УДК: 004.75, 004.41

К вопросу определения предметной области информационной безопасности технологий искусственного интеллекта

S.V. Zapechnikov, A.Yu. Shcherbakov

On the Issue of Determining the Subject Area of Information Security of Artificial Intelligence Technologies

Abstract. *The article is devoted to the analysis of issues related to the definition of the subject area of information security of artificial intelligence technologies. The analysis of the composition and content of the subject area is carried out, typical threats to artificial intelligence systems and typical models of adversaries are highlighted. The features of information processing using artificial intelligence technologies that determine the tasks of ensuring information security are analyzed. An informal formulation of the privacy-preserving machine learning problem is given. Classification features and criteria for evaluating privacy-preserving machine learning systems are highlighted. The classification of known methods and systems that ensure the privacy and verifiability of machine learning is carried out.*

Keywords: *artificial intelligence technologies, machine learning, deep learning, confidentiality, secure multi-party computations, secret sharing schemes, homomorphic encryption.*

С.В. Запечников[1]
А.Ю. Щербаков[2]

[1] Доктор технических наук, профессор Института интеллектуальных кибернетических систем, Национальный исследовательский ядерный университет «МИФИ», Вице-президент по научной работе Ассоциации специалистов в области криптовалют и цифровых финансовых активов.
E-mail: SVZapechnikov@mephi.ru

[2] Доктор технических наук, профессор, главный научный сотрудник РАН (ИТМиВТ им.С.А.Лебедева), президент Ассоциации специалистов в области развития криптовалют и цифровых финансовых активов.
E-mail: x509@ras.ru

Аннотация. *Статья посвящена анализу вопросов, связанных с определением предметной области информационной безопасности технологий искусственного интеллекта. Проводится анализ состава и содержания предметной области искусственного интеллекта, выделяются характерные угрозы системам искусственного интеллекта, типичные модели нарушителей. Анализируются особенности обработки информации с использованием технологий искусственного интеллекта, определяющие задачи обеспечения информационной безопасности. Приводится неформальная постановка задачи конфиденциального машинного обучения. Выделяются классификационные признаки и критерии оценки систем конфиденциального машинного обучения. Проводится классификация известных методов и систем, обеспечивающих конфиденциальность и проверяемость машинного обучения.*

Ключевые слова: *технологии искусственного интеллекта, машинное обучение, глубокое обучение, конфиденциальность, безопасные многосторонние вычисления, схемы разделения секрета, гомоморфное шифрование.*

ВВЕДЕНИЕ

Одна из самых значительных тенденций современных компьютерных наук и информационных технологий – бурное развитие технологий и систем искусственного интеллекта (ИИ). Немало говорится о том, что ИИ формирует ядро нового технологического уклада. Как любой комплекс новых технологий, он проходит путь от теоретической разработки, экспериментов и прототипов к широкому внедрению во многие сферы деятельности. Одним из важных критериев принятия обществом новых технологий, безусловно, является доверие к ним. В связи с этим возникло понятие доверенного ИИ. В России понятие доверенного ИИ с 2021 г. закреплено в стандарте ГОСТ Р 59276-2020 «Системы искусственного интеллекта. Способы обеспечения доверия. Общие положения». Процитируем определения из этого стандарта.

«Искусственный интеллект – 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.»

«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.»

«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.»

Информационная безопасность названа в этом стандарте в числе существенных характеристик, способных влиять на качество систем ИИ, а принятие эффективных мер по защите информации на стадии эксплуатации системы ИИ названо среди способов обеспечения доверия к системам ИИ.

Вместе с тем, предмет информационной безопасности технологий и систем ИИ до сих пор в значительной мере остаётся неопределенным, вызывая дискуссии среди экспертов в сфере ИИ и информационной безопасности. Ни в коем случае не претендуя на полноту и окончательный характер решения проблем информационной безопасности ИИ, в настоящей статье авторы делают попытку представить предварительные результаты своих исследований, позволяющие прояснить предмет обеспечения информационной безопасности применительно к современным технологиям и системам ИИ.

ПРЕДМЕТНАЯ ОБЛАСТЬ ТЕХНОЛОГИЙ И СИСТЕМ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА

Для решения проблем информационной безопасности прежде всего требуется уточнить саму предметную область ИИ. В настоящее время чаще всего её отождествляют с технологиями машинного обучения, иногда включая в неё также и технологии интеллектуального анализа данных, которые имеют высокую степень общности с технологиями машинного обучения. Не вызывает сомнения, что ядро технологий ИИ действительно составляет машинное обучение, в первую очередь глубокое обучение. Однако анализ таких авторитетных источников как [1, 2] позволяет констатировать, что в предметную область ИИ следует включить следующие основные разделы:

- методы поиска в пространствах состояний, включая генетические алгоритмы;
- мультиагентный поиск;
- марковские процессы принятия решений;
- теоретико-игровые методы принятия решений;
- исчисление высказываний (пропозициональная логика);
- исчисление предикатов (логика первого порядка);
- элементарные методы машинного обучения;
- классические нейронные сети;
- нейронные сети специальной архитектуры: свёрточные, рекуррентные и пр.;
- обучение без учителя, включая генеративные модели;
- обучение с подкреплением;
- вероятностные графовые модели;
- графы знаний;
- интеграция логических обоснований и обучения, включая трансферное обучение, непрерывное машинное обучение и пр.

Таким образом, предметная область ИИ в современном понимании включает в себя модели увеличивающейся выразительной силы: от рефлекторных моделей, основанных на состояниях, до моделей, основанных на переменных, и логики, и для каждой из них рассматриваются свои парадигмы моделирования – вывода – обучения.

Таким образом, обеспечение информационной безопасности, вообще говоря, должно пониматься шире, чем только защита систем машинного обучения. Очевидно, что ряд задач обеспечения информационной безопасности не будет отличаться от уже существующих ИТ-систем другого назначения и может быть решен традиционными методами. Однако выявление и формулирование новых задач обеспечения информационной безопасности, специфичных для систем ИИ, а также поиск методов и средств их решения должны стать предметом дальнейших исследований. Поэтому в настоящей статье мы сосредоточимся на задачах обеспечения информационной безопасности процессов машинного обучения.

Крайне важно не только принимать меры по

защите информации на стадии эксплуатации системы ИИ, как это отмечено в стандарте ГОСТ Р 59276-2020, но уже на стадии создания системы ИИ проектировать ее архитектуру, алгоритмическое обеспечение и выбирать способы и средства её реализации таким образом, чтобы априори обеспечить её информационную безопасность, тем самым внося вклад в обеспечение доверия к системам ИИ.

УГРОЗЫ СИСТЕМАМ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА

При анализе информационной безопасности системы ИИ принимают во внимание следующие потенциально опасные для её функционирования факторы:

- разглашение конфиденциальных данных, передаваемых между участниками системы;
- манипуляции данными, поступающими на вход моделей;
- «отравление» обучаемой модели посредством умышленной передачи на вход искаженных признаков объектов обучающей выборки (неверно размеченных объектов, систематически смещенных числовых признаков объектов и т.п.);
- частичная или полная реконструкция объекта анализа по признаковому описанию;
- деанонимизация субъекта по персональным данным (медицинским, финансовым и т.п., возможно, даже обезличенным) и иным признакам, используемым для обучения или тестирования модели;
- установление членства (или отсутствия членства) объекта с фиксированным набором признаков в выборке;
- инверсия обученной модели, т.е. восстановление значений признаков, на которых была обучена модель;
- для методов обучения, основанных на градиентном спуске, – восстановление объектов выборки по небольшой части градиентов, а также другие факторы.

Очевидно, что перечисленные угрозы относятся ко всем классическим аспектам информационной безопасности: доступности, целостности и конфиденциальности информации. Для систем ИИ, обрабатывающих информацию ограниченного распространения, наиболее актуальны угрозы конфиденциальности информации.

МОДЕЛИ НАРУШИТЕЛЕЙ В СИСТЕМАХ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА

При анализе проблем обеспечения информационной безопасности в системах ИИ выделяют две категории нарушителей: внешние и внутренние.

О внешнем нарушителе делаются стандартные предположения о влиянии его на информацию, передаваемую по каналам связи, а именно, о возможности прослушивания, модификации, удаления, вставки и задержки сообщений. Во всех случаях, когда делается предположение о присутствии внешнего нарушителя, предполагается, что передача данных осуществляется по защищенным коммуникационным каналам, обеспечивающим конфиденциальность и аутентичность.

Внутренние нарушители – это лица из числа участников системы ИИ, предпринимающие действия с целью получить преимущества над остальными участниками. Их принято подразделять на два типа:

- получестный (semi-honest) нарушитель;
- злоумышленный (malicious) нарушитель.

Получестный нарушитель также может считаться условно пассивным, так как он характеризуется теми же поведенческими шаблонами, что и классический пассивный нарушитель. Он в точности придерживается предписанного протокола взаимодействия с остальными участниками, однако пытается в процессе взаимодействия получить некоторые преимущества над другими участниками, например, выведать у них не предназначенные ему конфиденциальные данные путем генерации и отправки им специально подобранных наборов входных данных. Допускается применение получестным нарушителем любых методов и алгоритмов для получения преимущества над другими участниками, включая, в том числе, любые методы и алгоритмы интеллектуального анализа данных и машинного обучения.

Злоумышленный нарушитель определяется как участник протокола, который может про-

извольно отклоняться от предписанного хода выполнения протоколов с целью достижения своих индивидуальных целей, противоречащих целям других участников протокола, в том числе посредством произвольной модификации данных, передаваемых и принимаемых им по каналам связи, внедрения, удаления и задержки сообщений, отказа от продолжения участия в протоколе и т.п. способами.

При трех и более участниках вычислений следует уточнять, какая доля участников относится к нарушителям.

Если большинство участников вычислений предполагается честным, то говорят о *модели честного большинства* (honest-majority adversarial setting). Можно рассматривать две разновидности модели честного большинства: с получестными и злоумышленными нарушителями. Сложность защиты систем ИИ во втором случае существенно выше, так как в вычисления необходимо вводить избыточность, достаточную для того, чтобы честные участники могли довести процесс вычислений до конца, несмотря на действия злоумышленников.

Самой сильной моделью является *модель нечестного большинства* (dishonest majority): в этом случае половина и более от общего количества участников вычислений могут быть нарушителями. Сложность защиты систем ИИ в этом случае возрастает многократно по сравнению с моделью честного большинства, так как каждый из участников вычислений должен самостоятельно выполнить такой объем вычислений, который сопоставим с объемом вычислений всего пула вычислителей. Дополнительными криптографическими инструментами при проектировании протоколов в этой модели служат доказательства с нулевым разглашением.

ОСОБЕННОСТИ ОБРАБОТКИ ИНФОРМАЦИИ С ИСПОЛЬЗОВАНИЕМ ТЕХНОЛОГИЙ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА, ОПРЕДЕЛЯЮЩИЕ ЗАДАЧИ ОБЕСПЕЧЕНИЯ ИНФОРМАЦИОННОЙ БЕЗОПАСНОСТИ

Выделим особенности обработки информации с использованием технологий и систем ИИ, определяющие формулирование и решение задач обеспечения информационной безопасности.

1. *Архитектурные особенности систем обработки данных.* Для современных сложных систем обработки данных характерна распределенная конфигурация, как правило, совмещенная с гетерогенностью. Распределенная архитектура системы выражается как в пространственном разделении клиентских и серверных компонент системы, так и в разделении серверной компоненты на несколько физических составляющих, способных взаимодействовать в процессе решения функциональных задач по заданиям клиентов. Распространенный способ реализации распределенной архитектуры – облачные системы обработки данных.

Распределенный характер систем обработки данных определяет необходимость обеспечения конфиденциальности при обработке информации ограниченного распространения. Функции обеспечения конфиденциальности должны применяться при передаче данных между клиентскими и серверными компонентами, при обмене данными между параллельно функционирующими серверами, а также непосредственно в процессе обработки информации серверами, так как они чаще всего рассматриваются клиентами как недоверенные компоненты. Если первые две задачи решаются традиционными криптографическими методами защиты информации, то последняя требует специфических решений, которые реализуются в концепции безопасных многосторонних вычислений.

2. *Особенности технологий обработки конфиденциальных данных.* Следствием архитектурных особенностей являются особенности процессов обработки конфиденциальных данных в распределенных системах. Типичными приемами обработки данных становятся аутсорсинговые и проверяемые вычисления.

Аутсорсинговые вычисления (outsourcing computations) – это технология обработки конфиденциальных данных, при которой одна из сторон протокола владеет секретом, другая сторона – вычислительными ресурсами. При этом обе стороны, потенциально не доверяющие друг другу, заинтересованы в совместном решении некоторой общей задачи. Заказчик

вычислений (он же – получатель результата) не обладает достаточными вычислительными ресурсами для решения поставленной им задачи. В силу этого он вынужден обращаться для решения задачи, входные данные которой (и, возможно, результаты) являются конфиденциальными, к удаленному серверу.

Одним из основных криптографических примитивов для реализации аутсорсинговых вычислений является гомоморфное шифрование.

Проверяемые (верифицируемые) вычисления – это технология обработки конфиденциальных данных, при которой заказчик вычислений и/или внешние наблюдатели должны иметь возможность проверить корректность вычислений на основе проверочной информации, предоставленной исполнителем вычислений.

Одним из основных криптографических примитивов для реализации проверяемых вычислений являются вероятностные криптографические доказательства, в частности, доказательства с нулевым разглашением.

3. *Устойчивые приемы использования технологий ИИ для обработки массивов данных разных типов*:

• линейно упорядоченные массивы данных (текст, временные ряды, аудиопотоки, видеопотоки, траектории и пр.) обрабатываются преимущественно при помощи рекуррентных нейронных сетей.

• двумерные и многомерные упорядоченные массивы данных (цифровые изображения: рисунки, фотографии и т.п.) обрабатываются преимущественно при помощи сверточных нейронных сетей.

• сложноструктурированные данные, не имеющие сколько-нибудь выраженной регулярности, связанные отношениями между отдельными объектами, обрабатываются преимущественно при помощи методов машинного обучения на графах.

Выбор технологии в зависимости от типа и характера обрабатываемого массива данных является первичным. На основе выбранной технологии и в зависимости от предположения об угрозах и нарушителях информационной безопасности далее выбирается система конфиденциального машинного обучения.

ПОСТАНОВКА ЗАДАЧИ КОНФИДЕНЦИАЛЬНОГО МАШИННОГО ОБУЧЕНИЯ

Задачу конфиденциального машинного обучения (англ. privacy-preserving machine learning) можно определить как задачу обеспечения гарантий конфиденциальности данных каждого из участников системы машинного обучения в условиях, когда лица, предоставляющие обучающую выборку на этапе обучения модели (training) либо предоставляющие запросы к модели на этапе её эксплуатации (inference) и ожидающие получения ответов на свои запросы (далее будем называть их клиентами), дистанционно взаимодействуют с обладателем модели, способным выполнять вычисления с помощью этой модели (далее будем называть его сервером). При этом клиенты заинтересованы в неразглашении своих данных (обучающей выборки, запросов к модели, ответов на них) как друг другу, так и обладателю модели. В то же время обладатель модели заинтересован в неразглашении параметров своей модели клиентам. Конкретные ситуации, в которых возникает такая заинтересованность, могут различаться: например, при обработке его персональных данных, а также данных, составляющих врачебную, налоговую или банковскую тайну. Заинтересованность обладателя модели может возникать при предоставлении платных услуг (например, прогнозирования) с использованием модели. Такого рода услуги получили обобщенное наименование «машинное обучение как сервис» (MLaaS – Machine Learning as a Service).

В оригинале концепция конфиденциального машинного обучения называется privacy-preserving machine learning, т.е. буквально «машинное обучение, сохраняющее приватность». Как известно, конфиденциальность информации неформально можно определить как гарантии того, что она не станет известна лицам, которым она не предназначена. Однако применительно к системам машинного обучения, как и в ряде других случаев, в зарубежной литературе чаще всего используется понятие приватности (privacy). Под ним понимается обеспечение тайны частной жизни, гарантии

того, что никто не сможет узнать о владельце (источнике) какой-либо информации больше того, что он сам желает. Каждый владелец вправе регулировать, кто и какие сведения о нём получает. Это свойство применимо не только к конфиденциальным данным. Приватность включает в себя такие составляющие как конфиденциальность, разграничение доступа, анонимность, несвязываемость действий или событий, неразличимость инициатора события и ряд других в зависимости от конкретной ситуации. Современные технологии предоставляют много способов узнать о человеке больше, чем он сам того пожелает: прямые утечки данных, восстановление информации по косвенным признакам, интеллектуальный анализ данных с применением машинного обучения (выявление «тонких» и скрытых закономерностей в данных) и многие другие. В широком смысле слова приватность подразумевает защищенность от «излишне любопытного» нарушителя.

Конфигурацию, в которой осуществляется конфиденциальное машинное обучение, можно представить архитектурной моделью, изображенной на рис. 1.

Модель включает в себя три основных функциональных блока:

• блок генерации задачи, передаваемой на аутсорсинг, и получения результата решения задачи;

• блок преобразования (трансформации) задачи и проверки результата решения задачи;

• блок решения задачи.

Физическое соответствие функциональных блоков участникам вычислительного процесса может быть неоднозначным. Наиболее предпочтительной является схема, когда первые два блока расположены на доверенной системе-клиенте, третий блок – у недоверенной серверной стороны, которая может быть представлена одним или несколькими физически выделенными серверами либо виртуальными машинами. Однако такая конфигурация может приводить к высокой вычислительной нагрузке на клиента. Второй возможной конфигурацией является размещение первого блока на доверенной системе-клиенте, второго блока – на доверенной системе-шлюзе, третьего блока – на недоверенных серверах. Такая модель позволяет удобно разместить по блокам всю основную функциональность, обеспечивающую безопасность информации, но потенциально связана с большим количеством уязвимостей.

Исполнитель вычислений (серверная сторона) может быть представлен единственным физическим вычислителем либо кластером взаимосвязанных устройств, которые в этом случае должны выполнять между собой многосторонние протоколы безопасных вычислений.

Блоки подготовки задачи, а также обратного

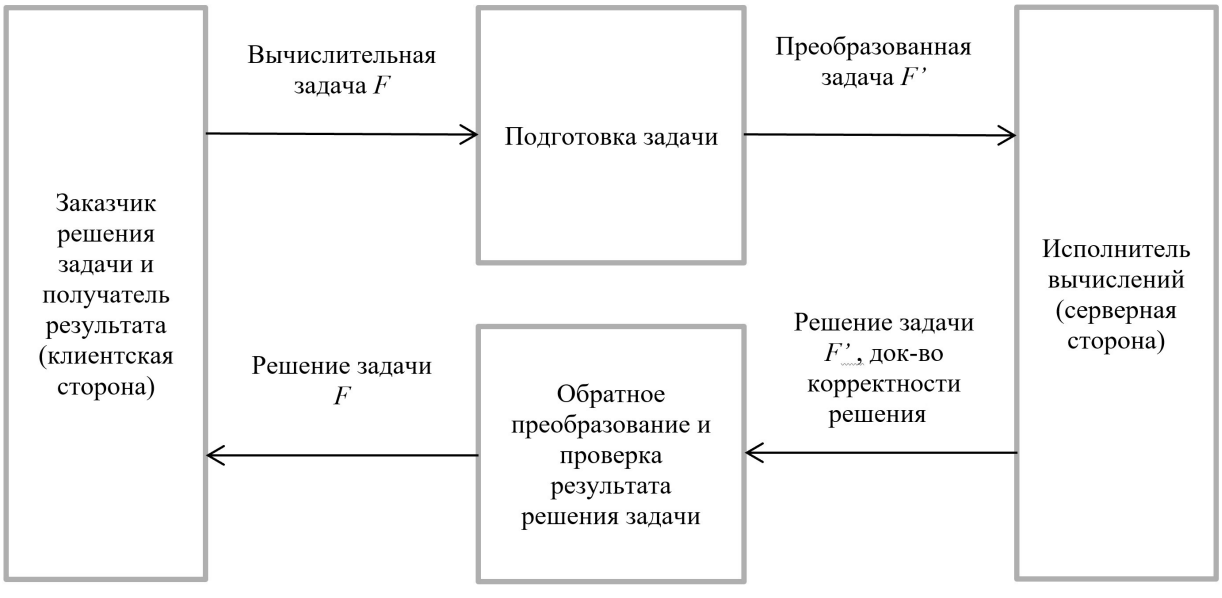

Рис. 1. Обобщенная модель архитектуры систем конфиденциального и проверяемого машинного обучения

преобразования и проверки решения задачи могут включать в себя разную функциональность в зависимости от метода решения задачи. Так, при использовании гомоморфного шифрования первый из этих блоков будет выполнять преобразование зашифрования входных данных, второй – расшифрования результата решения задачи. В случае использования безопасных многосторонних вычислений на основе схем разделения секрета первый блок будет разделять конфиденциальные входные данные на доли для последующей пересылки долей вычислителям, второй блок – собирать результат решения задачи из долей, переданных вычислителями. В ряде случаев к обратному преобразованию может добавляться проверка корректности решения задачи, например, с помощью доказательств с нулевым разглашением, прилагаемых вычислителями к своим долям решения.

В качестве промежуточной модели рассматривается также возможность участия в вычислениях клиента (при наличии у него достаточных вычислительных ресурсов) наряду с серверами (это характерно для протоколов на основе схем разделения секрета).

От недоверенных компонентов модели требуется следующая функциональность:

• выполнение алгоритмов обработки данных без раскрытия открытого текста данных;

• генерация доказательств корректного выполнения требуемых алгоритмов обработки данных.

От доверенных компонентов модели требуется следующая функциональность:

• подготовка исходных данных решаемой задачи для передачи недоверенным компонентам;

• прием результатов решения задачи от недоверенных компонентов и преобразование их в формат, пригодный для восприятия заказчиком вычислений;

• при необходимости – проверка корректности решения задачи недоверенным компонентом.

В связи с этим наиболее подходящими инструментами реализации функций, требуемых от недоверенных компонентов, выглядят следующие криптографические методы:

• полностью и частично гомоморфные схемы шифрования (в части алгоритмов выполнения арифметических операций над зашифрованными данными);

• GC-схемы (garbled circuits);

• схемы разделения секрета (операции над долями разделенных секретов);

• протоколы безопасных многосторонних вычислений (SMPC – secure multi-party computations);

• доказательства с нулевым разглашением (алгоритмы генерации доказательств).

Наиболее подходящими инструментами реализации функций, требуемых от доверенных компонентов, представляются следующие криптографические методы:

• полностью и частично гомоморфные схемы шифрования (в части алгоритмов зашифрования и расшифрования данных);

• схемы разделения секрета (операции по разделению секретов на доли и восстановлению секретов из долей);

• доказательства с нулевым разглашением (алгоритмы проверки доказательств).

КЛАССИФИКАЦИОННЫЕ ПРИЗНАКИ И КРИТЕРИИ ОЦЕНКИ СИСТЕМ КОНФИДЕНЦИАЛЬНОГО МАШИННОГО ОБУЧЕНИЯ

Анализ публикаций, посвященных разработке и реализации систем конфиденциального машинного обучения, позволяет выделить ряд критериев, по которым возможна их систематизация и оценка. К основным из них можно отнести признаки, перечисленные ниже.

1. *Количество сторон в протоколах*, реализующих функциональность систем конфиденциального машинного обучения:

• двусторонние;

• многосторонние (в настоящее время исключительно трех- и четырехсторонние).

Некоторые системы конфиденциального машинного обучения позволяют реализовывать функциональность посредством протоколов с разным числом участников, достигая компромисса между требованиями информационной безопасности и производительности.

2. *Криптографические примитивы*, используемые для реализации системы:

• схемы разделения секрета;
• GC-схемы (garbled circuits);
• схемы гомоморфного шифрования.

Для большинства систем характерно сочетание двух или даже всех трех перечисленных типов криптографических примитивов, хотя есть попытки построить системы конфиденциального машинного обучения, используя только один вид примитивов, но они, как правило, либо обладают ограниченной функциональностью, либо показывают неудовлетворительную производительность.

3. *Модель нарушителя*, в предположении о которой разрабатывалась система конфиденциального машинного обучения и в которой обеспечивается ее криптографическая стойкость:

• получестный нарушитель;
• злоумышленный нарушитель.

Системы, основанные на двусторонних криптографических протоколах, обеспечивают стойкость преимущественно к получестному нарушителю. В то же время значительное большинство систем, основанных на трех- и четырехсторонних протоколах обеспечивает стойкость как к получестному, так и к злоумышленному нарушителю.

4. *Поддержка стадий жизненного цикла машинного обучения*:

• обучение моделей (training);
• применение моделей для получения прогнозных ответов на запросы пользователей (inference).

Фаза обучения моделей на несколько порядков величины более трудоемка, чем фаза применения. При этом обучение модели – относительно нечасто выполняемая операция по сравнению с последующим применением обученной модели. Большая часть систем конфиденциального машинного обучения в настоящее время поддерживает лишь стадию применения уже обученных моделей, но ряд систем поддерживают обе стадии.

5. *Поддержка различных концепций, методов и алгоритмов машинного обучения*:

• относительно простых статистических и логических методов машинного обучения, таких как линейная регрессия, логистическая регрессия, кластеризация, решающие деревья;
• полносвязных нейронных сетей;
• глубоких нейронных сетей;
• специальных приемов обучения и применения нейронных сетей для повышения точности прогнозирования, производительности, сходимости параметров сети и т.п., таких как пакетная нормализация, оптимизация методом Adam и пр.

Среди систем конфиденциального машинного обучения преобладают разработки для обеспечения безопасности глубоких нейронных сетей, прежде всего, сверточных. Растет количество работ, посвященных обеспечению конфиденциальности при использовании специальных приемов обучения нейросетей, часто применяемых на практике.

6. *Обеспечиваемый системой набор свойств информационной безопасности*:

• конфиденциальность вычислений при получении прогнозных ответов на запросы пользователей к обученной модели;
• конфиденциальность вычислений при обучении моделей и последующем получении прогнозных ответов на запросы пользователей к этой модели;
• конфиденциальность и проверяемость (верифицируемость) вычислений при взаимодействии с моделью.

7. *Пригодность для использования в различных коммуникационных архитектурах*:

• локальных компьютерных сетях (LAN);
• глобальных компьютерных сетях (WAN).

Системы конфиденциального машинного обучения, которые реализуются посредством протоколов безопасных многосторонних вычислений с большой коммуникационной сложностью, а также со сбалансированными требованиями к коммуникационным и вычислительным ресурсам участников значительно лучше подходят для локальных сетей. Системы, предназначенные для глобальных сетей, должны минимизировать коммуникационные требования к участникам, сместив их в сторону более высоких вычислительных требований.

8. *Объем массивов данных*, использованных при апробации и тестировании систем конфиденциального машинного обучения:

• эталонные массивы данных (датасеты) относительно малого объема, такие как MNIST,

CIFAR-10 и т.п.;

• сверхбольшие и постоянно растущие массивы данных («большие данные»).

Многие системы конфиденциального машинного обучения, которые показывают хорошие результаты при экспериментах на относительно малых массивах данных, могут показать неприемлемо большое время работы на массивах данных, представляющих практический интерес. В связи с этим большое значение имеет апробация прототипов систем конфиденциального машинного обучения на массивах данных объема, сопоставимого с тем, который будет встречаться при практическом использовании.

9. *Архитектуры нейросетей*, для которых апробированы системы конфиденциального машинного обучения:

• относительно простые нейросети с небольшим количеством слоев (например, LeNet, AlexNet и т.п.);

• глубокие нейросети с числом слоев порядка 50 – 200 (например, VGG-16, ResNet, DenseNet).

Практический интерес представляют системы, способные работать с глубокими нейросетями (100 – 200 слоёв и более), получившими большее практическое применение.

КЛАССИФИКАЦИЯ МЕТОДОВ И СИСТЕМ, ОБЕСПЕЧИВАЮЩИХ БЕЗОПАСНОСТЬ МАШИННОГО ОБУЧЕНИЯ

Обеспечение конфиденциальности становится актуальным для систем машинного обучения, как только в процессе обучения участвуют две или более стороны, в частности, либо клиенты и обрабатывающие центры, либо несколько равноправных участников. Именно поэтому на ранних этапах развития машинного обучения, когда оно воспринималось почти всегда как локальный процесс, выполняемый одним исполнителем, проблемой конфиденциальности пренебрегали. Иными словами, проблема конфиденциальности возникает тогда, когда не совпадают по крайней мере две из трех ролей участников процесса машинного обучения:

• источника обучающих и/или тестовых выборок;

• вычислителя, обучающего либо применяющего модель;

• получателя ответов на запросы к обученной модели.

Основными инструментами обеспечения конфиденциальности в распределенных компьютерных системах являются криптографические методы и алгоритмы. В связи с этим далее в качестве основного классификационного признака примем криптографические методы, которые послужили основой алгоритмического обеспечения той или иной системы конфиденциального машинного обучения.

1. *Системы конфиденциального машинного обучения на основе схем гомоморфного шифрования*. Только системы конфиденциального машинного обучения на базе гомоморфного шифрования способны функционировать в односерверной конфигурации (хотя и не все из них). В то же время основной недостаток таких систем – очень высокие вычислительные затраты, приводящие к неприемлемому для практического использования времени обучения моделей и последующего получения с их помощью прогнозных ответов на запросы пользователей.

Примерами таких систем являются MiniONN [3], SecureML [4], Glyph [5], система Lee – Kang – Lee и др. [6].

2. *Системы конфиденциального машинного обучения на основе двусторонних протоколов безопасных вычислений*.

В качестве примеров таких систем можно привести ABY [7], ABY 2.0 [8], EzPC [9], CrypTFlow2 [10], Gazelle [11], Delphi [12], Muse [13].

3. *Системы конфиденциального машинного обучения на основе трехсторонних протоколов безопасных вычислений*.

Примеры таких систем – ABY3 [14], CrypTFlow [15], SecureNN [16], Falcon [17], система Attrapadung – Hamada – Ikarashi и др. [18].

3. *Системы конфиденциального машинного обучения на основе четырехсторонних протоколов безопасных вычислений*.

К таким системам относятся Flash [19], Trident [20], Tetrad [21].

4. *Системы конфиденциального машинного обучения, использующие доказательства с нулевым разглашением*. Такие системы, по-

мимо основной функции – обеспечения конфиденциальности информации в процессе использования технологий машинного обучения – обеспечивают дополнительное свойство – проверяемость (верифицируемость) вычислений. Большинство известных систем конфиденциального и проверяемого машинного обучения направлено на обеспечение свойств конфиденциальности и проверяемости для искусственных нейронных сетей. Это легко объяснимо, поскольку искусственные нейронные сети – самый популярный и распространенный инструмент машинного обучения.

В качестве примеров систем конфиденциального и проверяемого машинного обучения можно назвать системы VeriML [22], vCNN [23], ZEN [24], zkCNN [25], Mystique [26], систему Zhang – Fang – Zhang и др. [27].

Каждая из перечисленных систем – это, как правило, весьма нетривиальная алгоритмическая конструкция, заслуживающая отдельного рассмотрения. В то же время следует отметить, что практически все названные системы конфиденциального машинного обучения в настоящее время реализованы на стадии прототипов (экспериментальных образцов). Систем, находящихся в промышленной эксплуатации, пока не существует.

ЗАКЛЮЧЕНИЕ

В статье изложены предварительные результаты, касающиеся определения предметной области информационной безопасности технологий ИИ. Охарактеризованы основные составляющие подхода к обеспечению информационной безопасности технологий и систем ИИ.

В частности, выделены характерные угрозы системам ИИ, типичные модели нарушителей в системах ИИ, выделены особенности обработки информации с использованием технологий ИИ, определяющие задачи обеспечения информационной безопасности. Приведена постановка задачи конфиденциального машинного обучения. Выделены классификационные признаки и критерии оценки систем конфиденциального машинного обучения, а также проведена классификация известных методов и систем конфиденциального и проверяемого машинного обучения.

Перспективы продолжения исследований в этой области могут состоять как в разработке общетеоретических положений в сфере обеспечения информационной безопасности ИИ, взаимоувязанных с понятием доверенного ИИ, так и в прикладных исследованиях с целью создания алгоритмов и протоколов взаимодействия участников систем ИИ, обеспечивающих достижение различных свойств информационной безопасности: конфиденциальности, целостности, доступности, невозможности отказа, анонимности, разграничения доступа и др. В настоящее время такие конструкции ограничены лишь протоколами конфиденциального и проверяемого машинного обучения для некоторых элементарных методов машинного обучения, а также для свёрточных нейронных сетей. Представляется, что информационная безопасность ИИ как сфера научно-технической деятельности имеет большой потенциал развития в ближайшем будущем.

СПИСОК ЛИТЕРАТУРЫ

1. Aggarwal, C. Artificial intelligence: A textbook. Springer, 2021. 496 pp. ISBN 978-3-030-72356-9.
2. CS221: Artificial Intelligence: Principles and Techniques. Stanford university course. URL: https://stanford-cs221.github.io/autumn2020 (дата обращения: 13.09.2021).
3. Liu, J. Oblivious neural network predictions via MiniONN transformations / J. Liu, M. Juuti, Y. Lu, and N. Asokan // ACM CCS 2017, B. M. Thuraisingham, D. Evans, T. Malkin, and D. Xu, Eds. ACM Press, Oct. 2017, pp. 619–631.
4. Mohassel, P. SecureML: A system for scalable privacy-preserving machine learning / P. Mohassel, Y. Zhang // Cryptology ePrint Archive. 2017. 38 pp. URL: https://eprint.iacr.org/2017/396 (дата обращения:

13.09.2021).

5. Lou, Q. Glyph: Fast and accurately training deep neural networks on encrypted data / Q. Lou, B. Feng, G. C. Fox, L. Jiang // Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., 2020. URL: https://proceedings.neurips.cc/paper/2020/hash/685ac8cadc1be5ac98da9556bc1c8d9e-Abstract.html (дата обращения: 13.09.2021)

6. Lee, J.-W. Privacy-preserving machine learning with fully homomorphic encryption for deep neural network / J.-W. Lee, H.Kang, Y. Lee et al. // Cryptology ePrint Archive. 2021. 12 pp. URL: https://eprint.iacr.org/2021/783 (дата обращения: 13.09.2021)

7. Demmler, D. ABY – a framework for efficient mixed-protocol secure two-party computation / D. Demmler, T. Schneider, M. Zohner // 22nd Network and Distributed System Security Symposium (NDSS'15), Internet Society, San Diego, CA, USA, February 8-11, 2015. URL: https://encrypto.de/papers/DSZ15.pdf (дата обращения: 13.09.2021)

8. Patra A. ABY2.0: Improved mixed-protocol secure two-party computation / A. Patra, T. Schneider, A. Suresh et al. URL: https://ia.cr/2020/1225 (дата обращения: 13.09.2021)

9. Chandran, N. EzPC: Programmable and Efficient Secure Two-Party Computation for Machine Learning / N. Chandran, D. Gupta, A. Rastogi, R. Sharma, S. Tripathi // 2019 IEEE European Symposium on Security and Privacy (EuroS&P), Stockholm, Sweden, 2019, pp. 496-511, doi: 10.1109/EuroSP.2019.00043.

10. Rathee D. et al. CrypTFlow2: Practical 2-Party Secure Inference. arXiv preprint. 2020. 18 pp. URL: https://arxiv.org/pdf/2010.06457.pdf (дата обращения: 13.09.2021)

11. Juvekar, C. GAZELLE: A Low Latency Framework for Secure Neural Network Inference / C. Juvekar, V. Vaikuntanathan, A. Chandrakasan // Cryptology ePrint Archive. 2021. 17 pp. URL: https://eprint.iacr.org/2018/073.pdf (дата обращения: 13.09.2021).

12. Mishra, P. Delphi: A Cryptographic Inference Service for Neural Networks / P. Mishra, R. Lehmkuhl, A. Srinivasan et al. // Proc. of USENIX Security 2020 (USENIX Security Symposium). URL: https://www.usenix.org/system/files/sec20spring_mishra_prepub.pdf (дата обращения: 13.09.2021).

13. Lehmkuhl, R. Muse: Secure Inference Resilient to Malicious Clients.
/ R. Lehmkuhl, P. Mishra, A. Srinivasan et al. // Proc. of USENIX Security 2021 (USENIX Security Symposium). URL: https://people.eecs.berkeley.edu/ ~raluca/MUSEcamera.pdf (дата обращения: 13.09.2021).

14. Mohassel, P. ABY3: A mixed protocol framework for machine learning / P. Mohassel, P. Rindal // Cryptology ePrint Archive. 2018. – 40 pp. URL: https://eprint.iacr.org/2018/403 (дата обращения: 13.09.2021).

15. Kumar E. et al. CrypTFlow: Secure TensorFlow Inference. arXiv preprint. 2020. 18 pp. URL: https://arxiv.org/pdf/1909.07814v2.pdf (дата обращения: 13.09.2021)

16. Wagh, S. SecureNN: Efficient and private neural network training / S. Wagh, D. Gupta, N. Chandran // Cryptology ePrint Archive. 2018. 24 pp. URL: https://eprint.iacr.org/2018/442 (дата обращения: 13.09.2021).

17. Wagh, S. Falcon: Honest-Majority Maliciously Secure Framework for Private Deep Learning / Sameer Wagh, Shruti Tople, Fabrice Benhamouda, et al. // Proc. of Privacy Enhancing Technologies Symposium (PETS), June 2021. Pp. 1 – 21. URL: https://arxiv.org/pdf/ 2004.02229.pdf (дата обращения: 13.09.2021).

18. Attrapadung, N. Adam in Private: Secure and Fast Training of Deep Neural Networks with Adaptive Moment Estimation / N. Attrapadung, K. Hamada, D. Ikarashi, et al. // Cryptology ePrint Archive. 2021. 24 pp. URL: https://eprint.iacr.org/2021/736.pdf (дата обращения: 13.09.2021).

19. Byali, M. FLASH: Fast and robust framework for privacy-preserving machine learning / M. Byali, H. Chaudhari, A. Patra, et al. // Cryptology ePrint Archive. 2019. 29 pp. URL: https://eprint.iacr.org/2019/1365 (дата обращения: 13.09.2021).

20. Rachuri, R. Trident: Efficient 4PC framework for privacy preserving machine learning / R. Rachuri, A. Suresh // Cryptology ePrint Archive. 2019. 26 pp. URL: https://eprint.iacr.org/2019/1315 (дата обращения: 13.09.2021).

21. Koti, N. Tetrad: Actively Secure 4PC for Secure Training and Inference / N. Koti, A. Patra, R. Rachuri, et al.

// Cryptology ePrint Archive. 2021. 31 pp. URL: https://eprint.iacr.org/2021/755.pdf (дата обращения: 13.09.2021).

22. Zhao L., QianWang, CongWang, et al. VeriML: Enabling Integrity Assurances and Fair Payments for Machine Learning as a Service. - URL: https://arxiv.org/pdf/1909.06961v1.pdf (дата обращения: 13.09.2021).

23. Lee S., Ko H., Kim J., Oh H. vCNN: Verifiable convolutional neural network based on zk-SNARKs. URL: https://eprint.iacr.org/2020/584 (дата обращения: 13.09.2021).

24. Feng B., Qin L., Zhang Z. et al. ZEN: An optimizing compiler for verifiable, zero-knowledge neural network inferences. URL: https://eprint.iacr.org/2021/087 (дата обращения: 13.09.2021).

25. Liu T., Xie X., Zhang Y. zkCNN: Zero knowledge proofs for convolutional neural network predictions and accuracy. URL: https://eprint.iacr.org/2021/673 (дата обращения: 13.09.2021).

26. Weng C., Yang K., Xie X. et al. Mystique: Efficient conversions for zero-knowledge proofs with applications to machine learning. URL: https://eprint.iacr.org/2021/730 (дата обращения: 13.09.2021).

27. Zhang J., Fang Z., Zhang Y., Song D. Zero knowledge proofs for decision tree predictions and accuracy // CCS '20: Proceedings of the 2020 ACM SIGSAC Conference on Computer and Communications Security, 2020. P. 2039–2053. Doi: 10.1145/3372297.3417278.