*Математические структуры и моделирование*
*2025. № 3 (75). С. 116–121*

УДК 004.8
DOI 10.24147/2222-8772.2025.3.116-121

# СИНТЕТИЧЕСКИЕ ДАННЫЕ: ПРОБЛЕМЫ И ПУТИ ИХ РЕШЕНИЯ

**Т.Л. Салова**
к.т.н., доцент, e-mail: salova@mail.ru
**И.С. Суворов**
магистрант, e-mail: ioannsuvr@yandex.ru

Сочинский государственный университет, Сочи, Россия

**Аннотация.** Синтетические данные стали ключевым инструментом в машинном обучении, тестировании программного обеспечения и защите приватности. Однако их генерация и применение сопряжены с проблемами качества, этики и технической сложности. Например, применение синтетических данных в социологии, в форме синтетических респондентов, с одной стороны, позволяет имитировать реальное поведение исследуемой аудитории, с другой стороны, не исключает проблему предвзятости. В данной статье анализируются основные вызовы, связанные с синтетическими данными, и предлагаются пути их решения, включая усовершенствование алгоритмов генерации, внедрение методов валидации и соблюдение этических норм.

**Ключевые слова:** синтетические данные, этические нормы, генеративно-состязательные системы, синтетические респонденты, трансформеры, генеративные нейронные сети.

В эпоху цифровой трансформации синтетические данные перешли из категории экспериментальных инструментов в разряд критически важных ресурсов для технологических инноваций. В отличие от данных, собранных эмпирическим путём, они создаются алгоритмически с использованием методов машинного обучения (например, генеративно-состязательных сетей), физического моделирования или стохастических процессов, что позволяет гибко управлять их объёмом, разнообразием и параметрами распределения. Подобные возможности делают их незаменимыми в сценариях, где доступ к реальным данным ограничен правовыми нормами (медицина, финансы), требует значительных временных затрат (автономные транспортные системы) или сопряжён с высокими рисками (кибербезопасность) [1].

Несмотря на растущий интерес к технологии, её повсеместное внедрение сдерживается фундаментальными проблемами. Во-первых, сохраняются вопросы к качеству синтеза: даже современные алгоритмы не всегда корректно воспроизводят скрытые зависимости в данных, такие как пространственно-временные корреляции в геоданных или причинно-следственные связи в клинических исследованиях. Во-вторых, сохраняется парадокс анонимности: комбинирование синтетических наборов с внешними источниками информации в ряде случаев позволяет реконструировать прототипы исходных записей, что ставит под угрозу концепцию «этичного

ИИ». В-третьих, генерация данных, адекватно отражающих сложные доменные особенности (например, многоуровневые социальные взаимодействия в метавселенных), требует не только мощных вычислительных ресурсов, но и междисциплинарных знаний в предметной области [2].

Целью настоящей работы является структурирование накопленного опыта в области синтетических данных через призму преодоления указанных ограничений. В статье систематизированы современные подходы к оценке качества синтеза, проанализированы кейсы неудачного применения технологии, а также предложена архитектура комплексного решения, объединяющего технические, методологические и регуляторные аспекты. Особое внимание уделено роли человекоцентричного дизайна в создании систем генерации, обеспечивающего баланс между утилитарной ценностью данных и соблюдением этико-правовых норм.

Главным ограничением синтетических данных остаётся их неспособность полноценно воспроизводить многомерные зависимости и нюансы, присущие реальным данным. Современные алгоритмы, включая классические генеративно-состязательные сети (GAN), часто фокусируются на аппроксимации глобальных статистических характеристик, таких как средние значения или дисперсия, но игнорируют локальные аномалии, редкие события и контекстуальные взаимосвязи. Например, при моделировании финансовых временных рядов простые GAN могут корректно генерировать тренды, но не учитывать «чёрных лебедей» – экстремальных рыночных событий, критически важных для стресс-тестирования. Это приводит к тому, что модели, обученные на таких данных, демонстрируют завышенную точность в контролируемых условиях, но терпят неудачи при столкновении с реальными сценариями [3].

Особую остроту проблема приобретает в областях с высокой стоимостью ошибки. В медицинской диагностике, где синтетические данные используются для расширения выборок редких заболеваний, некорректный синтез аномалий (например, артефактов МРТ-снимков при опухолях мозга) искажает представление модели о паттернах болезни. В результате система может пропускать ранние стадии онкологии или, напротив, генерировать ложноположительные диагнозы, ставя под угрозу персонализированное лечение.

Для количественной оценки качества применяется комбинация метрик:

1. Статистические расстояния (KL-дивергенция, расстояние Вассерштейна) – измерение расхождения между распределениями реальных и синтетических признаков и выявление систематических смещений.

2. Семантическая согласованность – анализ сохранения смысловых связей между переменными (например, корреляция между возрастом пациента и частотой определённых диагнозов).

3. Адверсарное тестирование – проверка, способна ли модель-классификатор отличить синтетические записи от реальных при их смешивании.

Тем не менее ни одна метрика не гарантирует абсолютной достоверности. Так, KL-дивергенция может «не замечать» локальных расхождений в хвостах распределений, а адверсарные методы чувствительны к переобучению тестовой модели. Поэтому валидация требует привлечения экспертов предметной области – радиологов, финансистов, социологов – для ручной проверки содержательной целостности

данных.

Даже тщательно сгенерированные данные несут риск деанонимизации при интеграции с внешними источниками. Например, синтетическая медицинская запись, содержащая уникальную комбинацию возраста, региона проживания и истории вакцинации, может быть сопоставлена с публичными демографическими базами, раскрывая прототип реального пациента. Это ставит под вопрос этичность использования технологии в конфиденциальных сферах без дополнительных защитных механизмов, таких как дифференциальная приватность или синтез на уровне агрегированных признаков [4].

Генерация высококачественных данных для сложных доменов (например, мультимодальные наборы, сочетающие текст, изображения и сенсорные сигналы) требует не только продвинутых алгоритмов, но и значительных вычислительных ресурсов. Тренировка моделей на базе трансформеров или диффузионных архитектур для создания реалистичных синтетических видеопоследовательностей может занимать недели даже на кластерах с GPU, что ограничивает доступ к технологии для небольших исследовательских групп.

Создание синтетических данных, несмотря на их потенциальную безопасность, сопряжено с серьёзными этическими и правовыми рисками. Одной из ключевых проблем остаются атаки на повторную идентификацию. Даже искусственно сгенерированные данные могут содержать скрытые закономерности, позволяющие восстановить исходную информацию о реальных людях. Правовое регулирование ужесточает требования к защите персональных данных. Регламенты GDPR (Общий регламент по защите данных ЕС) и HIPAA (Закон о переносимости и подотчётности медицинского страхования США) предъявляют строгие требования к анонимизации информации. Например, GDPR требует, чтобы данные не могли быть повторно связаны с конкретным человеком даже при использовании дополнительной информации. Это создаёт парадокс: для обучения реалистичных генеративных моделей необходимы детальные исходные данные, но их может нарушить принцип анонимности. В результате разработчикам приходится искать компромисс между качеством синтетических данных и юридическими ограничениями [5].

Синтетические данные, применяемые в социологии в форме синтетических респондентов, когда искусственный интеллект имитирует поведение исследуемой аудитории, используются для проверки базовых гипотез, предсказания реального поведения. И здесь существует проблема приватности, встроенной в технологии. Поэтому важно, чтобы исследования были этически обоснованными и научно достоверными. «При осторожном использовании эти технологии будут способствовать методологическим инновациям в социальных науках» [6, с. 67].

Отдельную проблему представляет воспроизведение уникальных комбинаций признаков. Синтетические данные, созданные на основе редких или уникальных случаев (например, пациентов с экзотическими заболеваниями или комбинациями генетических мутаций), могут косвенно идентифицировать конкретных людей. Если в реальной популяции существует только один человек с определённым набором характеристик, его синтетический «двойник» сохраняет прямую связь с оригиналом. Это особенно критично в медицинских исследованиях, где утечка подобных паттернов может привести к разглашению диагноза, генетических особенностей

или других конфиденциальных сведений.

Современные методы генерации синтетических данных, такие как диффузионные модели или архитектуры на основе трансформеров, требуют колоссальных вычислительных ресурсов. Например, обучение модели StyleGAN3, разработанной NVIDIA для создания фотореалистичных изображений, занимает более 300 GPU-часов даже на высокопроизводительных серверах с графическими процессорами уровня Tesla V100 [7]. Это связано с необходимостью обработки миллионов параметров и многократной оптимизации модели для достижения правдоподобности результатов.

Помимо аппаратных затрат, возникают сложности с обеспечением репрезентативности данных. Генеративные модели склонны воспроизводить системные смещения, присутствующие в обучающих наборах. Например, если исходные данные содержат дисбаланс по расовому или гендерному признаку, синтетические данные усилят эти перекосы, что может привести к дискриминационным последствиям в системах искусственного интеллекта. Для устранения этой проблемы требуются дополнительные этапы валидации и коррекции, что ещё больше увеличивает время и стоимость разработки.

Кроме того, поддержание конфиденциальности в процессе генерации требует интеграции специализированных методов, таких как дифференциальная приватность или федеративное обучение. Эти технологии добавляют «шум» в данные или распределяют вычисления между устройствами, чтобы минимизировать риски утечек. Однако их внедрение снижает точность моделей и усложняет процесс настройки, создавая дилемму между безопасностью и практической применимостью синтетических данных.

Современные подходы, такие как диффузионные модели (DDPM), работают по принципу поэтапного «очищения» данных от шума, что позволяет создавать высокодетализированные объекты. Например, в генерации изображений это помогает избежать артефактов вроде размытых границ или неестественных текстур. Трансформеры, в свою очередь, за счёт механизма внимания анализируют контекстные зависимости, что критично для последовательностей (текст, временные ряды) [8].

Для задач, требующих соблюдения физических законов, применяют гибридные архитектуры. Например, в симуляции жидкостей нейросеть обучается на уравнениях Навье – Стокса, а в генерации медицинских данных – на анатомических атласах. Это позволяет синтетическим данным сохранять научную достоверность.

Метрики качества моделей (F1, AUC-ROC) используются в схеме тренировка-на-синтетике/тест-на-реальных. Если модель показывает близкие результаты на обоих наборах, данные считаются пригодными.

Для достижения неотличимости от реальных данных критически важны три аспекта.

Первый – контекстная достоверность, требующая адаптации генерации под специфику домена: тексты создаются через дообучение языковых моделей (GPT-4o, Claude) на целевых данных с последующим ручным устранением логических противоречий, а изображения генерируются в высоком разрешении (1024x1024+) с добавлением артефактов, имитирующих шум камеры или сжатие JPEG.

Второй принцип – динамическая вариативность, которая обеспечивает есте-

ственность за счёт внедрения контролируемой случайности: временные ряды обогащаются фликкер-шумом и «сбоями», аналогичными помехам датчиков, а изображения – вариациями освещения и ракурсов.

Третий элемент – бесшовная интеграция в экосистему, где синтетика смешивается с реальными данными (например, 30 % искусственных транзакций в финансовых выборках) и соблюдает стандарты метаданных (EXIF, ISO8601), что позволяет использовать их в единых аналитических пайплайнах. Комбинация этих подходов не только обеспечивает статистическую и семантическую близость к реальности, но и усложняет обнаружение алгоритмами-детекторами (GPTZero, Anti-Forgery Toolkit) за счёт имитации «неидеальностей», присущих естественным данным [9].

Синтетические данные стали ключевым инструментом в эпоху цифровой трансформации, позволяя преодолевать ограничения реальных наборов через внедрение диффузионных моделей, трансформеров и гибридных архитектур, интегрирующих предметные знания. Однако их применение требует строгой этической дисциплины – от внедрения дифференциальной приватности до аудита на смещения, – а также оптимизации вычислительных ресурсов через квантованные модели и облачные решения. Будущие направления зависят от баланса между инновациями (квантовые вычисления, DaaS-платформы) и регуляторной зрелостью, чтобы синтетика оставалась не заменой, а безопасной средой для ответственных экспериментов, где технологический прогресс не противоречит социальным и правовым нормам.

## Литература

1. Копылов Д.А., Агешин Е.С., Хомутская О.В. Формирование синтетических данных для обучения системы компьютерного зрения // Автоматизация и моделирование в проектировании и управлении. 2022. № 4. С. 18–19.
2. Пчелинцев С., Юляшков М. А., Ковалева О. А. Метод создания синтетических наборов данных для обучения нейросетевых моделей распознаванию объектов // Информационно-управляющие системы. 2022. № 3. С. 9–17.
3. Узких Г. Ю. Применение генеративно-состязательных сетей (gan) в обработке изображений // Вестник науки. 2024. № 8. С. 182–184.
4. Воронов Ю.П., Свиридов М.А Новые средства измерения межрегиональных различий // Интерэкспо Гео-Сибирь. 2019. № 1. С. 71–76.
5. Колесникова Г.И. Искусственный интеллект: проблемы и перспективы // Видеонаука. 2018. № 2. С. 34–38.
6. Драч В.Е., Торкунова Ю.В. Использование генеративного искусственного интеллекта для социологических исследований. // ДИСКУРС. 2025. Т. 11. № 1. С. 52–70.
7. Кумратова А. М., Борлакова М. А., Сайкинов В. Е., Когай И. Е. Использование диффузионных моделей для разработки приложений, генерирующих изображения на основе текстовых запросов // Вестник Адыгейского государственного университета. Серия 4: Естественно-математические и технические науки. 2022. № 4. С. 66–70.
8. Гайнетдинов А.Ф. Исследование влияния трансформеров на улучшение генерации изображений // Universum: технические науки. 2024. № 4. С. 45–48.
9. Малышев И.О., Смирнов А.А. Обзор современных генеративных нейросетей: отечественная и зарубежная практика // Международный журнал гуманитарных и естественных наук. 2024. № 1–2. С. 168–171.

# SYNTETIC DATA: PROBLEMS AND SOLUTIONS

**T.L. Salova**
Ph.D. (Tech.), Associate Professor, e-mail: salova@mail.ru
**I.S. Suvorov**
Master's Degree Student, e-mail: ioannsuvr@yandex.ru

Sochi State University, Sochi, Russia

**Abstract.** Synthetic data has become a key tool in machine learning, software testing, and privacy protection. However, their generation and application are fraught with issues of quality, ethics, and technical complexity. For example, the use of synthetic data in sociology, in the form of synthetic respondents, on the one hand, allows imitating the real behavior of the audience under study, on the other hand, does not exclude the problem of bias. This article analyzes the main challenges associated with synthetic data and suggests ways to solve them, including improving generation algorithms, implementing validation methods, and complying with ethical standards.

**Keywords:** synthetic data, ethical norms, generative-adversarial systems, synthetic respondents, transformers, generative neural networks.

*Дата поступления в редакцию: 16.05.2025*