

1 РАЗРАБОТАН Акционерным обществом «Всероссийский научно-исследовательский институт сертификации» (АО «ВНИИС»), Обществом с ограниченной ответственностью «ТВпортал» (ООО «ТВпортал»)

2 ВНЕСЕН Техническим комитетом по стандартизации ТК 164 «Искусственный интеллект»

3 УТВЕРЖДЕН И ВВЕДЕН В ДЕЙСТВИЕ Приказом Федерального агентства по техническому регулированию и метрологии от 23 декабря 2020 г. № 1371-ст

4 ВВЕДЕН ВПЕРВЫЕ

Правила применения настоящего стандарта установлены в статье 26 Федерального закона от 29 июня 2015 г. № 162-ФЗ «О стандартизации в Российской Федерации». Информация об изменениях к настоящему стандарту публикуется в ежегодном (по состоянию на 1 января текущего года) информационном указателе «Национальные стандарты», а официальный текст изменений и поправок — в ежемесячном информационном указателе «Национальные стандарты». В случае пересмотра (замены) или отмены настоящего стандарта соответствующее уведомление будет опубликовано в ближайшем выпуске ежемесячного информационного указателя «Национальные стандарты». Соответствующая информация, уведомление и тексты размещаются также в информационной системе общего пользования — на официальном сайте Федерального агентства по техническому регулированию и метрологии в сети Интернет (www.gost.ru)

Содержание

Введение

Доверие к системам искусственного интеллекта является важнейшим условием, определяющим возможность применения этих систем при решении ответственных задач обработки данных. Примерами таких задач являются поддержка принятия врачебных решений, беспилотное управление транспортными средствами и некоторые другие, ошибки при решении которых могут привести к тяжким последствиям, связанным с угрозой для жизни и здоровья людей, серьезным экономическим и экологическим ущербом.

В настоящем стандарте:

- определено понятие доверия к системам искусственного интеллекта;
- приведена классификация факторов, влияющих на качество и способность систем искусственного интеллекта вызывать доверие на стадиях жизненного цикла;
- формализована взаимосвязь качества и способности систем искусственного интеллекта вызывать доверие;
- приведена классификация основных способов обеспечения доверия к системам искусственного интеллекта.

НАЦИОНАЛЬНЫЙ СТАНДАРТ РОССИЙСКОЙ ФЕДЕРАЦИИ

Системы искусственного интеллекта

СПОСОБЫ ОБЕСПЕЧЕНИЯ ДОВЕРИЯ

Общие положения

Artificial intelligence systems. Methods for ensuring trust. General

Дата введения — 2021—03—01

1 Область применения

В настоящем стандарте рассматриваются вопросы обеспечения доверия к системам искусственного интеллекта со стороны потребителей результатов работы этих систем и, при необходимости, со стороны организаций, ответственных за регулирование вопросов создания и применения систем искусственного интеллекта, на основе подтверждения их качества, включая:

- виды существенных характеристик систем искусственного интеллекта, подтверждение значений которых установленным требованиям обеспечивает доверие к этим системам;
- виды факторов, приводящих к снижению качества систем искусственного интеллекта;
- особенности соотношения понятий «качество систем искусственного интеллекта» и доверия к этим системам;
- особенности обеспечения доверия на стадиях жизненного цикла систем искусственного интеллекта, включая стадии создания и эксплуатации этих систем;
- вопросы обеспечения доверия на различных уровнях архитектуры систем искусственного интеллекта.

Стандарт распространяется на системы искусственного интеллекта, обеспечивающие решение конкретных практически значимых задач, и не может быть использован для систем «сильного» или «общего» искусственного интеллекта.

2 Нормативные ссылки

В настоящем стандарте использованы нормативные ссылки на следующие стандарты:

ГОСТ 27.002—2015 Надежность в технике (ССНТ). Термины и определения

ГОСТ 34.601 Информационная технология. Комплекс стандартов на автоматизированные системы. Автоматизированные системы. Стадии создания

ГОСТ 34003—90 Информационная технология. Комплекс стандартов на автоматизированные системы. Термины и определения

ГОСТ Р ИСО 9000 Системы менеджмента качества. Основные положения и словарь

ГОСТ Р ИСО/МЭК 9126 Информационная технология. Оценка программной продукции. Характеристика качества и руководства по их применению

ГОСТ Р ИСО/МЭК 12207 Информационная технология. Системная и программная инженерия. Процессы жизненного цикла программных средств

ГОСТ Р 53622—2009 Информационные технологии (ИТ). Информационно-вычислительные системы. Стадии и этапы жизненного цикла, виды и комплектность документов

ГОСТ Р 57194.1—2016 Трансфер технологий. Общие положения

Издание официальное

Примечание — При пользовании настоящим стандартом целесообразно проверить действие ссылочных стандартов в информационной системе общего пользования — на официальном сайте Федерального агентства по техническому регулированию и метрологии в сети Интернет или по ежегодному информационному указателю «Национальные стандарты», который опубликован по состоянию на 1 января текущего года, и по выпускам ежемесячного информационного указателя «Национальные стандарты» за текущий год. Если заменен ссылочный стандарт, на который дана недатированная ссылка, то рекомендуется использовать действующую версию этого стандарта с учетом всех внесенных в данную версию изменений. Если заменен ссылочный стандарт, на который дана датированная ссылка, то рекомендуется использовать версию этого стандарта с указанным выше годом утверждения (принятия). Если после утверждения настоящего стандарта в ссылочный стандарт, на который дана датированная ссылка, внесено изменение, затрагивающее положение, на которое дана ссылка, то это положение рекомендуется применять без учета данного изменения. Если ссылочный стандарт отменен без замены, то положение, в котором дана ссылка на него, рекомендуется применять в части, не затрагивающей эту ссылку.

3 Термины и определения

В настоящем стандарте применены следующие термины с соответствующими определениями.

3.1

автоматизированная система: Система, состоящая из персонала и комплекса средств автоматизации его деятельности, реализующая информационную технологию выполнения установленных функций.

[ГОСТ 34003—90, статья 1.1]

3.2

вычислительные средства (средства вычислительной техники): Технические средства, непосредственно осуществляющие обработку данных.

[ГОСТ Р 53622—2009, статья 3.4]

3.3 **доверие к системе искусственного интеллекта:** Уверенность потребителя, и при необходимости, 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.

3.4 **доверенная система искусственного интеллекта:** Система искусственного интеллекта, в отношении которой потребитель и, при необходимости, организации, ответственные за регулирование вопросов создания и применения систем искусственного интеллекта, проявляют доверие.

3.5 **информационная технология,** ИТ: Методы, способы, приемы и процессы обработки (сбора, накопления, ввода-вывода, приема-передачи, хранения, поиска, регистрации, преобразования, предоставления, отображения, распространения и уничтожения) информации с применением программного обеспечения и аппаратных средств.

3.6 **искусственный интеллект,** ИИ: Способность технической системы имитировать когнитивные функции человека (включая самообучение и поиск решений без заранее заданного алгоритма) и получать при выполнении конкретных практически значимых задач обработки данных результаты, сопоставимые, как минимум, с результатами интеллектуальной деятельности человека.

3.7

надежность: Свойство объекта сохранять во времени способность выполнять требуемые функции в заданных режимах и условиях применения, технического обслуживания, хранения и транспортирования.

[ГОСТ 27.002—2015, статья 3.1.5]

3.8 **объяснимость** (explainability): Свойство системы искусственного интеллекта, заключающееся в возможности представления причин, приводящих к тому или иному решению системы, в виде, понятном человеку.

3.9 **показатель качества системы искусственного интеллекта**: Степень соответствия представительного набора существенных (значимых) характеристик системы искусственного интеллекта требованиям, то есть потребностям или ожиданиям, которые установлены, обычно предполагаются или являются обязательными для этой системы.

3.10 **понятность** (transparency): Свойство системы искусственного интеллекта, заключающееся в возможности открытого, исчерпывающего, доступного, четкого и понятного представления информации.

3.11 **предвзятость, необъективность** (bias): Свойство системы искусственного интеллекта, заключающееся в принятии ошибочных решений, связанных со статистической смещенностью обучающей выборки исходных данных.

3.12 **предсказуемость** (predictability). Свойство системы искусственного интеллекта, заключающееся в способности принимать решения ожидаемым (естественным, приемлемым) для человека способом.

3.13 **представительный набор существенных характеристик:** Минимально необходимая и достаточная совокупность характеристик системы искусственного интеллекта, позволяющая потребителю, организациям, ответственным за регулирование вопросов создания и применения систем искусственного интеллекта, или любой другой заинтересованной стороне достоверно оценивать качество системы при решении конкретной прикладной задачи.

3.14

программное обеспечение: Упорядоченная последовательность инструкций (кодов) для вычислительного средства, находящаяся в памяти этого средства и представляющая собой описание алгоритма управления вычислительными средствами и действий с данными.

[ГОСТ Р 53622—2009, статья 3.8]

3.15 **сильный (общий) искусственный интеллект:** Способность технической системы, подобно человеку, мыслить, взаимодействовать, адаптироваться к изменяющимся условиям и решать другие задачи в области обработки информации, ассоциирующиеся с естественным интеллектом человека.

3.16 **система искусственного интеллекта:** Техническая система, в которой используются технологии искусственного интеллекта и обладающая искусственным интеллектом.

3.17 **существенные (значимые) характеристики системы искусственного интеллекта:** Характеристики системы искусственного интеллекта, определяющие его качество при решении конкретной прикладной задачи, подтверждение соответствия которых установленным требованиям может быть выполнено потребителем системы, организациями, ответственными за регулирование вопросов создания и применения систем искусственного интеллекта, или любой другой заинтересованной стороной.

Примечание — Характеристики систем искусственного интеллекта, подтверждение соответствия которых установленным требованиям может быть выполнено исключительно разработчиком системы, не относятся к существенным.

3.18

техническая система: Целостная совокупность конечного числа взаимосвязанных материальных объектов, имеющая последовательно взаимодействующие сенсорную и исполнительную функциональные части, модель их предопределенного поведения в пространстве равновесных устойчивых состояний и способная при нахождении хотя бы в одном из них (целевом состоянии) самостоятельно в штатных условиях выполнять предусмотренные ее конструкцией потребительские функции.

[ГОСТ Р 57194.1—2016, статья 3.8]

3.19

технические средства: Аппаратные и программные средства, используемые для сбора, обработки, хранения, манипуляции и выдачи данных.

[ГОСТ Р 53622—2009, статья 3.12]

3.20 **технологии искусственного интеллекта**: Комплекс технологических решений, позволяющий имитировать когнитивные функции человека (включая самообучение и поиск решений без заранее заданного алгоритма) и получать результаты, сопоставимые, как минимум, с результатами интеллектуальной деятельности человека при решении задач компьютерного зрения, обработки естественного языка, распознавания и синтеза речи, поддержки принятия решений и других практически значимых задач обработки данных.

4 Сокращения

В настоящем стандарте использованы следующие сокращения:

АС — автоматизированная система;

ЖЦ — жизненный цикл;

ИНС — искусственная нейронная сеть;

ИТ — информационная технология;

ПО — программное обеспечение;

ПС — программное средство;

ТС — транспортное средство.

5 Существенные характеристики систем искусственного интеллекта

Качество объекта (продукта или услуги) является комплексным показателем, определяющим потребительские свойства объекта (ГОСТ Р ИСО 9000).

Понятие качества существует в конкретном прикладном контексте и может быть определено лишь для систем искусственного интеллекта, обеспечивающих решение конкретных практически значимых задач. Приведенное определение качества не распространяется на системы «сильного» или «общего» искусственного интеллекта, претендующих на повторение естественных интеллектуальных способностей человека вне зависимости от решаемой прикладной задачи.

Существенные характеристики системы искусственного интеллекта (ИИ) могут быть присущими или присвоенными, но при оценке качества учитываются только присущие (ГОСТ Р ИСО 9000). При этом существуют различные классы характеристик, такие как:

а) физические (например, механические, электрические, химические или биологические характеристики);

б) органолептические (например, связанные с запахом, осязанием, вкусом, зрением, слухом);

в) этические;

г) характеристики, связанные со временем;

д) эргономические;

е) функциональные (например, максимальная скорость движения беспилотного ТС, вероятность ошибок первого и второго рода);

ж) информационной безопасности.

Представительный набор существенных характеристик определяется решаемой прикладной задачей и условиями применения системы ИИ. Например, функциональные характеристики для различных систем ИИ могут быть выбраны следующим образом:

- для систем распознавания речи — величина уровня пословной ошибки, фактор реального времени преобразования речи в текст на конкретном сервере обработки данных и другие характеристики;

- для систем дешифрирования аэрокосмических изображений — точность обнаружения объектов, полнота обнаружения объектов, точность локализации объекта, вероятность правильной классификации объекта и другие характеристики;

- для систем беспилотного управления уровней 1—5 ТС — вероятность дорожно-транспортного происшествия при автономном управлении, уровень комфорта управления движением для пассажиров ТС, отношение времени автономного управления к времени вынужденного перехода на ручное управление за все время движения ТС и другие характеристики.

При выборе представительного набора существенных характеристик системы ИИ целесообразно руководствоваться следующими принципами (ГОСТ Р ИСО/МЭК 9126):

- достаточность набора характеристик для принятия решения о возможности использования системы ИИ при решении конкретной прикладной задачи;

- простота и возможность измерения значений характеристик;

- отсутствие перекрытия между используемыми характеристиками;

- соответствие установившимся понятиям и терминологии;

- возможность последующего уточнения и детализации характеристик.

Существенные характеристики и субхарактеристики систем ИИ (см. таблицу 1) в зависимости от возможностей их измерения, представления и интерпретации могут быть определены на различных шкалах, включая шкалу наименований, порядковую шкалу, интервальную шкалу и шкалу отношений.

Таблица 1 — Существенные характеристики систем искусственного интеллекта (ГОСТ Р ИСО/МЭК 9126)

Характеристика	Субхарактеристика
1 Функциональные возможности	1.1 Пригодность 1.2 Корректность (правильность) 1.3 Способность к взаимодействию 1.4 Согласованность 1.5 Защищенность
2 Надежность	2.1 Стабильность 2.2 Устойчивость к ошибке 2.3 Восстанавливаемость
3 Эффективность	3.1 Характер изменения во времени 3.2 Характер изменения ресурсов
4 Практичность	4.1 Понятность 4.2 Изучаемость 4.3 Простота использования
5 Сопровождаемость	5.1 Анализируемость 5.2 Изменяемость 5.3 Устойчивость 5.4 Тестируемость
6 Мобильность	6.1 Адаптируемость 6.2 Простота внедрения 6.3 Соответствие 6.4 Взаимозаменяемость

Требования к представительному набору существенных характеристик задаются для конкретных условий эксплуатации системы ИИ. Описание условий эксплуатации может включать, например:

- требования к сенсорам и качеству (точности, полноте, достоверности) исходных данных, поступающих на вход системы ИИ;
- требования к информационным характеристикам других устройств, с которыми взаимодействует система ИИ;
- требования к области применения, в котором должна быть обеспечена работа системы ИИ, включая ограничения на возможные преднамеренные и непреднамеренные искажения исходных данных;
- требования к квалификации персонала, эксплуатирующего систему ИИ и др.

6 Факторы снижения качества на стадиях жизненного цикла систем искусственного интеллекта

ЖЦ системы ИИ может быть определен последовательностью стадий и этапов создания и эксплуатации АС (ГОСТ 34.601), ПС (ГОСТ Р ИСО/МЭК 12207), этапов ЖЦ систем ИИ или другим рациональным способом (см. таблицу 2).

Таблица 2 — Стадии и этапы жизненного цикла системы искусственного интеллекта

Стадии ЖЦ АС (ГОСТ 34.601)	Этапы ЖЦ АС (ГОСТ 34.601)	Стадии ЖЦ ПС (ГОСТ Р ИСО/МЭК 12207)	Этапы ЖЦ системы ИИ
Создание системы ИИ			
Формирование требований к АС	Обследование объекта и обоснование необходимости создания АС. Формирование требований пользователя к АС. Оформление отчета о выполненной работе и заявки на разработку АС (тактико-технического задания)	Концепция ПС	Определение облика системы ИИ

Окончание таблицы 2

Стадии ЖЦ АС (ГОСТ 34.601)	Этапы ЖЦ АС (ГОСТ 34.601)	Стадии ЖЦ ПС (ГОСТ Р ИСО/МЭК 12207)	Этапы ЖЦ системы ИИ
Разработка концепции АС	Изучение объекта. Проведение необходимых научно-исследовательских работ. Разработка вариантов концепции АС, удовлетворяющего требованиям пользователя. Оформление отчета о выполненной работе		
Техническое задание	Разработка и утверждение технического задания на создание АС	Разработка ПС	Разработка
Эскизный проект	Разработка предварительных проектных решений по системе и ее частям. Разработка документации на АС и ее части		
Технический проект	Разработка проектных решений по системе и ее частям. Разработка документации на АС и ее части. Разработка и оформление документации на поставку изделий для комплектования АС и (или) технических требований (технических заданий) на их разработку. Разработка заданий на проектирование в смежных частях проекта объекта автоматизации		
Рабочая документация	Разработка рабочей документации на систему и ее части. Разработка или адаптация программ	Производство ПС	
Эксплуатация системы ИИ			
Ввод в действие	Подготовка объекта автоматизации к вводу АС в действие. Подготовка персонала. Комплектация АС поставляемыми изделиями (ПС и аппаратными средствами, программно-аппаратными комплексами, информационными изделиями). Строительно-монтажные работы. Пусконаладочные работы. Проведение предварительных испытаний. Проведение опытной эксплуатации. Проведение приемочных испытаний	Применение ПС по назначению	Внедрение Верификация и валидация
Сопровождение АС	Выполнение работ в соответствии с гарантийными обязательствами. Послегарантийное обслуживание	Поддержка ПС	Эксплуатация и сопровождение
		Прекращение применения ПС	Вывод из эксплуатации

При описании жизненного цикла системы ИИ предпочтительной является модель ЖЦ, ранее выбиравшаяся организациями, участвующими в разработке и применении системы, в других проектах. Такой подход к выбору стадий и этапов ЖЦ обеспечивает наиболее эффективное управление процессом создания и применения системы ИИ на протяжении всего ЖЦ (ГОСТ Р ИСО/МЭК 12207).

Следует выделять две группы стадий и этапов, различающихся обязательными субъектами управления ЖЦ системы ИИ.

1) создание системы ИИ — управление ЖЦ осуществляется с обязательным участием разработчика системы, участие потребителя системы является опциональным;

2) эксплуатация системы ИИ — управление ЖЦ осуществляется с обязательным участием потребителя системы, участие разработчика системы является опциональным.

На каждой стадии и каждом этапе ЖЦ существуют факторы (причины), приводящие к снижению качества системы ИИ (факторы снижения качества системы ИИ). Каждый фактор снижения качества

связан с возможными отклонениями одной или нескольких существенных характеристик системы ИИ от установленных требований.

В зависимости от стадии ЖЦ, на которой может быть устранена та или иная причина снижения качества, факторы разделяют на две группы:

1) факторы снижения качества на стадии создания системы ИИ;

2) факторы снижения качества на стадии эксплуатации системы ИИ.

Примеры факторов для последовательных стадий жизненного цикла системы ИИ приведены в таблице 3.

Таблица 3 — Факторы снижения качества на стадиях создания и эксплуатации систем искусственного интеллекта

Стадия ЖЦ	Фактор снижения качества системы ИИ
Создание системы ИИ	
Концепция	Недостаточная полнота выбранного набора функциональных характеристик системы ИИ (прикладных характеристик, характеристик безопасности, надежности и других), не позволяющая считать выбранный набор характеристик представительным
Разработка	Недостаточная представительность обучающей выборки, использованной при создании системы ИИ. Смещенность обучающей выборки, способная привести к предвзятости (необъективности) результатов работы системы ИИ. Неоптимальность используемой модели данных. Недостаточный уровень унификации и низкая интероперабельность разрабатываемой системы
Производство	Недостаточная надежность создаваемой системы ИИ. Чрезмерная стоимость владения системой ИИ. Недостаточная понятность, объяснимость, предсказуемость и др. Недостаточная защищенность информации о модели данных
Эксплуатация системы ИИ	
Применение по назначению	Применение системы ИИ не по назначению. Недостаточная представительность выборки, используемой при тестировании системы ИИ. Недостаточная периодичность тестирования системы ИИ. Отсутствие средств автоматического самотестирования после каждого обучения, дообучения системы ИИ. Недостаточная защищенность информации о функционировании системы ИИ. Недостаточная защищенность информации о модели данных, используемой в системе ИИ. Недостаточная защищенность обрабатываемых персональных данных
Поддержка	Утрата актуальности модели данных
Прекращение применения	Нарушение конфиденциальности персональных данных при выводе системы ИИ из эксплуатации

Факторы снижения качества могут быть связаны с естественными (непреднамеренное снижение качества) или искусственными (преднамеренное снижение качества) причинами. Примерами преднамеренного снижения качества, специфичными для систем ИИ, реализованных на ИНС, являются:

- на стадии создания системы — наличие преднамеренных искажений в обучающей выборке системы распознавания изображений, приводящих к ошибкам в работе системы распознавания, вызванным специальными, заранее определенными искажениями в исходных данных, включая «состязательные» атаки;

- на стадии эксплуатации системы — отсутствие достоверных и представительных оценок устойчивости системы распознавания изображений к воздействию преднамеренных «состязательных» атак, приводящее к неустойчивой работе системы в процессе ее эксплуатации.

Примерами непреднамеренного снижения качества систем ИИ являются:

- на стадии создания системы ИИ — использование статистически смещенной обучающей выборки, приводящей к появлению «предвзятостей» в результатах работы системы ИИ;

- на стадии эксплуатации системы — нарушение конфиденциальности обрабатываемых данных в условиях, когда уровень конфиденциальности данных существенно и неконтролируемо возрос в процессе эксплуатации системы ИИ вследствие накопления и обобщения информации.

7 Доверие и качество систем искусственного интеллекта

Проверка доверия к системе ИИ обеспечивается подтверждением соответствия представительного набора существенных характеристик системы ИИ требованиям (рисунок 1), установленным:

- разработчиком системы ИИ в том случае, если существует возможность подтверждения соответствия этим требованиям любой заинтересованной стороной, а не только самим разработчиком системы ИИ. Такие требования разработчика являются открытыми требованиями в отличие от внутренних требований, подтверждение соответствия которым может быть выполнено только самим разработчиком системы ИИ;

- потребителем системы ИИ. Подтверждение соответствия потребительским требованиям осуществляется в процессе испытаний системы ИИ;

- организацией, ответственной за регулирование вопросов создания и применения систем ИИ в соответствии с принятыми национальными нормами (регулятором). Данные требования являются опциональными и, как правило, устанавливаются в том случае, если некорректная работа системы ИИ может привести к угрозам безопасности людей, окружающей природной среды, материальных и нематериальных активов. В этом случае подтверждение соответствия требованиям, установленным регулятором, осуществляется в ходе сертификации системы ИИ.

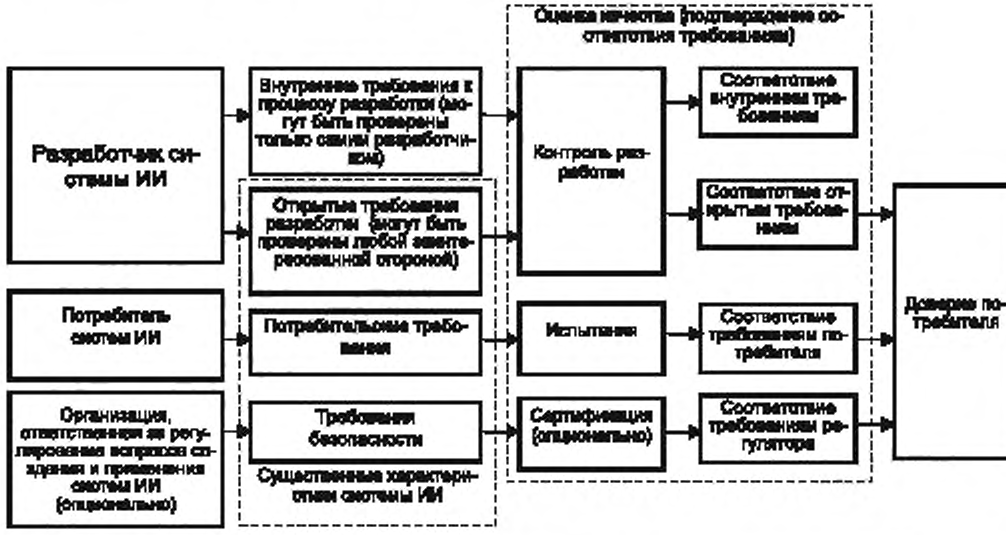

Рисунок 1 — Процессы обеспечения доверия и оценки качества (подтверждения соответствия требованиям) систем искусственного интеллекта

Способность вызывать доверие может изменяться со временем: это свойство может быть приобретено и потеряно системой ИИ на одной из стадий ее ЖЦ. В связи с этим целесообразно определять контрольные точки, в которых будет осуществляться проверка доверия к системе ИИ. Такие контрольные точки можно устанавливать регулярно на протяжении всего ЖЦ системы ИИ (периодический контроль), в моменты, когда система ИИ используется для принятия решений (итоговый контроль) и/или в другие моменты.

8 Способы обеспечения доверия на стадиях жизненного цикла систем искусственного интеллекта

Процедура подтверждения доверия включает:

- выбор достаточного (представительного) набора существенных характеристик системы ИИ (разработчиком, потребителем и, при необходимости, организацией, ответственной за регулирование вопросов создания и применения систем ИИ);

- установление требований к представительному набору существенных характеристик системы ИИ (открытых требований разработчика, потребительских требований и, при необходимости, требований безопасности, установленных организацией, ответственной за регулирование вопросов создания и применения систем ИИ);

- организацию процедур подтверждения качества системы ИИ, то есть подтверждения соответствия представительного набора существенных характеристик системы ИИ установленным требованиям;

- реализацию мероприятий по обеспечению соответствия представительного набора существенных характеристик системы ИИ установленным требованиям (доведению качества системы ИИ до требуемого уровня за счет устранения причин, приводящих к снижению качества), — способов обеспечения доверия.

Способы обеспечения доверия, направленные на устранение факторов, приводящих к снижению качества систем ИИ, могут быть реализованы на последовательных стадиях ЖЦ разработчиками, потребителями систем ИИ или третьей стороной (например, органом по сертификации) по инициативе разработчиков или потребителей систем (таблица 4).

Способы обеспечения доверия на стадии разработки заключаются в обеспечении соответствия открытым требованиям разработчика системы ИИ (см. рисунок 1). Способы обеспечения доверия на стадии эксплуатации заключаются в обеспечении соответствия требованиям потребителя системы ИИ и требованиям организации, ответственной за регулирование вопросов создания и применения систем ИИ в соответствии с принятыми национальными нормами (опционально).

Таблица 4 — Способы обеспечения доверия к системам искусственного интеллекта на соответствующих стадиях жизненного цикла

Фактор снижения качества системы ИИ	Способ обеспечения доверия к системам ИИ
Создание системы ИИ	
Недостаточная полнота выбранного перечня существенных характеристик системы ИИ (прикладных характеристик, характеристик безопасности, надежности и других)	Выбор представительного набора существенных характеристик системы и корректных правил их определения
Недостаточная представительность обучающей выборки, использованной при создании системы ИИ	Формирование представительной обучающей выборки (например, для биометрических систем ИИ) (ГОСТ Р 57194.1)
Статистическая смещенность обучающей выборки, способная привести к предвзятости (необъективности) результатов работы системы	Очистка набора данных различными способами. Статистический анализ наборов исходных данных и оценка их представительности и качества. Кросс-валидация выборки, полученной при разметке данных людьми. Непосредственная корректировка модели. Наложение ограничений на допустимую область применения системы ИИ
Неоптимальность используемой модели данных	Разработка оптимальной модели данных
Недостаточная надежность создаваемой системы ИИ. Чрезмерная стоимость владения системой ИИ	Использование рациональной ИТ-инфраструктуры, обеспечивающей надежную реализацию алгоритмов обработки данных при приемлемой стоимости владения системой ИИ
Недостаточная понятность, объяснимость, предсказуемость и др.	Использование интеллектуальных алгоритмов обработки данных, обеспечивающих принятие системой объяснимых, предсказуемых и так далее решений

Окончание таблицы 4

Фактор снижения качества системы ИИ	Способ обеспечения доверия к системам ИИ
Недостаточная защищенность информации о модели данных	Принятие эффективных мер по защите информации о модели данных на стадии разработки системы ИИ
Эксплуатация системы ИИ	
Применение системы ИИ не по назначению	Соблюдение допустимой области применения системы ИИ
Недостаточная представительность выборки, используемой при тестировании системы ИИ	Выбор представительной тестовой выборки при подтверждении соответствия системы установленным функциональным требованиям
Недостаточная защищенность информации о модели данных, используемой в системе ИИ. Недостаточная защищенность обрабатываемых персональных данных	Принятие эффективных мер по защите информации на стадии эксплуатации системы ИИ
Утрата актуальности модели данных	Своевременное выявление существенных отклонений в условиях эксплуатации системы ИИ и принятие мер по актуализации модели данных
Нарушение конфиденциальности персональных данных при выводе системы ИИ из эксплуатации	Принятие эффективных мер по защите персональных данных при выводе системы ИИ из эксплуатации

При создании и применении систем ИИ необходимо стремиться к применению всей совокупности способов обеспечения доверия, так как при этом обеспечивается устранение максимального количества факторов, приводящих к снижению качества работы системы. Способы обеспечения доверия на стадиях создания и эксплуатации дополняют друг друга, что объясняется наличием преимуществ и недостатков, присущих способам различных типов (таблица 5).

Таблица 5 — Особенности способов обеспечения доверия на разных стадиях жизненного цикла систем искусственного интеллекта

Тип характеристик	Преимущества	Недостатки
1 Способы обеспечения доверия на стадии создания системы ИИ	Позволяют заблаговременно (на этапах проектирования и разработки) гарантировать определенные свойства системы. Гарантируют высокую повторяемость свойств создаваемых систем	Предполагают достаточно полный доступ к процессу создания системы (основная часть требований разработчика должна быть открытой, то есть подлежащей проверке любой заинтересованной стороной), что может быть затруднительно для разработчика системы. Показательны для разработчика, но, как правило, неинформативны для потребителя системы
2 Способы обеспечения доверия на стадии эксплуатации системы ИИ	Хорошо интерпретируются в терминах потребительских свойств, более понятны потребителю системы	Результаты измерения характеристик не всегда могут быть экстраполированы на реальные условия эксплуатации системы (проблема представительности результатов тестирования)

9 Доверие и уровни архитектуры в системах искусственного интеллекта

В системах ИИ следует выделять три уровня архитектуры:

1) физический уровень — уровень сенсоров и исполнительных устройств систем ИИ, благодаря которым система осуществляет физическое взаимодействие с окружающей средой и объектами;

2) инфраструктурный уровень (уровень информационной инфраструктуры) — уровень, включающий аппаратные средства хранения, обработки и передачи информации, включая облачную инфраструктуру, а также системное ПО;

3) прикладной уровень — уровень прикладного ПО, реализующего алгоритмы интеллектуальной обработки данных.

Для каждого из трех уровней доверия могут быть определены свои существенные характеристики систем ИИ.

Доверие на физическом уровне, как правило, основывается на комбинации требований к надежности, безопасности и функциональности, поскольку существенные характеристики, требования к которым устанавливаются, основаны в данном случае на физических измерениях или тестах. Например, успешно пройденный технический контроль ТС делает ТС и его системы заслуживающими доверия. В этом контексте уровень доверия может быть определен посредством выполнения проверочных мероприятий. Кроме того, некоторые процессы, такие как калибровка датчиков, могут гарантировать правильность измерений и, следовательно, корректность полученных данных.

Доверие на инфраструктурном уровне заключается, как правило, в выполнении требований безопасности ИТ-инфраструктуры, таких как контроль доступа и другие меры для поддержания целостности и доступности системы ИИ, а также обеспечения конфиденциальности обрабатываемых данных.

Доверие на прикладном уровне системы ИИ требует, среди прочего, подтверждения надежности и безопасности ПО. Разработка ПО предполагает реализацию процессов его верификации и валидации. Обеспечение доверия к системам ИИ включает подходы, применимые к обычным информационным системам, но не ограничивается ими. Так, например для систем ИИ, основанных на машинном обучении, способность вызывать доверие подразумевает также непредвзятость (объективность) функционирования системы, что соответствует отсутствию необоснованного смещения формируемых оценок.

Доверие к системе ИИ достигается в том случае, если выполняются требования к представительному набору существенных характеристик систем ИИ на всех трех уровнях. При этом используются способы обеспечения доверия на всех стадиях ЖЦ системы ИИ.

Требования к существенным характеристикам могут предъявляться отдельно для разных уровней архитектуры системы ИИ или комплексно, для системы в целом.

Способы обеспечения доверия к различным уровням архитектуры системы ИИ могут приводить к устранению причин снижения качества на одном или на нескольких уровнях системы ИИ. Например, применение ПО, сертифицированного по требованиям информационной безопасности, обеспечивает устранение факторов снижения качества, связанных с нарушением конфиденциальности обрабатываемых персональных данных, как на уровне ИТ-инфраструктуры, так и на прикладном уровне.

УДК 656.072:681.3:006.354 ОКС 35.240.01

Ключевые слова: технологии искусственного интеллекта, искусственный интеллект, надежность, доверие, способы обеспечения доверия

Редактор *Н.А. Аргунова*
Технический редактор *И.Е. Черепкова*
Корректор *И.А. Королева*
Компьютерная верстка *Л.А. Круговой*

Сдано в набор 24.12.2020. Подписано в печать 18.01.2021. Формат 60×84 1/8. Гарнитура Ариал.
Усл. печ. л. 1,86. Уч.-изд. л. 1,68.
Подготовлено на основе электронной версии, предоставленной разработчиком стандарта

Создано в единичном исполнении во ФГУП «СТАНДАРТИНФОРМ»
для комплектования Федерального информационного фонда стандартов, 117418 Москва, Нахимовский пр-т, д 31, к. 2.
www.gostinfo.ru info@gostinfo.ru