Предложения
в проект требований по обеспечению информационной безопасности в системах, реализующих искусственный интеллект

I. Общие замечания

Данные предложения в проект требований по обеспечению информационной безопасности в системах, реализующих искусственный интеллект (далее – Предложения), разработаны на основе сформированной Академией криптографии Российской Федерации научной базы для современных защищенных технологий и систем искусственного интеллекта, применяемых в государственных информационных системах. С учетом необходимости эффективного внедрения в рамках национальной программы «Цифровая экономика Российской Федерации» сквозной технологии искусственного интеллекта (далее – ИИ) для взаимодействия государства, бизнеса и науки в Предложениях не рассматриваются требования к системам, в которых обрабатываются сведения, составляющие государственную тайну.

Обеспечение информационной безопасности технологий искусственного интеллекта, и, в частности, наиболее распространенного варианта таких технологий – машинного обучения, является новой научно-технической задачей. Это связано с зависимостью алгоритмов искусственного интеллекта от обучающих данных, что отличает их от классических детерминированных алгоритмов и порождает спектр новых угроз безопасности информации.

Как следствие, для систем, реализующих технологии искусственного интеллекта, актуальными являются как традиционные угрозы информационной безопасности, так и угрозы, определяемые специфическими свойствами обработки данных в таких системах.

Как было отмечено, в настоящее время технологии машинного обучения являются доминирующими и наиболее практически значимыми. В связи с этим требования к обеспечению информационной безопасности технологий машинного обучения являются базовыми (основополагающими) для внедряемых в настоящее время систем ИИ. В системах, реализующих ИИ, применяются различные типы, методы и схемы машинного обучения, опирающиеся на общую структуру жизненного цикла данных. Это позволяет рассматривать Предложения как универсальный механизм/методический подход обеспечения минимального/базового уровня информационной безопасности систем, реализующих ИИ, вне зависимости от класса защищенности системы, конкретных способов проведения атак и методов защиты от них.

Представленная далее обобщенная модель угроз безопасности информации содержит модель нарушителя, перечень свойственных технологиям машинного обучения типовых угроз безопасности информации и методологические принципы их реализации, а также предложения в проект требований по обеспечению информационной безопасности систем, реализующих ИИ.

Формирование требований к информационной безопасности систем, реализующих ИИ, в конкретных условиях отраслевого применения должно осуществляться соответствующими профильными организациями на основе известных данных о классах защищенности этих систем и актуальных моделях угроз безопасности с учетом действующих нормативных, методических и нормативно-технических документов в сферах защиты информации, искусственного интеллекта и создания информационных систем, утвержденных уполномоченными органами Российской Федерации. В частности, с учетом приказа ФСБ России от 24 октября 2022 г. № 524 «Об утверждении Требований о защите информации, содержащейся в государственных информационных системах, с использованием шифровальных (криптографических) средств» и приказа ФСТЭК России от 11 февраля 2013 г. № 17 «Об утверждении Требований о защите информации, не составляющей государственную тайну, содержащейся в государственных информационных системах».

Обоснование необходимости использования конкретных организационно-технических мер и средств криптографической защиты информации должно отражаться в модели угроз безопасности информации, технических проектах и технических заданиях на создание данных систем в конкретных условиях отраслевого применения.

Настоящие Предложения могут использоваться в качестве вспомогательных материалов при формировании нормативно обязательных требований к информационной безопасности систем, реализующих ИИ в конкретных условиях отраслевого применения, в которых не хранятся и не обрабатываются сведения, составляющие государственную тайну.

II. Используемые понятия

Для единообразного понимания данных Предложений используются следующие понятия.

Алгоритм машинного обучения: алгоритм определения параметров модели машинного обучения в соответствии с заданными критериями на основе данных.

Атака: целенаправленные действия нарушителя с целью нарушения функционирования и/или заданных характеристик системы, реализующей ИИ, или с целью создания условий для этого.

Валидационная выборка: набор данных, который используется в процессе обучения для подбора оптимального набора гиперпараметров.

Входные данные: данные, на основе которых система, реализующая ИИ, получает в качестве результата прогноз или логический вывод.

Гиперпараметр: параметр алгоритма машинного обучения, влияющий на процесс обучения модели.

Примечание 1. Гиперпараметры выбираются до начала обучения модели и могут использоваться для помощи в оценке ее параметров.

Примечание 2. Примерами гиперпараметров могут служить количество слоев нейронной сети, ширина каждого слоя, тип функции активации, метод

оптимизации, скорость обучения нейронных сетей; выбор функции ядра в методе опорных векторов; количество листьев или высота дерева; значение параметра K при кластеризации методом K-средних; максимальное количество итераций алгоритма максимизации ожидания; количество гауссианов в гауссовой смеси.

Искусственный интеллект: экспериментальная междисциплинарная научно-инженерная область знаний, входящая в комплекс компьютерных наук и относящаяся к информационным технологиям, направленная на исследование, разработку и внедрение методов, моделей и программных средств, позволяющих искусственным системам и устройствам реализовывать разумные рассуждения и целенаправленное поведение, свойственное человеку, а также решать комплекс задач, для которых не существует эффективных алгоритмов (в математическом смысле понятия «алгоритм»).

Контролируемая зона: территория или пространство, на которых установлены сервера или базы данных системы, реализующей ИИ, и осуществляется контроль за пребыванием и действиями лиц.

Метка: значение целевой переменной, присвоенное неделимому элементу размеченных входных данных.

Метрика качества: функционал, значение которого показывает, насколько хорошо модель способна решать поставленную задачу.

Модель машинного обучения: математическая конструкция, обладающая набором внутренних настраиваемых параметров, генерирующая логический вывод или прогноз на основе входных данных.

Нарушитель: физическое лицо или группа лиц, которые в результате предумышленных и/или непредумышленных действий обеспечивают реализацию угроз системе, реализующей ИИ, на различных этапах жизненного цикла данных.

Обучающая выборка: совокупность данных, которые используются для обучения модели.

Обучение: процесс оптимизации параметров модели с помощью вычислительных методов таким образом, чтобы поведение модели отражало данные и/или опыт.

Подготовка данных: процесс преобразования описания данных из одного формата в другой, в том числе для возможности их передачи на вход модели.

Признак: элемент вектора (как правило, вещественных значений) данных, подаваемого на вход модели и являющегося описанием анализируемого объекта.

Проверка качества данных: процесс, в ходе которого данные проверяются на полноту, предвзятость и наличие иных факторов, влияющих на их полезность для системы, реализующей ИИ.

Промпт: текстовый запрос специального вида к нейросети, содержащий инструкцию, определяющую задачу, которую должна решить сеть, и, возможно, некую дополнительную управляющую информацию (формат выдачи результата и т.п.).

Разметка данных: процесс присвоения каждому элементу данных конкретной метки из заранее заданного ограниченного множества.

Система, реализующая искусственный интеллект: техническая система, реализующая технологии ИИ, например, машинного обучения.

«Теневая» модель машинного обучения: модель машинного обучения, построенная на основе анализа ответов на запросы к целевой модели, которая (в смысле заранее определенной метрики качества) близка к целевой модели.

Тестовая выборка: набор данных, на которых модель не обучалась, но который используется для оценки ее качества и точности предсказаний.

Технологии искусственного интеллекта: приемы, способы и методы применения искусственного интеллекта (в частности, машинного обучения) при выполнении функций сбора, хранения, обработки, передачи и использования данных и знаний в процессе решения конкретных прикладных задач.

Угроза безопасности системы, реализующей ИИ: потенциально возможное событие, действие или процесс, которые могут повлиять на процесс функционирования этой системы.

Целевая модель: модель, в отношении которой реализуется атака.

Эксплуатационные данные, рабочие данные: данные, приобретенные на стадии эксплуатации системы, реализующей ИИ, для которых развернутая система получает в качестве результата прогноз или логический вывод.

III. Жизненный цикл данных в системе, реализующей ИИ

Для целей дальнейшего изложения используется обобщенная модель жизненного цикла данных в системе, реализующей ИИ, состоящая из этапов, приведенных в Таблице 1. Такая модель не является линейной в том смысле, что при создании указанных систем некоторые этапы или последовательности этапов могут многократно повторяться в процессе поиска оптимального решения поставленной задачи.

Таблица 1.

№ п/п	Наименование этапа
1	Описания характеристик системы
2	Выбор программно-аппаратных средств машинного обучения
3	Сбор данных
4	Предварительная обработка и статистический анализ собранных данных
5	Выбор модели, алгоритмов машинного обучения и метрики качества
6	Приведение исходных данных к виду, который может быть подан на вход программам анализа данных
7	Отбор информативных признаков

8	Обучение
9	Эксплуатация
10	Вывод из эксплуатации

IV. Модель нарушителя

В условиях многообразия отраслевого применения систем, реализующих ИИ, принадлежность нарушителя к какой-либо профессиональной или социальной группе не может рассматриваться как универсальная характеристика, определяющая его потенциал. Типы нарушителей и их потенциал целесообразно определять исходя из функциональных возможностей, которыми нарушитель может обладать на этапах обучения и эксплуатации модели машинного обучения [1].

Нарушители реализуют угрозы путем непосредственного доступа к системе в пределах контролируемой зоны (внутренние нарушители), а также путем проведения атак из-за пределов системы, например, используя сети общего пользования (внешние нарушители).

1. Типы нарушителя

1.1 Внешний нарушитель

Предполагается, что параметры модели машинного обучения внешнему нарушителю неизвестны, однако он может обладать общими сведениями о предназначении и архитектуре модели, количестве параметров и используемых алгоритмах обучения. Это, в частности, связано с тем, что при решении прикладных задач могут использоваться имеющиеся в свободном доступе заранее предобученные модели из некоторого известного класса. Внешний нарушитель может иметь информацию о типе, форматах входных данных, распределении таких данных. Он может иметь сведения о распределении обучающих данных, может иметь доступ к части обучающих данных и полный доступ к рабочим данным. Нарушитель может получать и анализировать ответы модели на вводимые им данные.

На этапе обучения и предшествующих этапах жизненного цикла данных внешний нарушитель с целью нарушения целостности и/или доступности модели может реализовывать атаки путем добавления в обучающие выборки специальным образом сформированных данных с учетом доступной ему информации о модели и ее предназначении, но не может иметь доступа к результатам ее обучения. В частности, для построения атак он может использовать модели из того же класса, которому принадлежит атакуемая модель.

Нарушитель может формировать данные следующим образом:

– добавлением в обучающую выборку новых данных, как естественных, так и синтезированных;

– модификацией меток существующих данных в случае использования обучения с учителем;

– модификацией данных обучающей выборки;

– удалением определенных данных из обучающей выборки.

Целью нарушителя на данном этапе может являться:

– нарушение целостности ответов модели для заранее заданных входных примеров, формируемое «отравлением» обучающей выборки;

– нарушение качества ответов модели для заранее заданных входных примеров, формируемое «отравлением» обучающей выборки.

Параметрами атак в данном случае являются: объем сформированных данных, вероятность возникновения навязываемого нарушителем события (классификации) в случае нарушения целостности или при нарушении доступности модели – разность характеристики качества модели без сформированных данных и с ними.

На этапе эксплуатации модели внешний нарушитель имеет возможность подавать на ее вход данные и получать результаты работы модели, при этом он может формировать очередные входные данные в зависимости от результатов анализа предшествующих запросов, составляя таким образом адаптивную последовательность запросов к модели. Под результатом работы модели понимается значение некоторой функции от выходного вектора модели, например, латентного вектора, вектора вероятности классов и т.п. Вид функции определяется функциональным предназначением системы и реализуемыми мерами по контролю доступа. В наиболее благоприятном для нарушителя случае ему доступен выходной вектор полностью, в наиболее неблагоприятном – прикладной результат (номер определенного класса, ответ вида «да/ нет»).

Целью нарушителя на данном этапе может являться:

– восстановление параметров модели;

– проверка принадлежности обучающему множеству;

– обращение модели;

– извлечение данных из генеративных моделей;

– построение состязательного примера.

1.2 Внутренний нарушитель

Предполагается, что внутреннему нарушителю известны обучающие данные и характеристики модели машинного обучения (процедура обучения, архитектура сети и т.д.). Внутренний нарушитель может иметь полный доступ к целевой модели и получать всю информацию, включая параметры обученной модели машинного обучения.

Внутренний нарушитель может проводить те же атаки, что и внешний, но в отличии от последнего может наблюдать промежуточные результаты вычислений модели, ее параметры и строить запросы с учетом таких знаний.

2. Потенциал нарушителя

На этапе обучения (в контексте данного документа рассматриваемого как совокупность этапов сбора, подготовки и анализа данных, и собственно обучения) целесообразно различать три типовых нарушителя (ТН) в зависимости от их возможностей:

ТН1 – доступ на чтение данных;

ТН2 – доступ на запись данных (добавление и удаление);

ТН3 – прямой контроль процесса обучения.

На этапе эксплуатации системы, реализующей ИИ, нарушитель напрямую взаимодействует с уже обученной моделью. Здесь можно различать случаи, когда нарушитель обладает знаниями о целевой модели («белый ящик») и не обладает этими знаниями («черный ящик»).

В случае «черного ящика» все, что способен делать нарушитель – это подавать на вход разные примеры и анализировать предсказания модели.

В случае «белого ящика» нарушителю могут быть известны структура и тип модели, а также ее параметры. Указанные знания помогают ему быстрее и эффективнее достигать своих целей.

Отдельно следует отметить фактор обладания знаниями о конкретных платформах машинного обучения, которые использовались для подготовки и обучения модели. Уязвимости программного обеспечения таких платформ могут приводить к отказу систем ИИ.

На этапе эксплуатации можно различать следующих типовых нарушителей в зависимости от их возможностей:

ТН4 – не обладающие конкретными знаниями о модели;

ТН5 – обладающие информацией о типе модели и ее архитектуре;

ТН6 – обладающие знаниями о параметрах модели и платформе машинного обучения, на которой она реализована.

В зависимости от потенциала, требуемого для реализации атак на системы, реализующие ИИ, нарушители могут подразделяться на нарушителей с низким, средним и высоким потенциалом нападения (Таблица 2).

Таблица 2.

Потенциал	**Типы нарушителей**
Низкий	ТН1, ТН4
Средний	ТН2, ТН5
Высокий	ТН3, ТН6

В Таблице 3 приведены примерные данные об информации, которая может быть доступна нарушителю на этапах жизненного цикла данных в зависимости от его потенциала.

Таблица 3.

№ п/п	Этап	**Внешний нарушитель**	**Потенциал**	**Внутренний нарушитель**	**Потенциал**
1	Описание характеристик системы	Не имеет возможность воздействия.		Имеет возможность выбора требований к безопасности создаваемой системы	высокий
2	Выбор программно-аппаратных средств машинного обучения	Имеет доступ к репозиториям свободно распространяемых библиотек, к документации и технологическому процессу производства аппаратных компонентов.	низкий	Имеет доступ к информации о целях и задачах создаваемой системы, требуемых характеристиках.	высокий
		Имеет возможность вносить изменения в программные модули фреймворков машинного обучения, спецификации аппаратных компонентов.	средний	Имеет возможность выбора программных и аппаратных компонентов для дальнейшего использования.	высокий
3	Сбор данных	Имеет доступ к данным одного или нескольких внешних источников.	низкий	Имеет доступ ко всем собранным данным.	средний
		Имеет возможность манипулировать данными внешних источников, воздействовать на канал загрузки данных в систему, реализующую ИИ.	средний	Имеет возможность манипулировать данными источников, извлекать из них информацию (например, персональные данные).	высокий
4	Предварительная обработка и статистический анализ собранных данных	Не имеет доступа к данным.		Имеет доступ к используемым процедурам, функциям и значениям результатов их применения.	средний
		Не имеет возможность воздействия.		Имеет возможность навязывать используемые процедуры и функции, модифицировать значения результатов их применения, извлекать из данных информацию (например, персональные данные).	высокий

5	Выбор модели, алгоритмов машинного обучения и метрики качества	Не имеет доступа к данным.		Имеет доступ к используемым моделям, алгоритмам и метрикам.	средний
		Не имеет возможность воздействия.		Имеет возможность навязывать используемые модели, алгоритмы и метрики.	высокий
6	Приведение исходных данных к виду, который может быть подан на вход программам анализа данных	Не имеет доступа к данным.		Имеет доступ к используемым статистическим процедурам, функциям и значениям результатов их применения.	средний
		Не имеет возможность воздействия.		Имеет возможность навязывать используемые для статистического анализа процедуры и функции, модифицировать значения результатов их применения, извлекать из данных информацию (например, персональные данные).	высокий
7	Отбор информативных признаков	Не имеет доступа к данным.		Имеет доступ к процедурам отбора признаков и значениям результатов их применения.	средний
		Не имеет возможность воздействия.		Имеет возможность навязывать используемые для машинного обучения признаки, извлекать из данных информацию (например, персональные данные)	высокий
8	Обучение	Не имеет доступа к данным.		Имеет доступ к процедурам и алгоритмам машинного обучения и результатам их работы.	средний
		Не имеет возможность воздействия.		Имеет возможность навязывать используемые процедуры машинного обучения, нарушать штатные режимы их работы, модифицировать результаты применения, извлекать из данных	высокий

				информацию (например, персональные данные).	
9	Эксплуатация	Имеет доступ к поступающим на вход модели данным и соответствующим результатам их обработки моделью машинного обучения.	средний	Имеет доступ к процедурам и алгоритмам машинного обучения и результатам их работы.	средний
		Имеет возможность формировать входные данные специального вида, для реконструкции обученной модели по ответам, определения членства объекта в обучающей выборке, формирования некорректных результатов.	средний	Имеет возможность модифицировать используемые процедуры машинного обучения, нарушать штатные режимы их работы, модифицировать результаты применения, извлекать из данных информацию (например, персональные данные).	высокий
10	Вывод из эксплуатации	Не имеет возможность воздействия.		Имеет возможность извлекать из данных информацию (например, персональные данные).	средний

Исходя из возможностей нарушения конфиденциальности и целостности данных, а также обученных моделей на различных этапах жизненного цикла данных, можно считать, что основное множество атак на системы, реализующие ИИ, приходится на этапы обучения и эксплуатации.

При этом внутренний нарушитель, являющийся аналитиком, обладает широким спектром возможностей по доступу к обучающим и промежуточным данным алгоритмов машинного обучения, к моделям машинного обучения и их параметрам, а также к модификации данных и алгоритмов.

Внешний нарушитель обладает существенно меньшими возможностями, поскольку имеет возможность влиять на функционирование системы, реализующей ИИ, в режиме «черного ящика» в основном за счет модификации входных данных.

V. Типовые угрозы информационной безопасности системам, реализующим ИИ

1. Нарушение конфиденциальности информации

Модели машинного обучения в настоящее время имеют большое число настраиваемых параметров, в связи с чем они обладают достаточной способностью запоминать информацию об обучающей выборке (т. н. эффект запоминания) [2, 3, 5, 6]. Это связано с тем, что наборы обучающих данных имеют конечный размер, и модели машинного обучения неоднократно обучаются в течение нескольких эпох (часто от десятков до сотен) на одних и тех же экземплярах обучающих данных. Так как параметры модели хранят статистически коррелированную информацию о конкретных записях данных в использованном наборе обучающих данных, то модели более точно предсказывают/классифицируют экземпляры обучающих данных по сравнению с элементами, не входящими в обучающую выборку [3, 4].

1.1 Нарушение конфиденциальности информации о гиперпараметрах обученной модели

В модели «черного ящика», когда нарушителю не известны параметры обученной модели, но он имеет возможность опрашивать ее путем подачи на вход данных и анализа ответов, возникает угроза восстановления нарушителем неизвестных параметров модели и построения для последующего использования имитирующей модели-аналога, т. н. «теневой» модели [7]. Для ряда моделей машинного обучения, например, использующих функции активации ReLU, такое восстановление может быть проведено достаточно эффективно при сравнительно малом числе запросов к модели [8].

1.2 Нарушение конфиденциальности обучающей выборки

1.2.1 Нарушение конфиденциальности обучающей выборки через проверку принадлежности примера обучающей выборке

Данный тип угроз [7, 9] реализуется через опрос нарушителем обученной модели машинного обучения, с целью выяснения присутствия экземпляров данных (с помощью которых производится опрос) в обучающей выборке. Может реализовываться как в модели «белого ящика», т. е. нарушителю известна обученная модель, так и в модели «черного ящика». В последнем случае нарушитель сначала создает «теневые» модели, которые имитируют поведение атакуемой модели, а затем уже для них проводит проверку принадлежности.

1.2.2 Нарушение конфиденциальности обучающей выборки через обращение обученной модели

Угроза возникает, когда нарушители для получения конфиденциальной информации из моделей машинного обучения используют результат работы модели и некоторую вспомогательную информацию для восстановления исходных входных данных [10]. Атаки, реализующие угрозу обращения обученной модели, могут реализовываться различными способами: через построение

обратной модели к атакуемой, построение генеративно-состязательной сети, в которой атакуемая сеть является классификатором, а генератор обучается в процессе атаки формировать примеры обучающей выборки.

1.2.3 Нарушение конфиденциальности обучающей выборки через извлечение данных из генеративных моделей путем формирования запросов специального вида (промптов)

Благодаря особенностям функционирования генеративных моделей, основанных на больших языковых моделях, нарушитель может восстановить данные обучающей выборки модели, даже если они появляются в обучающих материалах только один или несколько раз, посредством построения промптов – целевых запросов к модели [14]. Существуют также варианты проверки принадлежности обучающему множеству для моделей данного класса [12]. Чтобы языковая модель могла обрабатывать тексты, их обычно преобразуют в векторные представления. Инверсия векторного представления направлена на восстановление исходного входного текста. Атаки инверсии векторного представления особенно актуальны в контексте приложений, использующих языковые модели. В этих случаях необходимые для работы данные хранятся в виде вложений в соответствующих векторных базах данных, которые часто размещаются у внешних поставщиков услуг. В ряде случаев для получения требуемого ответа возможна адаптивная корректировка промптов [13].

2. Нарушение целостности информации

2.1 Нарушение целостности ответов модели для заранее заданных входных примеров, формируемое «отравлением» обучающей выборки

Угрозы данного типа связаны с созданием «бэкдоров» в обученных моделях. В этом случае в обучающую выборку вставляются специально сформированные данные, нацеленные на точечное искажение целевых показателей системы. Это может достигаться изменением меток данных или искажении самих примеров, например, добавлением шума [16, 17]. Подобные угрозы возникают и при трансферном обучении, когда модель обучается на основе заранее предобученной модели. В этом случае существует угроза переноса «бэкдоров» из базовой в новую модель [18].

2.2 Нарушение целостности ответов модели, формируемое входными данными специального вида (состязательными примерами)

Вероятностная природа моделей машинного обучения создает угрозу возможности формирования на этапе функционирования модели входных данных специального вида, некорректно обрабатываемых моделью [19, 21]. Для графических и аудио данных, это может быть достигнуто добавлением шума к некоторому исходному входному примеру. Для генеративных языковых моделей подобная угроза формируется благодаря так называемым промпт-инъекциям (запросам специального вида, заставляющим модель отклоняться от заданных правил работы), которые, например, могут приводить к отключению блокировок на выдачу нежелательного контента [14].

Можно рассматривать косвенные промпт-инъекции, которые, как и обычные указанные выше промпт-инъекции, направлены на изменение заранее определенного или усвоенного поведения модели посредством конкретных входных данных. Но их ключевое отличие состоит в том, что манипуляции производятся косвенно, через сторонние (непроверяемые) источники, а не самими пользователями [20]. Эта угроза реализуется, когда модель используется вместе с внешними источниками и приложениями для расширения его функциональности, позволяя данным из этих источников быть частью входных данных. В таких случаях нарушители могут использовать склонность языковых моделей к интерпретации текста как инструкции пользователя к действию.

3. Нарушение доступности информации

3.1 Нарушение качества ответов модели для заранее заданных входных примеров, формируемое «отравлением» обучающей выборки

Угрозы данного класса направлены на снижение доступности системы, реализующей ИИ. Для этого нарушитель, разбавляет «нормальную» обучающую выборку «аномальными» данными, которые приводят к общему снижению целевых качеств системы – точности распознавания, проценту ложноположительных срабатываний и т. д. [16, 17].

VI. Методологические принципы реализации типовых угроз информационной безопасности системам, реализующим ИИ

Для реализации угроз, нарушителю необходимо организовать взаимодействие с атакуемой моделью или, в случае невозможности такого взаимодействия – с моделью-аналогом (т. н. «теневой» моделью), которая в некотором приближении имитирует работу атакуемой модели.

Взаимодействие с моделью предполагает подачу на вход специально подготовленных данных и последующий анализ результатов работы модели. Это, в частности, позволяет создавать адаптивно сформированные последовательности запросов.

Эффективность атаки характеризуется вероятностью ее успешной реализации и средним количеством запросов к модели.

В случае отсутствия непосредственного доступа к модели нарушитель может пытаться сформировать «теневую» модель. Для этого он может использовать знания о распределении обучающих данных, а также использовать доступные модели из класса моделей, например, базовую модель класса, которому принадлежит атакуемая модель.

Нарушитель может сформировать т. н. «теневой» набор данных, содержащий записи данных того же типа и с тем же распределением, что и обучающие наборы. «Теневой» набор может быть получен путем синтеза с использованием методов математической статистики, когда распределение данных известно, или синтеза на основе модели машинного обучения, когда распределение данных неизвестно.

С использованием полученного «теневого» набора данных, нарушитель может пытаться обучить «теневую» модель.

На этапе обучения системы, реализующей ИИ, внутренний нарушитель, как правило, пытается повлиять на качество целевой модели. В самом простом случае у него есть доступ на чтение данных из обучающей выборки. Такой доступ может не только сам по себе наносить ущерб конфиденциальности данных, но и позволяет обучать свои «теневые» модели для анализа потенциальных уязвимых мест атакуемой системы.

В случае, если у нарушителя есть доступ на запись данных, то он может как изменять существующие данные, так и расширять обучающий набор своими данными. В задачах с обучением с подкреплением можно проводить модификации отдельных частей, например, окружения. Отдельно следует учитывать, насколько затратным для нарушителя может быть добавление собственных данных. Зачастую возможность напрямую добавлять цифровые данные отличается от искажения данных сенсоров в физическом мире, например, нарушителю гораздо проще напрямую добавлять графические файлы, чем воздействовать на видеокамеру в физическом мире.

Наконец, в самом неблагоприятном случае внутренний нарушитель может полностью контролировать процесс обучения модели. Например, у него есть возможности: влиять на выбор конкретного алгоритма, влиять на статистические распределения в обучающем и тестовом наборах, задавать гиперпараметры или менять метрику качества. В таком случае он полностью контролирует модель и потенциально всю систему.

VII. Предложения в проект требований по обеспечению информационной безопасности систем, реализующих ИИ

При формировании требований к обеспечению информационной безопасности систем, реализующих ИИ в конкретных условиях отраслевого применения, следует учитывать следующее.

Угрозы безопасности функционирования систем, реализующих ИИ, связаны с возможным влиянием нарушителей на входные данные моделей ИИ, на параметры обученных моделей и на сами модели машинного обучения.

Целями нарушителя могут быть как нарушение корректной работы системы, реализующей ИИ, так и извлечение данных из обучающих выборок и обученных моделей.

Системы, реализующие ИИ, имеют высокую уязвимость к внутреннему нарушителю, который имеет возможность доступа как к входным данным, так и к данным, формируемым в процессе обучения, а также непосредственно к обученным или обучаемым моделям. Внешний нарушитель имеет ограниченный доступ к модели в основном через вводимые им данные.

Использование стандартных моделей машинного обучения и процедур обучения таких моделей делает возможным реализацию широкого спектра атак, направленных как на извлечение данных, так и на изменение штатного функционирования системы, реализующей ИИ.

Требования к обеспечению информационной безопасности систем, реализующих ИИ, целесообразно рассматривать как требования к доступу и обработке данных, а также к доступу к моделям машинного обучения на этапах жизненного цикла данных.

Требования должны быть направлены на обеспечение защиты данных, обрабатываемых и формируемых на каждом из этапов их жизненного цикла, на защиту моделей машинного обучения и включать, в том числе, требования:

- выбора доверенных программно-аппаратных средств машинного обучения;
- аутентификации источника данных и их целостности при передаче в систему, реализующую ИИ;
- обеспечения конфиденциальности персональных или других подлежащих защите данных;
- к корректности разметки;
- обоснования корректности и полноты используемых критериев качества данных, а также корректности их реализаций;
- обеспечения защиты от навязывания некорректных результатов анализа данных, моделей, алгоритмов и метрик качества;
- обоснования корректности и адекватности выбора модели (класса модели), используемых алгоритмов обучения и метрик;
- обоснования корректности выбора и реализации используемых алгоритмов нормализации и очистки данных;
- обоснования корректности выбора информативных признаков;
- обоснования корректности реализаций используемых алгоритмов обучения;
- конфиденциальности обучаемой модели с целью минимизации рисков ее компрометации;
- журналирования процесса обучения и эксплуатации;
- контроля доступа пользователей к модели и данным.

На каждом этапе жизненного цикла данных описанные далее требования следует рассматривать с учетом архитектуры конкретной системы, в том смысле, что каждое требование может предъявляться, если архитектура системы и разработанная модель угроз подразумевают наличие условий возникновения соответствующих угроз безопасности информации.

1. Этап описания характеристик системы

Основой создания системы, реализующей ИИ, является определение архитектуры управления данными, ролей участников, типов и характеристик обрабатываемых данных, а также предполагаемых к решению разрабатываемой системой задач. Основными угрозами в данном случае являются неполное описание архитектуры и функционала системы, а также некорректная оценка угроз безопасности. В случае наличия внутреннего нарушителя, имеющего доступ к обучающим данным, это несет потенциальную угрозу безопасности системы в целом.

Должна быть определена архитектура системы, включая:

- описание используемых (предполагаемых к использованию) данных:
  - числовые, текстовые, графические и т. д.;
  - конфиденциальные/свободно распространяемые.
- источники данных:
  - внешние/внутренние;
  - доверенные/не доверенные.
- задача или набор задач, решаемая(ых) с помощью методов машинного обучения.
- перечень показателей качества решения такой(их) задач(и) и соответствующих пороговых значений.
- взаимодействие модели с внешними системами.
- описание ролей и полномочий обслуживающего персонала информационной системы:
  - инженеров по работе с данными;
  - аналитиков;
  - специалистов по машинному обучению;
  - администраторов;
  - пользователей.
- тип информационной системы:
  - централизованная/распределенная;
  - персональная/групповая/корпоративная.
- частные требования по защите информации используемых систем, реализующих ИИ.

Требования по защите информации (в т. ч. модели угроз и нарушителя) должны быть согласованы с требованиями по защите информации (в т. ч. моделями угроз и нарушителя), предъявляемыми к информационной системе в целом.

На данном этапе должна быть обеспечена конфиденциальность разработанных требований по безопасности системы, реализующей ИИ.

Должна быть обоснована релевантность разработанной модели угроз выбранной архитектуре и предполагаемым условиям эксплуатации. В частности, с целью обеспечения релевантности должен быть сформирован и на постоянной основе поддерживаться и обновляться перечень известных атак на системы, реализующие ИИ.

2. Этап выбора программно-аппаратных средств

При разработке требований необходимо учитывать, что реализация механизмов защиты в системах, реализующих ИИ, требует дополнительных ресурсов в виде физического оборудования, дополнительного объема памяти для хранения данных обучающих, валидационных и тестовых выборок, результатов промежуточных вычислений, а также дополнительной процессорной мощности на реализацию ресурсоемких механизмов защиты, например, таких как гомоморфное шифрование, операции над разделенными секретами и т. п.

Также следует исходить из того, что внешний нарушитель может иметь доступ к репозиториям свободно распространяемых библиотек, к документации и технологическому процессу производства аппаратных компонентов, а также может иметь возможность вносить изменения в программные модули фреймворков машинного обучения, спецификации аппаратных компонентов. Внутренний нарушитель может иметь доступ к информации о целях и задачах создаваемой системы, требуемых характеристиках, а также имеет возможность влиять на выбор программных и аппаратных компонентов для дальнейшего использования.

На данном этапе необходимо обеспечить выполнение следующих требований:

- Программные средства, используемые для предобработки данных, должны быть проверены на корректность реализации и отсутствие недекларированных возможностей.
- Должен быть обеспечен выбор доверенных программно-аппаратных средств машинного обучения, проведена их проверка на наличие недекларированных возможностей.
- В качестве требований по обеспечению корректности функционирования аппаратных платформ и операционных систем может выступать требование наличия сертификата безопасности для конкретной платформы и операционной системы.
- В отношении используемых фреймворков машинного обучения должно быть предусмотрено проведение поиска уязвимостей в заимствованном коде по открытым источникам, статический (автоматический анализ исходных текстов программы без реального выполнения кода) и динамический анализ (фаззинг-тестирование) или использование таких фреймворков из доверенных источников.
- Должно быть предусмотрено использование программно-аппаратных средств идентификации и аутентификации пользователей, разграничения прав доступа, журналирования событий, средств криптографической защиты информации и других средств защиты.
- Должно быть реализовано разграничение программно-аппаратных средств системы, реализующей ИИ, и остальных компонентов информационной системы.

3. Этап сбора данных

На данном этапе внешний нарушитель может иметь доступ к данным одного или нескольких внешних источников, а также иметь возможность манипулировать данными внешних источников, воздействовать на канал загрузки данных в систему, реализующую ИИ.

Внутренний нарушитель может иметь доступ ко всем собранным данным внешних источников и возможность манипулировать данными внешних источников, извлекать из них информацию (например, персональные данные).

На данном этапе необходимо обеспечить выполнение следующих требований:

- Должны быть обеспечены идентификация и аутентификация обслуживающего персонала, разграничение прав доступа и журналирование действий.
- Должны быть обеспечены идентификация и аутентификация источников данных, а также целостность данных при передаче в систему, реализующую ИИ, с целью недопущения внесения нарушителем отравленных данных.
- При сборе данных должны использоваться доверенные источники и сертифицированные устройства регистрации.
- Сбор данных должен осуществляться при соблюдении заранее установленных условий. Должен быть обеспечен контроль условий сбора данных. Заимствование данных из внешних источников должно осуществляться после обязательной оценки соблюдения источником установленных условий сбора и фильтрации.
- Данные должны поставляться в заранее определенном формате, включающем служебную информацию, которая должна содержать сведения об источниках, условиях сбора данных и их модификации. Должен быть обеспечен контроль целостности такой информации.
- При хранении и передаче данных, в том числе по каналам связи, выходящим за пределы контролируемой зоны, должны быть обеспечены контроль конфиденциальности и целостности данных.
- При наличии в обучающих данных информации, подлежащей защите в соответствии с действующим законодательством, например, персональных данных, должна быть обеспечена их конфиденциальность криптографическими методами, разграничением прав доступа, методами обезличивания. При этом параметры криптографических методов и методов разграничения прав доступа должны быть взаимоувязаны с соответствующими параметрами требований к информационной системе в целом; параметры методов обезличивания – с параметрами требований к информационной системе в целом и пороговыми значениями показателей качества системы, реализующей ИИ.

4. Этап предварительной обработки и статистического анализа собранных данных

На данном этапе осуществляются такие операции с данными как фильтрация, нормализация, проверка на качество (на отравление), формирование репрезентативной выборки, обогащение, разметка. Также для выбора возможных моделей машинного обучения осуществляется получение предварительных выводов о свойствах наблюдаемых данных на основе их статистического анализа.

Необходимо обеспечить защиту собранных данных от внутреннего нарушителя с целью предотвращения навязывания некорректных результатов анализа данных.

На данном этапе необходимо обеспечить выполнение следующих требований:

- Должны быть обеспечены идентификация и аутентификация обслуживающего персонала, разграничение прав доступа и журналирование действий.
- При наличии в обучающих данных информации, подлежащей защите в соответствии с действующим законодательством, например, персональных данных, должна быть обеспечена их конфиденциальность криптографическими методами, разграничением прав доступа, методами обезличивания. При этом параметры криптографических методов и методов разграничения прав доступа должны быть взаимоувязаны с соответствующими параметрами требований к информационной системе в целом; параметры методов обезличивания – с параметрами требований к информационной системе в целом и пороговыми значениями показателей качества системы, реализующей ИИ.
- Должна быть разработана методика оценки обеспечения качества данных, включающая набор тестов и граничных параметров таких тестов. Наборы тестов и граничные параметры должны быть взаимоувязаны с показателями качества создаваемой системы, реализующей ИИ, и пороговыми значениями таких показателей.
- В случае предъявления требований по защите от отравления данных методика оценки обеспечения качества данных должна включать тесты выявления отравленных данных. Наборы тестов и граничные параметры должны быть согласованы с перечнем известных методов построения отравленных данных, а также с показателями качества создаваемой системы, реализующей ИИ, и пороговыми значениями таких показателей.
- В случае предъявления требований по защите от состязательных атак формируемые наборы данных должны содержать наборы размеченных состязательных примеров. Количество и состав таких наборов должны определяться исходя из перечня известных состязательных атак и с учетом параметров требований по безопасности информационной системы в целом, а также быть взаимоувязаны с показателями качества создаваемой системы, реализующей ИИ, и пороговыми значениями таких показателей.
- При осуществлении экспертной разметки данных, должна быть разработана методика оценки качества разметки. Параметры такой методики должны быть согласованы с показателями качества создаваемой системы, реализующей ИИ, и пороговыми значениями таких показателей.
- При осуществлении экспертной разметки данных должна быть обеспечена идентификация и аутентификация экспертов, а также журналирование их действий. Должна быть разработана методика оценки качества разметки. Параметры такой методики должны быть согласованы с показателями качества создаваемой системы, реализующей ИИ, и пороговыми значениями таких показателей.

- При осуществлении автоматизированной разметки данных должна проводиться последующая оценка ее качества. Должна быть разработана методика оценки. Параметры такой методики должны быть согласованы с показателями качества создаваемой системы, реализующей ИИ, и пороговыми значениями таких показателей.
- При хранении и передаче данных, в том числе по каналам связи, выходящим за пределы контролируемой зоны, контроль целостности и защита от несанкционированного доступа должны быть обеспечены криптографическими методами.
- После каждого этапа обработки данных должен быть обеспечен контроль целостности промежуточного результата, вспомогательных данных, а также целостность сформированной обучающей выборки.

5. Этап выбора модели, алгоритмов машинного обучения и метрики качества

Следует учитывать, что внешний нарушитель на данном этапе не имеет доступа к данным и не имеет возможности воздействия. Внутренний нарушитель имеет доступ к используемым моделям, алгоритмам и метрикам, и имеет возможность навязывать используемые модели, алгоритмы и метрики.

На данном этапе необходимо обеспечить выполнение следующих требований:

- Выбор модели машинного обучения, алгоритмов машинного обучения и метрики качества должны определяться конкретной решаемой прикладной задачей и значением метрики качества. При этом необходимо обосновать корректность и адекватность выбора модели (класса модели), используемых алгоритмов обучения и метрик.
- При использовании ранее предобученных моделей необходимо обеспечить контроль их происхождения. Используемые модели машинного обучения должны быть получены из доверенных источников. Должны быть обоснованы свойства используемых моделей, в частности, с точки зрения противодействия их возможному отравлению.

6. Этап подготовки собранных данных (приведение исходных данных к виду, который может быть подан на вход программам анализа данных)

Внешний нарушитель, при осуществлении обработки данных в пределах контролируемой зоны, не имеет доступа к данным и не имеет возможность воздействия. Внутренний нарушитель имеет доступ к используемым статистическим процедурам и функциям и значениям результатов их применения. Имеет возможность навязывать используемые для статистического анализа процедуры и функции, модифицировать значения результатов их применения, извлекать из данных информацию (например, персональные данные).

На данном этапе необходимо обеспечить выполнение следующих требований:

- Необходимо обеспечить применение механизмов идентификации и аутентификации обслуживающего персонала, разграничения доступа и журналирования действий с целью предотвращения модификации данных при их очистке и нормализации.
- В случае использования подлежащих защите данных (например, персональных) необходимо обеспечить конфиденциальность их обработки с использованием криптографических методов, разграничения доступа, методов обезличивания. При этом параметры криптографических методов и методов разграничения прав доступа должны быть взаимоувязаны с соответствующими параметрами требований к информационной системе в целом; параметры методов обезличивания – с параметрами требований к информационной системе в целом и пороговыми значениями показателей качества системы, реализующей ИИ.
- Необходимо обосновать корректность выбора и реализации используемых методов нормализации и очистки. Параметры таких методов должны быть взаимоувязаны с пороговыми значениями показателей качества системы, реализующей ИИ.
- При хранении и передаче данных, в том числе по каналам связи, выходящим за пределы контролируемой зоны, контроль целостности и защита от несанкционированного доступа должны быть обеспечены криптографическими методами.

7. Этап отбора информативных признаков

Внешний нарушитель не имеет доступа к данным и не имеет возможность воздействия. Внутренний нарушитель имеет доступ к процедурам отбора признаков и значениям результатов их применения, имеет возможность навязывать используемые для машинного обучения признаки и извлекать из данных информацию (например, персональные данные).

На данном этапе необходимо обеспечить выполнение следующих требований:

- Необходимо обеспечить применение механизмов идентификации и аутентификации обслуживающего персонала, разграничения доступа и журналирования действий с целью предотвращения модификации данных.
- Необходимо обосновать корректность выбора информативных признаков. Выбранные информативные признаки должны обеспечивать достижение пороговых значений показателей качества системы, реализующей ИИ.
- В случае использования подлежащих защите данных (например, персональных), необходимо обеспечить конфиденциальность их обработки с использованием криптографических методов, разграничения доступа, методов обезличивания. При этом параметры криптографических методов и методов разграничения прав доступа должны быть взаимоувязаны с соответствующими параметрами требований к информационной си-

стеме в целом; параметры методов обезличивания – с параметрами требований к информационной системе в целом и пороговыми значениями показателей качества системы, реализующей ИИ.

- Необходимо обосновать выбор объемов обучающих, валидационных и тестовых наборов данных. Выбранные объемы должны обеспечивать достижение пороговых значений показателей качества системы, реализующей ИИ.
- Необходимо обеспечить контроль целостности промежуточных результатов, а также контроль целостности сформированных обучающих, валидационных и тестовых наборов данных.

8. Этап обучения

На этапе обучения внешний нарушитель, при осуществлении обработки данных в пределах контролируемой зоны, не имеет доступа к данным и не имеет возможность воздействия. Внутренний нарушитель имеет доступ к процедурам и алгоритмам машинного обучения и результатам их работы. Имеет возможность навязывать используемые процедуры машинного обучения, нарушать штатные режимы их работы, модифицировать результаты применения, извлекать из данных информацию (например, персональные данные).

На данном этапе необходимо обеспечить выполнение следующих требований:

- При хранении и передаче данных, в том числе по каналам связи, выходящим за пределы контролируемой зоны, должны быть обеспечены криптографическими методами контроль целостности и защита от несанкционированного доступа.
- Должны быть обеспечены идентификация и аутентификация обслуживающего персонала, разграничение прав доступа и журналирование действий.
- Должно быть обеспечено журналирование и контроль целостности промежуточных результатов обучения.
- В случае использования подлежащих защите данных (например, персональных), необходимо обеспечить конфиденциальность их обработки с использованием криптографических методов, разграничения доступа, методов обезличивания. При этом параметры криптографических методов и методов разграничения прав доступа должны быть взаимоувязаны с соответствующими параметрами требований к информационной системе в целом; параметры методов обезличивания – с параметрами требований к информационной системе в целом и пороговыми значениями показателей качества системы, реализующей ИИ.
- В случае предъявления требований по защите от атак, направленных на извлечение данных об обучающей выборке или параметрах модели, должны использоваться методы обучения, направленные на снижение эффективности таких атак. Характеристики таких методов должны быть согласованы с перечнем известных методов построения отравленных

данных, а также с показателями качества создаваемой системы, реализующей ИИ, и пороговыми значениями таких показателей.

- В случае предъявления требований по защите от состязательных атак должны использоваться методы обучения, направленные на снижение эффективности таких атак. Характеристики таких методов должны быть согласованы с перечнем известных методов построения отравленных данных, а также с показателями качества создаваемой системы, реализующей ИИ, и пороговыми значениями таких показателей.
- В случае предъявления требований по защите от отравляющих атак должны использоваться методы обучения, направленные на снижение эффективности таких атак. Характеристики таких методов должны быть согласованы с перечнем известных методов построения отравленных данных, а также с показателями качества создаваемой системы, реализующей ИИ, и пороговыми значениями таких показателей.
- Необходимо провести оценку устойчивости модели относительно известных атак указанных типов, если соответствующие требования предъявляются. По результатам такой оценки должны быть определены требования к параметрам доступа пользователей к модели, определяющие максимальное количество запросов пользователя к системе, а также объем информации, получаемый им о результате работы системы. Эти значения должны быть согласованы с показателями качества создаваемой системы, реализующей ИИ, и пороговыми значениями таких показателей.

9. Этап эксплуатации

Внешний нарушитель может иметь доступ к поступающим на вход модели данным и соответствующим результатам их обработки. Может формировать входные данные специального вида, для реконструкции обученной модели по ответам, определения членства объекта в обучающей выборке, формирования некорректных результатов работы модели, отказа от обслуживания.

Внутренний нарушитель может иметь доступ к процедурам и алгоритмам машинного обучения и результатам их работы. Может модифицировать используемые процедуры машинного обучения, нарушать штатные режимы их работы, модифицировать результаты применения, извлекать из данных информацию (например, персональные данные).

На данном этапе необходимо обеспечить выполнение следующих требований:

- В случае использования подлежащих защите данных (например, персональных), необходимо обеспечить конфиденциальность их обработки с использованием криптографических методов, разграничения доступа, методов обезличивания. При этом параметры криптографических методов и методов разграничения прав доступа должны быть взаимоувязаны с соответствующими параметрами требований к информационной системе в целом; параметры методов обезличивания – с параметрами тре-

бований к информационной системе в целом и пороговыми значениями показателей качества системы, реализующей ИИ.

- Необходимо обеспечить контроль целостности обученной модели.
- Необходимо обеспечить контроль доступа пользователей к модели по количеству запросов и/или точности ответов на них с целью противодействия восстановлению параметров модели с учетом значений указанных параметров, определенных на этапе обучения.
- Если предъявляются требования защиты от состязательных атак и атак извлечения информации об обучающей выборке, то необходимо обеспечить фильтрацию запросов к модели.
- Необходимо обеспечить защиту интерфейсов взаимодействия модели с внешними системами.
- При хранении и передаче данных, в том числе по каналам связи, выходящим за пределы контролируемой зоны, должны быть обеспечены криптографическими методами контроль целостности и защита от несанкционированного доступа.
- Должны быть обеспечены идентификация и аутентификация обслуживающего персонала и пользователей, разграничение прав доступа и журналирование действий.
- Необходимо обеспечить контроль качества работы модели.

10. Этап вывода из эксплуатации

При выводе модели из эксплуатации должны быть обеспечены уничтожение или конфиденциальность (в случае необходимости последующего использования) обученной модели и обучающих данных.

Список источников

1. Г. Маршалко. К вопросу построения требований к системам ДИИ. Форум «Технологии доверенного искусственного интеллекта», 2023 г.
2. N. Carlini, F. Tramer, E. Wallace, M. Jagielski, A. Herbert-Voss, K. Lee, A. Roberts, N. Brown, D. Song, U. Erlingsson, A. Oprea, C. Raffel (2020). Extracting Training Data from Large Language Models. 10.48550/arXiv.2012.07805.
3. S. K. Murakonda and R. Shokri (2020). ML Privacy Meter: Aiding Regulatory Compliance by Quantifying the Privacy Risks of Machine Learning. 10.48550/arXiv.2007.09339.
4. M. Nasr, R. Shokri, and A. Houmansadr. Comprehensive privacy analysis of deep learning: Passive and active white-box inference attacks against centralized and federated learning. In 2019 IEEE Symposium on Security and Privacy (SP). IEEE, May 2019.

5. C. Song, T. Ristenpart, and V. Shmatikov (2017). Machine Learning Models that Remember too Much. 10.48550/arXiv.1709.07886.
6. X. Yuan, K. Chen, J. Zhang, W. Zhang, N. Yu, and Y. Zhang (2023). Pseudo Label-Guided Model Inversion Attack via Conditional Generative Adversarial Network. 10.48550/arXiv.2302.09814.
7. R. Shokri, M. Stronati, C. Song, V. Shmatikov (2017). Membership Inference Attacks Against Machine Learning Models. 10.1109/SP.2017.41.
8. N. Carlini, M. Jagielskim I. Mironov (2020). Cryptanalytic Extraction of Neural Network Models. 10.1007/978-3-030-56877-1_7.
9. H. Hu, Z. Salcic, L. Sun, G. Dobbie, P. S, X. Zhang (2021). Membership Inference Attacks on Machine Learning: A Survey. 10.48550/arXiv.2103.07853.
10. H. Fang, Y. Qiu, H. Yu, W. Yu, J. Kong, B. Chong, B. Chen, X. Wang, S.T. Xia (2024). Privacy leackage on DNNs: a survey of model inversion attacks and defences. arXiv:2402.04013v2.
11. M. Nasr, et al (2023). Scalable Extraction of Training Data from (Production) Language Model. arXiv:2311.17035v1.
12. W. Fu et al (2023). Practical Membership Inference Attacks against Fine-tuned Large Language Models via Self-prompt Calibration. arXiv:2311.06062.
13. J. Morris et al 2023. Text Embeddings Reveal (Almost) as Much as Text. 12448-12460. 10.18653/v1/2023.emnlp-main.765.
14. Generative AI Models. Opportunities and Risks for Industry and Authorities. V.1.1. Federal Office for Information Security, Germany, Bonn, 2024.
15. I. Zlobaite (2010). Learning under Concept Drift: an Overview. CoRR. abs/1010.4784.
16. W. Qi (2022). A Survey on Poisoning Attacks Against Supervised Machine Learning. 10.48550/arXiv.2202.02510.
17. M.A. Ramirez, S.-K. Kim, H.Al Hamadi, E. Damiani, Y.-J. Byon, T.Y. Kim, C.S. Cho, C.Y. Yeun (2022). Poisoning Attacks and Defenses on Artificial Intelligence: A Survey. 10.48550/arXiv.2202.10276.
18. T. Gu, K. Liu, B. Dolan-Gavitt, S. Garg (2019). BadNets: Evaluating Backdooring Attacks on Deep Neural Networks. IEEE Access. 7. 47230-47244. 10.1109/ACCESS.2019.2909068.
19. A. Chakraborty, M. Alam, V. Dey, A. Chattopadhyay, D. Mukhopahyay (2018). Adversarial attacks and defences: a survey. arXiv:1810.00069v1.
20. K. Greshake, S. Abdelnabi, S. Mishra, C. Endres, T. Holz, M. Fritz (2023). More than you've asked for: A Comprehensive Analysis of Novel Prompt Injection Threats to Application-Integrated Large Language Models. 10.48550/arXiv.2302.12173.
21. H. Li, D. Namiot, A Survey of Adversarial Attacks and Defenses for image data

on Deep Learning, International Journal of Open Information Technologies ISSN: 2307-8162 vol. 10, no. 5, 2022