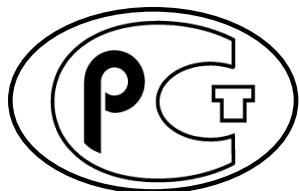

**ФЕДЕРАЛЬНОЕ АГЕНТСТВО
ПО ТЕХНИЧЕСКОМУ РЕГУЛИРОВАНИЮ И МЕТРОЛОГИИ**



**ПРЕДВАРИТЕЛЬНЫЙ
НАЦИОНАЛЬНЫЙ
СТАНДАРТ
РОССИЙСКОЙ
ФЕДЕРАЦИИ**

**ПНСТ
—
2026**

**СИНТЕЗ ДАННЫХ
Основные положения**

Издание официальное

**Москва
Российский институт стандартизации
2026**

Предисловие

1 РАЗРАБОТАН Некоммерческой организацией «Ассоциация больших данных» (НО «АБД»)

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 _____ 202_ г. № _____

Правила применения настоящего стандарта и проведения его мониторинга установлены в ГОСТ Р 1.16—2011 (разделы 5 и 6).

Федеральный орган исполнительной власти в сфере стандартизации собирает сведения о практическом применении настоящего стандарта. Данные сведения, а также замечания и предложения по содержанию стандарта можно направить не позднее чем за 4 мес до истечения срока его действия разработчику настоящего стандарта по адресу: 123112 Москва, Пресненская набережная, д. 10, стр. 2, электронная почта: info@rubda.ru и/или в Федеральное агентство по техническому регулированию и метрологии по адресу: 123112 Москва, Пресненская набережная, д. 10, стр. 2.

В случае отмены настоящего стандарта соответствующая информация будет опубликована в ежемесячном информационном указателе «Национальные стандарты» и также будет размещена на официальном сайте Федерального агентства по техническому регулированию и метрологии в сети Интернет (www.rst.gov.ru)

© Оформление. ФГБУ «Институт стандартизации», 2026

Настоящий стандарт не может быть полностью или частично воспроизведен, тиражирован и распространен в качестве официального издания без разрешения Федерального агентства по техническому регулированию и метрологии

Содержание

1	Область применения
2	Нормативные ссылки.....
3	Термины и определения
4	Основные концепции синтеза данных.....
4.1	Синтез данных с использованием ИИ
4.2	Свойства синтетических данных
4.3	Классификация типов синтетических данных.....
4.4	Основные принципы осуществления синтеза данных
	Библиография

Введение

В современных условиях развития информационных технологий искусственный интеллект (ИИ) становится неотъемлемой частью многих областей деятельности, включая бизнес, науку, медицину и государственное управление, а синтетические данные находят все большее применение в ИИ, статистике и других сферах, требующих использования данных для моделирования и анализа. Рост объемов данных и необходимость их эффективной обработки требуют унификации подходов и терминологии для улучшения взаимодействия между различными участниками рынка обмена данными. В то же время синтетические данные помогают восполнить нехватку реальных данных для обучения моделей и анализа.

В связи с этим возросла актуальность создания стандартов, регулирующего основные понятия и методы работы с синтетическими данными, которые находят применение в различных областях, включая ИИ, статистический анализ и тестирование систем (см. [1]).

ПРЕДВАРИТЕЛЬНЫЙ НАЦИОНАЛЬНЫЙ СТАНДАРТ РОССИЙСКОЙ ФЕДЕРАЦИИ

СИНТЕЗ ДАННЫХ

Основные положения

Data synthesis. Part 1. Terms and definitions of the synthesis process.
General provisions

Срок действия – с __-__-__

до __-__-__

1 Область применения

Настоящий стандарт распространяется на процессы создания полных синтетических данных табличного вида с использованием нейронных сетей.

Настоящий стандарт не распространяется на синтетические данные, создаваемые методами статистического и компьютерного моделирования на основе цифровых моделей реальных сценариев.

Стандарт применим к организациям любых форм собственности и масштабов, использующим технологии искусственного интеллекта (ИИ) и синтетические данные.

Издание официальное

2 Нормативные ссылки

В настоящем стандарте использованы нормативные ссылки на следующие стандарты:

ГОСТ 33707 (ISO/IEC 2382:2015) Информационные технологии.
Словарь

ГОСТ ISO/IEC 29100 Информационные технологии. Методы и средства обеспечения безопасности. Основы защиты персональных данных

ГОСТ Р 50922 Защита информации. Основные термины и определения

ГОСТ Р 51904 Программное обеспечение встроенных систем. Общие требования к разработке и документированию

ГОСТ Р 52292 Информационная технология. Электронный обмен информацией. Термины и определения

ГОСТ Р 59276 Системы искусственного интеллекта. Способы обеспечения доверия. Общие положения

ГОСТ Р 59407 Информационные технологии. Методы и средства обеспечения безопасности. Базовая архитектура защиты персональных данных

ГОСТ Р 59926/ISO/IEC TR 20547-2:2018 Информационные технологии. Эталонная архитектура больших данных. Часть 2. Варианты использования и производные требования

ГОСТ Р 70462.1/ISO/IEC TR 24029-1 Информационные технологии. Интеллект искусственный. Оценка робастности нейронных сетей. Часть 1. Обзор

ГОСТ Р 70889–2023 (ИСО/МЭК 8183:2023) Информационные технологии. Искусственный интеллект. Структура жизненного цикла данных

ГОСТ Р 71476 (ИСО/МЭК 22989:2022) Искусственный интеллект. Концепции и терминология искусственного интеллекта

ГОСТ Р ИСО 8000-2 Качество данных. Часть 2. Словарь

ГОСТ Р ИСО/МЭК 20546 Информационные технологии. Большие данные. Обзор и словарь

ГОСТ Р ИСО 22514-1 Статистические методы. Управление процессами. Часть 1. Общие принципы

ГОСТ Р ИСО/МЭК 24029-2 Искусственный интеллект. Оценка робастности нейронных сетей. Часть 2. Методология использования формальных методов

ПНСТ 840—2023 (ISO/IEC TR 24368:2022) Искусственный интеллект. Обзор этических и общественных аспектов

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

ПНСТ

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 термины по [2], [3], ГОСТ Р 51904, ГОСТ Р 52292, ГОСТ Р 50922, ГОСТ 33707, ГОСТ Р 59407, ГОСТ Р ИСО/МЭК 20546, ГОСТ Р ИСО/МЭК 24029-2, ГОСТ Р 70462.1, ГОСТ Р 71476, а также следующие термины с соответствующими определениями:

3.1 Основные понятия

3.1.1 синтез данных (data synthesis): Процесс создания синтетических данных на основе заданной статистической модели.

3.1.2 синтетические данные (synthetic data): Данные, искусственно созданные для имитации формата и свойств реальных данных, но которые не соответствуют напрямую каким-либо реальным объектам.

Примечания

1 Синтетические данные могут быть использованы для различных целей, включая обучение моделей машинного обучения, тестирование систем, а также обеспечение повышенной конфиденциальности данных.

2 Настоящий стандарт описывает полностью синтетические данные, которые в дополнение к указанным в определении свойствам не являются модификацией исходных данных.

3 Синтетические данные могут не всегда точно воспроизводить все свойства реальных данных, что может ограничивать их применение в определенных сценариях.

3.1.3 система синтеза данных (synthesis system): Совокупность инструментов и алгоритмов, предназначенная для создания синтетических данных на основе заданных моделей и методов.

Примечание — Система синтеза осуществляет управление процессом генерации данных, обеспечивая соответствие синтетических данных статистическим характеристикам исходных данных при одновременном соблюдении требований конфиденциальности.

3.1.4 статистическая модель данных (statistical data model): Математическое представление распределений, зависимостей и других статистических характеристик набора данных, применяемое для анализа, моделирования или генерации данных.

Примечания

1 В контексте синтетических данных под статистической моделью данных понимается частный случай статистических моделей, применяемых к наборам данных. Такие модели широко используются для генерации синтетических данных, так как позволяют воспроизводить ключевые характеристики и распределения реальных данных.

2 Примерами статистических моделей данных являются модели линейной регрессии, модели временных рядов и байесовские сети, которые применяются при синтезе для воспроизведения распределений и взаимосвязей между переменными в наборе данных.

3 Использование статистических моделей в синтезе данных способствует улучшению качества и достоверности синтетических данных, что важно для обеспечения их полезности в различных приложениях, включая машинное обучение и анализ данных.

ПНСТ

4 Статистическая модель данных отличается от общей модели данных, представляющей схему данных, структурированную в базе данных согласно ГОСТ Р ИСО/МЭК 20546.

3.1.5 полностью синтетические данные (fully synthetic data):

Искусственно созданные данные, сохраняющие характеристики и статистические свойства реальных данных, не содержащие объектов реальных данных и не допускающие восстановления исходных данных через статистические особенности в практическом применении.

Примечание — Под статистическими свойствами в контексте настоящего стандарта понимаются параметры распределения, зависимости между переменными, меры вариации, частотные характеристики и иные статистически наблюдаемые свойства выборки, применимые для сравнения и оценки синтетических данных.

3.1.6 уровень доверия к технологиям синтеза данных (level of trust in synthesis technologies): Комплексная оценка гарантий приватности и надежности, обеспечиваемых методиками, алгоритмами и инфраструктурой генерации синтетических данных.

Примечание — Применяется в контексте использования доверенных технологий ИИ и может определяться через вероятность утечки данных.

3.1.7 частично синтетические данные (partially synthetic data):

Данные, созданные путем замены или дополнения части реальных данных искусственно созданными данными, которые сохраняют характеристики и статистические свойства исходных данных.

3.2 Термины в области искусственного интеллекта

3.2.1 вариационный автокодировщик (variational autoencoder): Тип нейронной сети, используемый для генерации новых данных, которые следуют распределению, похожему на исходные данные.

Примечание — Вариационный автокодировщик состоит из энкодера, преобразующего входные данные в скрытое представление (латентное пространство) и декодера, генерирующего новые данные из латентного пространства, восстанавливая их в исходное состояние с учетом вероятностных характеристик.

3.2.2 генеративно-сопоставительная сеть; ГСС (generative adversarial network): Нейронные сети, содержащие один или несколько генераторов, пытающихся создать наиболее репрезентативные для набора данных образцы, и один или несколько дискриминаторов, пытающихся отличить сгенерированные образцы от реальных.

Примечания

1 Компоненты генератора и дискриминатора обучаются вместе, чтобы усовершенствовать внутренние представления сети.

2 ГСС могут быть использованы для выполнения различных заданий (например, классификации), а также для других целей, таких как генерация модельных данных или адаптация к новой области.

3 Генератор создает искусственные данные, а дискриминатор пытается отличить реальные данные от искусственных.

4 ГСС представляет собой важный способ генерации синтетических данных и обладает свойством обратного распространения ошибки, что исключает наличие явной обратной функции восстановления исходных данных, хотя при отсутствии механизмов приватности возможны атаки на реконструкцию.

3.3 Термины в области использования доверенных технологий

3.3.1 бюджет приватности (privacy budget): Концепция в дифференциальной приватности, определяющая количество операций или запросов, которые можно выполнить с данными до того, как вероятность раскрытия информации станет значимой.

Примечание — Бюджет приватности представляет собой количественную меру (например, параметр ϵ), которая расходуется при каждой операции доступа к данным. По мере выполнения последовательных запросов бюджет уменьшается, и при его исчерпании обработка приостанавливается или перенастраивается. Контроль бюджета обеспечивает ограничение совокупного риска раскрытия информации.

3.3.2 вероятность утечки данных (probability of original data leakage): Оценка (вероятностная мера) возможности несанкционированного раскрытия, копирования или распространения исходных данных, характеризующая возможное снижение уровня доверия к используемым технологиям синтеза.

Примечания

1 Вероятность утечки данных может использоваться для оценки степени риска, если синтетические данные создаются не с использованием доверенных технологий или содержат информацию, которую можно сопоставить с реальными данными.

2 Качественные проблемы могут быть связаны с возможностью того, что синтетические данные не полностью отражают характеристики реальных данных, что может привести к ошибочным выводам или неэффективности моделей машинного обучения.

3 Правовые и этические проблемы возникают, если компрометация синтетических данных приводит к их использованию для целей, не соответствующих первоначальному замыслу, или если они нарушают нормативные требования и стандарты обработки данных.

4 Управление негативными последствиями включает в себя внедрение соответствующих доверенных технологий (например, контроля доступа, шифрования данных), а также регулярную оценку и мониторинг качества и соответствия методов синтеза требованиям и общим стандартам безопасности.

3.3.3 гарантии приватности (privacy guarantees): Совокупность технических и организационных мер, включая правовые меры, направленных на снижение риска несанкционированного раскрытия информации, обеспечивающих достижение уровня приватности, соответствующего установленным требованиям при обработке или синтезе данных.

3.3.4 дифференциальная приватность (differential privacy): Формальная модель обеспечения приватности, гарантирующая, что результаты обработки данных остаются практически неизменными при добавлении или удалении любой отдельной записи, тем самым защищая информацию о конкретных субъектах данных независимо от внешних знаний третьей стороны.

Примечания

1 Механизмы, реализующие дифференциальную приватность, могут использовать различные методы, включая добавление калиброванного шума (механизм Лапласа, Гаусса), экспоненциальный механизм, механизм отчета.

2 Формально, механизм M является (ϵ, δ) -дифференциально приватным, если для всех $S \subseteq \text{Range}(M)$ и для всех соседних наборов данных D и D' : $\Pr[M(D) \in S] \leq e^\epsilon \cdot \Pr[M(D') \in S] + \delta$,

где M — рандомизированный механизм обработки данных;

ϵ — параметр приватности, определяющий верхнюю границу различия в распределениях выходных данных;

δ — параметр, допускающий малую вероятность нарушения гарантии приватности;

D, D' — соседние наборы данных, отличающиеся не более чем одной записью;

ПНСТ

$Range(M)$ — множество возможных выходных значений механизма M ;

S — произвольное подмножество выходных значений;

$Pr[\cdot]$ — вероятность события.

3 В контексте синтеза данных дифференциальная приватность обеспечивает формальные гарантии защиты от атак реидентификации, вывода членства и вывода атрибутов.

3.3.5 приватность данных (data privacy): Свойство процессов обработки данных, характеризующее степень соответствия установленным правилам и ограничениям для реализации прав субъектов данных на контроль касающейся их информации.

Примечания

1 Правила и ограничения устанавливаются текущим законодательством и могут реализовываться через процедуры и политики организации, обрабатывающей данные, включая требования к реализации систем синтеза данных.

2 В контексте синтеза данных приватность данных может быть измерена вероятностью утечки данных, в том числе косвенными способами.

3 Термин «приватность данных» используется в настоящем стандарте для обозначения права субъекта данных на контроль над информацией о себе и ее защиту от несанкционированного раскрытия. В отличие от термина «конфиденциальность информации» по ГОСТ Р 50922, характеризующей режим ограничения доступа к информации, приватность ориентирована на защиту интересов субъекта данных.

3.3.6 составная приватность (compositional privacy): Свойство системы, при котором уровни приватности, обеспеченные отдельными процессами или алгоритмами обработки данных, могут быть формально объединены для вычисления совокупной гарантии приватности всей системы.

3.3.7 этика обработки данных (data processing ethics): Свойство процессов обработки данных, определяемое степенью соответствия установленным этическим принципам, таким как справедливость, прозрачность, конфиденциальность и подотчетность.

Примечания

1 Оценка этичности может проводиться с помощью количественных и качественных метрик, измеряющих, например, предвзятость, репрезентативность данных и объяснимость использованных алгоритмов.

2 Требования к этичности обработки данных являются дополнительными по отношению к правовым и техническим требованиям и не подменяют их.

4 Основные концепции синтеза данных

4.1 Синтез данных с использованием ИИ

Синтетические данные становятся важным инструментом в развитии систем ИИ [2], обеспечивая приемлемое использование данных для обучения моделей машинного обучения в тех случаях, когда доступ к реальным данным ограничен [3]. Такие данные позволяют преодолевать ограничения, связанные с конфиденциальностью, доступностью и качеством реальных данных. Синтетические данные могут воспроизводить сложные структуры и взаимосвязи (см. [4]), что делает их незаменимыми для тестирования и улучшения алгоритмов ИИ, а также для обеспечения их надежности и приватности.

Статистическая модель данных является одним из ключевых элементов синтеза данных и обеспечивает формальное представление распределений, зависимостей, корреляций, ограничений и иных статистических характеристик исходного набора данных, которые

ПНСТ

должны быть сохранены или воспроизведены в синтетических данных. При применении ИИ такая модель может использоваться как основа для обучения генеративных алгоритмов, контроля качества синтеза и оценки соответствия синтетического набора данным исходной предметной области. Рекомендуется, чтобы используемая статистическая модель была документирована в части учитываемых признаков, зависимостей, допущений, ограничений применимости и критериев оценки качества синтезированных данных.

4.2 Свойства синтетических данных

Синтетические данные создаются искусственным путем для воспроизведения характеристик и статистических свойств реальных данных. Такие данные генерируются (синтезируются) с использованием различных методов, включая алгоритмы машинного обучения и традиционные статистические методы. Целью создания синтетических данных является обеспечение возможности обучения и тестирования моделей машинного обучения, проведение аналитических исследований и выполнение других задач, требующих данных, при этом минимизируя негативные последствия, связанные с утечкой данных и доступностью реальных данных.

Примечание — Негативные последствия могут включать утечку данных, в том числе через модели машинного обучения, дискриминацию, снижение доверия к результатам анализа, а также нарушение требований законодательства и регуляторов.

Синтез данных может осуществляться с использованием методов ИИ, таких как генеративно-состязательные сети, вариационные автокодировщики, а также других моделей машинного обучения. Эти методы позволяют создавать данные, которые сохраняют важные

статистические характеристики оригинальных данных, обеспечивая их полезность для аналитики и моделирования, разработки и тестирования (см. [5]).

В случаях, когда данные проходят несколько последовательных или взаимосвязанных процессов обработки, каждый из которых вносит собственный вклад в общий уровень приватности, используется понятие составной приватности. Такая концепция позволяет формально учитывать накопление рисков раскрытия защищаемых данных при применении нескольких алгоритмов, механизмов защиты или этапов синтеза данных и определять совокупную гарантию приватности всей системы. При использовании составной приватности следует документировать отдельные гарантии приватности, правила их объединения, допущения применяемой модели и итоговую оценку совокупного риска.

Особенностью реализации систем синтеза данных является тесная связь системы синтеза данных и самих синтетических данных в процессе единого жизненного цикла, свойственного системам ИИ в соответствии с ГОСТ Р 70889—2023 (пункт 6.1).

Основные свойства синтетических данных:

- искусственность происхождения: данные создаются с использованием алгоритмов и моделей, а не собираются из реальных источников;
- сохранение статистических свойств: синтетические данные сохраняют важные статистические характеристики оригинальных данных, что делает их полезными для аналитики и моделирования;

ПНСТ

- приватность: синтетические данные (в том числе полные синтетические данные) не содержат идентифицируемой информации, что минимизирует вероятность утечки данных; в частности, синтетические данные при использовании доверенных технологий ИИ не содержат исходных (персональных) данных.

Синтетические данные (полные синтетические данные) отличаются от реальных (персональных) данных по своему происхождению и содержанию. В то время как реальные данные собираются из реальных источников и могут содержать идентифицируемую информацию, полные синтетические данные создаются искусственно и, при соблюдении должной методики, не содержат информации, позволяющей идентифицировать конкретные лица или объекты. Дополнительное применение доверенных технологий для предотвращения остаточных следов исходных данных, например дифференциальной приватности, делает синтетические данные надежным и эффективным инструментом для множества приложений, где использование реальных данных ограничено из-за соображений приватности.

В отличие от полных синтетических данных, частично синтетические данные формируются путем замены или дополнения отдельных элементов реального набора искусственно созданными значениями и поэтому требуют отдельной оценки остаточных рисков, связанных с сохранением реальных фрагментов данных.

Примечание — Синтетические данные также могут ускорить процесс подготовки данных и их использования (например, при тестировании информационных систем).

4.3 Классификация типов синтетических данных

Синтетические данные можно классифицировать по нескольким различным измерениям:

а) по уровню синтеза:

- 1) полностью синтетические данные: искусственно созданные данные, которые сохраняют характеристики и статистические свойства реальных данных, не содержат объектов реальных данных и формируются без их использования. При должной реализации не допускают восстановления исходных данных через статистические особенности в практическом применении;
- 2) данные, созданные путем замены или дополнения части реальных данных искусственно созданными данными, сохраняющими характеристики и статистические свойства исходных данных;

б) по методу синтеза:

- 1) генеративные модели машинного обучения: используют модели, такие как генеративно-состязательные сети и вариационные автокодировщики, для создания данных, имитирующих реальные;
- 2) алгоритмы, основанные на статистических методах: применяют традиционные статистические методы для генерации данных, основываясь на распределении и взаимосвязях исходных данных (см. [6]);

в) по области применения:

- 1) общие синтетические данные: синтетические данные, используемые в широком спектре приложений, таких как обучение и тестирование моделей машинного обучения (см. [5]);
 - 2) отраслевые синтетические данные: синтетические данные, предназначенные для конкретных областей, таких как здравоохранение (см. [7]), финансы или маркетинг, и учитывающие специфические требования и особенности этих областей, а также проблемы обмена данными (см. [8], [9]);
- г) по цели использования:
- 1) для обучения систем ИИ: данные, предназначенные для обучения моделей машинного обучения и повышения их производительности;
 - 2) для тестирования и валидации моделей: данные, используемые для тестирования и оценки моделей, а также для проверки их устойчивости и надежности, для тестирования общих информационных систем (см. [1], [10]);
 - 3) для демонстраций и представлений: данные, созданные для презентаций, демонстрации возможностей систем и проведения обучающих семинаров;
 - 4) для анализа, исследований и разработки новых алгоритмов и моделей: синтетические данные позволяют экспериментировать с различными параметрами моделей и алгоритмов без необходимости сбора новых реальных данных, а также могут использоваться для разработки новых метрик оценки качества моделей;

5) для разработки и тестирования новых информационных систем: синтетические данные помогают осуществлять отладку и проверку производительности пользовательских интерфейсов, архитектур интеграции и других современных компонентов согласно ГОСТ Р 59926.

Список применений синтетических данных может быть значительно расширен и не исчерпывается перечисленным.

Примечание — В качестве генеративных моделей могут использоваться не только генеративно-состязательные сети или вариационные автокодировщики, но и другие архитектуры, включая трансформеры (например, большие языковые модели), нормализующие потоки и диффузионные модели, в зависимости от типа данных и задачи.

4.4 Основные принципы осуществления синтеза данных

Синтетические данные играют определяющую роль в качестве функционирования систем ИИ согласно ГОСТ Р ИСО 8000-2, построенных с их использованием. Приватность и этичность синтетических данных, полученных с использованием систем ИИ, согласно ГОСТ Р 59276, основываются на базовых принципах, охватывающих аспекты приватности, ответственности и технической устойчивости, чтобы гарантировать, что синтетические данные служат в интересах общества и способствуют устойчивому развитию согласно ПНСТ 840—2023 в части:

а) благополучия общества и конкретных людей:

ПНСТ

- 1) синтетические данные служат инструментом для улучшения принятия решений человеком, сохраняя неприкосновенность частной жизни согласно ГОСТ ISO/IEC 29100;
 - 2) важное значение имеет внедрение эффективных механизмов контроля и аудита процессов генерации, использования и управления синтетическими данными, что гарантирует, что действия, основанные на синтетических данных, соответствуют намерениям и интересам людей (см. [11]);
- б) технической надежности и приватности:
- 1) системы синтеза являются технически надежными и отвечают принципам использования доверенных технологий, особенно при интеграции в критические сектора, такие как здравоохранение, транспорт и финансы согласно ГОСТ ISO/IEC 29100 (см. также [11];
 - 2) разработка систем синтеза, способных справляться с неопределенностями, обеспечивает необходимый уровень доверия и устойчивость к внешним атакам и внутренним сбоям согласно ГОСТ Р ИСО 22514-1;
- в) приватности и управления данными:
- 1) механизмы управления данными должны гарантировать, что данные собираются, хранятся и обрабатываются в соответствии с регулятивными нормами и уважением к личности (см. [4]);
 - 2) использование техник дифференциальной приватности, обезличивания для снижения вероятности утечки данных (см. [12]);
- г) прозрачности:

1) для обеспечения доверия пользователей к синтетическим данным процессы и алгоритмы генерации являются прозрачными;

2) информация о методах генерации данных и алгоритмах, использованных для их создания, излагается ясно и доступно;

д) разнообразия, недискриминации и справедливости:

1) синтетические данные свободны от искажений и справедливы во избежание предвзятости и дискриминации согласно ГОСТ ISO/IEC 29100;

2) проводится разработка алгоритмов, обеспечивающих справедливость в принятии решений, основанных на синтетических данных;

3) поддерживается экологическое и общественное благополучие;

4) влияние синтетических данных на общество и окружающую среду должно быть положительным;

5) осуществляется учет общественных и экологических последствий при разработке и использовании синтетических данных согласно ПНСТ 840—2023;

е) ответственности:

1) разработчики и пользователи синтетических данных несут ответственность за результаты функционирования своих алгоритмов; реализуемые механизмы направлены на обеспечение возможности исправления ошибок и устранения предвзятости;

ПНСТ

2) внедряются механизмы для исправления ошибок и механизмов подотчетности (см. [8]).

При осуществлении синтеза данных следует учитывать этику обработки данных как один из базовых принципов, определяющих допустимость и качество применяемых процедур. Процессы синтеза должны проектироваться, выполняться и оцениваться с учетом справедливости, прозрачности, конфиденциальности и подотчетности, включая проверку отсутствия недопустимых смещений, рисков нарушения прав субъектов данных, а также наличия объяснимости применяемых методов и результатов синтеза. Рекомендуется документировать этические допущения, критерии их оценки, выявленные ограничения, меры контроля и ответственность участников процесса за результаты синтеза и последующее использование синтетических данных.

Библиография

- [1] ISO/IEC TR 29119-11:2020 Системная и программная инженерия. Тестирование программного обеспечения. Часть 11. Руководящие указания по тестированию систем искусственного интеллекта (AI) (Software and systems engineering — Software testing — Part 11: Guidelines on the testing of AI-based systems)
- [2] Национальная стратегия развития искусственного интеллекта на период до 2030 года (В редакции Указа Президента Российской Федерации от 15 февраля 2024 г. № 124)
- [3] Федеральный закон от 27 июля 2006 г. № 152-ФЗ «О персональных данных»
- [4] ИСО/МЭК 18023-1:2006 Информационные технологии. Языковая привязка SEDRIS. Часть 1. Функциональные технические условия (Information technology — SEDRIS. Part 1: Functional specification)
- [5] ISO/IEC TR 24372:2021 Информационная технология. Искусственный интеллект (AI). Обзор вычислительных методов для систем искусственного интеллекта

- [Information technology — Artificial intelligence (AI) — Overview of computational approaches for AI systems]
- [6] ИСО/МЭК 4922-1:2023 Информационная безопасность. Протокол конфиденциального вычисления. Часть 1. Общие требования (Information security — Secure multiparty computation — Part 1: General)
- [7] ИСО 18374:2025 Стоматология. Анализ двумерных рентгенограмм на основе искусственного интеллекта (AI) и дополненного интеллекта (Aul). Генерация данных, аннотирование данных и их обработка [Dentistry — Artificial intelligence (AI) and augmented intelligence (Aul) based 2D radiograph analysis — Data generation, data annotation and data processing]
- [8] ИСО/МЭК 23751:2022 Информационная технология. Облачные вычисления и распределенные платформы. Рамочное соглашение об обмене данными (DSA) [Information technology — Cloud computing and distributed platforms —

- Data sharing agreement (DSA) framework]
- [9] ИСО 25237:2017 Информатика в здравоохранении. Псевдонимизация (Health informatics — Pseudonymization)
- [10] ISO/IEC TR 24028:2020 Информационная технология. Искусственный интеллект. Обзор достоверности систем искусственного интеллекта (Information technology— Artificial intelligence — Overview of trustworthiness in artificial intelligence)
- [11] ISO/IEC TR 5469:2024 Искусственный интеллект. Функциональная безопасность и системы искусственного интеллекта (Artificial intelligence — Functional safety and AI system)
- [12] ИСО/МЭК 20889:2018 Терминология и классификация методов деидентификации данных повышенной конфиденциальности (Privacy enhancing data de-identification terminology and classification of techniques)

Ключевые слова: синтез данных, основные положения

Руководитель организации-разработчика

Ассоциация больших данных
наименование организации

Исполнительный директор
должность

личная подпись

А. В. Нейман
инициалы, фамилия

Руководитель разработки

должность

личная подпись

инициалы, фамилия

Исполнитель:

должность

личная подпись

инициалы, фамилия