ИНФОРМАЦИОННАЯ БЕЗОПАСНОСТЬ В СИСТЕМАХ С ИСКУССТВЕННЫМ ИНТЕЛЛЕКТОМ

Гумерова Л.Д[1], Ефимова Ю.А [1,] Файзуллин Р.В[1]

[1]МИРЭА – Российский технологический университет, Москва, Россия

В статье анализируются актуальные вопросы в области безопасности информационных технологий (ИТ) и описываются проблемы, которые могут возникнуть в результате использования алгоритмов или искусственного интеллекта в бизнес-приложениях и их применения злоумышленниками. В статье указаны на возможные критические и фундаментальные трудности развития цифровых технологиях с точки зрения информационной безопасности.

Введение

Одну из ключевых ролей в развитии цифровой экономики играет информационная безопасность искусственного интеллекта [2,3]. Существует ряд причин необходимости обеспечения безопасности систем:

1. Защита конфиденциальной информации. Искусственный интеллект на используются во многих сферах, использующийся для обработки большого количества информации, в том числе и конфиденциальной. Свободный и\или не защищённый доступ к такой информации может привести к серьёзным проблемам – мошенничество, кража личных данных и т.д.
2. Обеспечение бесперебойных бизнес – процессов. Автоматизированные системы – это системы, работающие в режиме «реального времени», и должны работать без перебоев во избежание поломок оборудования, аварий и т.д.
3. Повышение эффективности работы систем. Искусственный интеллект может значительно повысить эффективность процессов, сохраняя безопасность данных.
4. Рост числа угроз и нарушений. Вместе с развитием технологий развивается и число угроз, поэтому важность в обеспечении безопасности так же возрастает.

Обеспечении информационной безопасности систем искусственного интеллекта – одна из основных задач цифровой экономики. Реализация безопасности систем позволяет улучшать новые технологии, повышать эффективность бизнес-процессов и защищать конфиденциальность данных.

В более узком смысле термин «искусственный интеллект» (ИИ) относится к попытке воспроизвести человекоподобные структуры принятия решений в условиях неопределённости. Компьютеры запрограммированы на самостоятельную обработку, решение проблем и принятие решений.

Методы основаны на машинном обучении. Это относится к машинной генерации знаний на основе опыта. Система учится на примерах (обучающих данных) и может обобщать их после завершения этапа обучения, то есть применять их к другим новым совокупностям (пользовательским данным). Система «распознает» паттерны и закономерности в обучающих данных и применяет их для обработки пользовательских данных.

Обучение происходит за счет демонстрации результата, но не за счет указания критериев (закономерностей). Система сама извлекает их, что и составляет обучение и последующий «интеллект». Одна из возможностей реализации — имитация нейронных сетей. Здесь из множества входных параметров формируются некоторые выходные сигналы. Это делается путем определения весов входных сигналов (с помощью параметров) и соединения всех входных и всех выходных сигналов с помощью простых правил расчета.

При обучении определяются значения всех параметров. В слишком сложных случаях их число может превышать 100 миллионов. Это означает, что мы имеем дело с огромной системой уравнений, которую практически невозможно представить полностью. Но это также означает, что мы не можем знать, по каким критериям, закономерностям, законам система определила результат. Даже то, какие входные параметры действительно важны. Вы видите только результат, то есть взаимодействие с ИИ можно сравнить с использованием черного ящика.

Иногда можно заметить в ретроспективе, что что-то пошло не так. Например, компания Amazon экспериментировала с искусственным интеллектом в процессе найма для отбора кандидатов. Оказалось, что женщины систематически оказывались в невыгодном положении. По сути, система Amazon научила себя тому, что кандидаты мужского пола предпочтительнее. Он отказывал резюме, в которых содержалось слово «женский» [6].

В принципе, искусственный интеллект также имеет дело с решениями, которые являются правильными в среднем, но могут быть совершенно неверными в отдельных случаях. Если люди во многих случаях «понимают», то есть распознают взаимосвязи между причиной и следствием и основывают на них свои решения, то с искусственным интеллектом дело обстоит иначе. Система фильтрует статистические корреляции, которые не следует путать с причинно-следственными связями. По перечисленным причинам можно усомниться в «интеллектуальности» таких систем.

Теоретический анализ

Основные методы известны уже давно. Однако сегодня эти технологии переживают бум, поскольку 1) вычислительные технологии и технологии хранения данных значительно развились с точки зрения их производительности и доступности и 2) огромные объемы данных доступны во многих областях с «Большими данными» для обучения искусственного интеллекта, в том числе с концепцией «Интернет-вещей»[7]

Все известные меры и методы обеспечения безопасности ИТ основаны на определении целевого состояния, контроле фактического состояния и целенаправленном корректирующем вмешательстве. Но есть и другие предпосылки или важные детали к уже упомянутым. Три из них описаны ниже.

1. Известные информационные потоки (по всей системе): необходимо знать ИТ-инфраструктуру и поддерживаемые ею сценарии применения (желаемые информационные потоки и т. д.). Затем аналитик по информационной безопасности может приступить к поиску путей атаки и уязвимостей и оценить их с точки зрения вероятности их появления. На основе этого анализа могут быть выбраны для внедрения решения по безопасности, которые закрывают соответствующую уязвимость и тем самым снижают или полностью устраняют риски. Если ИИ сам определяет информационный поток, то трудно отличить намеренный или функциональный информационный поток от подозрительного или враждебного информационного потока. Также трудно предотвратить, чтобы определенные информационные потоки привели к нарушениям безопасности. Итак, если говорить коротко, трудно определить целевое состояние в отношении информационных потоков.

2. Понимание функциональности ИТ (объекта ИТ): Необходимость знать целевое состояния, особенно их программных компонентов, в том числе сами приложения. Здесь, однако, речь идет о функции (т. е. о потоках информации к объекту ИТ и от него, а также внутри него). Программное обеспечение, принадлежащее системе, отличается от враждебного программного обеспечения своей функциональностью. Часто по происхождению программного обеспечения можно сделать вывод о том, является ли оно подозрительным или враждебным или нет. Если это нелегко сделать, решение между хорошим и плохим принимается на основе опыта или анализа функциональности программного обеспечения. Дальнейшие уязвимости устраняются путем контроля и ограничения доступа с помощью централизованно управляемых цифровых идентификаторов. То же самое достигается путем шифрования информации или использования мер по обеспечению целостности, таких как подписи. Уязвимости также устраняются путем фильтрации и замены данных и команд. Контроль доступа, шифрование, обеспечение целостности и (манипуляционные) фильтры могут быть использованы только в том случае, если предполагаемая функциональность ИТ (объекта ИТ) более или менее полностью понятна. Искусственный интеллект характеризуется именно тем, что его функциональность не является полностью прозрачной или простой для понимания. В результате некоторые из сегодняшних стандартных решений в области ИТ-безопасности потеряют эффективность или будут ограничены в возможностях применения.

3. Ситуация противостояния или дуэли: нападающий и защищающийся стоят прямо напротив друг друга (дуэль). Атакующий пытается напрямую преодолеть или найти бреши в мерах безопасности, которые создал и поддерживает защитник. Цель – нарушить конфиденциальность, целостность и/или доступность ценностей или подготовить дальнейшие шаги атаки. В случае систем с искусственным интеллектом атакующий и обороняющийся не обязательно сталкиваются друг с другом напрямую. Они также не обязательно взаимодействуют. Злоумышленник может, например, влиять или манипулировать пользователями или данными пользователей, которые они генерируют. Однако они используются системой безопасности защитника на основе ИИ, например, для обучения. Манипулируя ими, злоумышленник может замаскировать свою последующую атаку.

В целом, ИИ работает правильно только в том случае, если обучающие данные и данные пользователя, используемые в работе, качественно совпадают, т. е. с точки зрения их распределения. Кроме того, обучающие данные и пользовательские данные должны обладать признаком, который необходимо отфильтровать. В результате возникают предпосылки для успешного использования, описанные ниже, которые в то же время указывают на проблемы или возможности для атаки:

1. Стабильность: это означает, что ситуация не должна быстро меняться. В этом случае систему необходимо заново обучить, используя новые данные.

2. Целостность обучающих данных. Обучающие данные также должны обладать целостностью в том смысле, что ими не манипулировал злоумышленник. Манипуляция может происходить, например, путем влияния на пользователей, а также за счет того, что злоумышленник уже активен на этапе обучения, так что впоследствии он не будет распознан.

3. Целостность процесса обучения. Также должно быть невозможно, чтобы злоумышленник намеренно манипулировал процессом обучения, вводя неблагоприятные примеры. Следует помнить, что «восприятие» систем ИИ сильно отличается от человеческого. Изменение нескольких пикселей на изображении может превратить лицо в автомобиль для ИИ, в то время как человек сразу же распознает разницу.

4. Маркировка. Часто не хватает информации, необходимой для обучения. В сложных ИТ-системах это трудно определить и параметризировать. Есть много вопросов, на которые необходимо ответить: как определяются примеры «хорошего случая»? Какие возможности существуют для создания этих «хороших примеров» в смысле максимизации результатов с течением времени и предоставления их для изучения? И какие рамочные условия, жесткие и мягкие, должны быть предоставлены такой системе ИИ?

5. Для успешного обучения необходим определенный объем данных, который включает «хорошие» и «плохие» случаи. Система не сможет научится распознавать ошибки, если их никогда не встречала.

Проблемы ИТ-безопасности в связи с ИИ

Представим себе сложную систему ИИ в ИТ, которая может обучаться и принимать решения автономно в рамках заданных параметров. На разных уровнях модели Open System Interconnection (OSI) использование ИТ-ресурсов меняется с течением времени [4]. Взаимодействие элементов ИТ друг с другом адаптируется соответствующим образом. Понятная причина такого изменения, а также доказательство целостности учебных данных и процесса обучения не являются обязательными.

Распознать передовую постоянную угрозу в изменениях можно будет только в том случае, если такая измененная модель поведения уже существовала в сопоставимой среде ИИ в прошлом, и это было успешно проанализировано и задокументировано. Поэтому обнаружение с помощью искусственного интеллекта обычно не работает, вопреки многим маркетинговым заявлениям.

Активные системы безопасности в ИТ, управляемых ИИ, такие как брандмауэры и управление идентификацией и доступом, могут быстро стать помехой для работы ИИ и, соответственно, ИТ. Администраторы, управляющие системами безопасности, могут оказаться не в состоянии достаточно быстро предвидеть изменение потребностей и адаптировать правила. Поэтому необходимы механизмы, которые помогут контролировать доступ пользователей к системе. Они могут быть реализованы в виде системы учетных записей, двухфакторной аутентификации или биометрической идентификации.

Классические системы управление информацией и событиями в области информационной безопасности (Security information and event management (SIEM)) также быстро достигают своих пределов в сложной системе искусственного интеллекта [5]. Важно определять аномалии, в том числе производя непрерывный мониторинг за поведением пользователей ИТ. Поэтому SIEM-решения все чаще могут не оправдывать ожиданий. Уже сейчас их эффективность зависит от того, что они должны быть адаптированы к сценариям применения и потокам данных, которые были изменены человеком, например, при установке новых приложений. Однако если сценарии приложений и потоки данных будут все чаще контролироваться автоматически и без вмешательства человека с помощью алгоритмов или искусственного интеллекта, адаптация вышеупомянутых решений для обеспечения ИТ-безопасности будет терять свои позиции.

Наиболее практичным подходом будет ограничение ИТ-безопасности защитой интерфейсов системы ИИ с ее окружением — ценой мониторинга любой активности внутри системы ИИ. Так неужели придется считать зону, контролируемую ИИ, «черным ящиком»? Не очень приятная перспектива, потому что это не может быть подходом к защите сложных систем искусственного интеллекта.

Рассмотрим конкретный пример роста важности информационной безопасности в ИТ, например, планирование и контроль производства в рамках цифровизации экономики. Следующий этап развития цифровой экономики – это революция (Индустрия 4.0), обещающая единичное производство в рамках серийного производства.[1] В автомобильной промышленности, например, один и тот же автомобиль уже редко выпускается дважды за производственный год. Количество аксессуаров, которые можно выбрать, количество их вариантов и комбинаций слишком велико. За каждым аксессуаром стоят поставщики с соответствующей логистикой. Все это должно поступать на сборочный конвейер в определенной последовательности в соответствии с жестким графиком, иначе автомобиль не может быть произведен.

Система искусственного интеллекта с таким широким полем деятельности представляет собой весьма критический объект для атаки. Атака не может «только» привести к простою производства или бракованной продукции. Это также необходимо для обеспечения безопасности сотрудников и пользователей. Защита только интерфейсов системы искусственного интеллекта с внешним миром не является решением проблемы. Сама система искусственного интеллекта должна быть защищена. Это включает, помимо прочего, мониторинг данных об обучении и процесса обучения. Сфера действия системы ИИ не должна быть пространством, свободным от безопасности. Если некоторые из традиционных компонентов безопасности окажутся бесполезными, следует подумать о расширении других и распространении их на всю систему ИИ. Это означает модернизацию зон, в которых в настоящее время мало или вообще нет компонентов безопасности. Например, физические производственные и транспортные подразделения могут быть модернизированы для повышения их устойчивости к возможным злоупотреблениям.

Основное утверждение – кибератака является вопросом не «если», а «когда» – относится ко всем компонентам системы ИИ, а также к самому ИИ. Поэтому необходимо подумать о том, как мы можем минимизировать или даже устранить последствия атаки. В идеале, конечно, это должно быть сделано без обнуления системы.

Зачастую, в результате кибератаки, необходимо восстанавливать систему полностью. Резервное копирование и восстановление системы является пассивным способом защиты и не должен заменять активный мониторинг.

Повышение безопасности ИТ с помощью искусственного интеллекта

Искусственный интеллект или «умные алгоритмы» уже используются во многих местах для повышения безопасности ИТ. Одним из примеров является обнаружение вредоносного кода (anti-malware). Однако, по крайней мере на начальном этапе, эти решения не обнаружили. Ярким примером является вредоносная программа используемая киберпреступниками для получения выкупа — WannaCry [8]. Хорошая реализация может, по крайней мере, обнаружить большинство вредоносных программ автоматизированным и безопасным способом. Тем не менее, придется бороться с ложными срабатываниями («ошибочными срабатываниями»).

При наличии соответствующих данных искусственный интеллект или «умные алгоритмы» также могут быть использованы для улучшения управления доступом. Это предполагает динамическое добавление информации к правам доступа, которая исключает, например, «нелогичный» доступ. Это подводит нас вплотную к четвертой области применения. Она не затрагивает непосредственно ИТ-безопасность, но связана с ней. Искусственный интеллект или «умные алгоритмы» можно вполне использовать для защиты от мошенничества и выявления «оскорбительного» поведения, например, в социальных сетях. Однако при всех этих применениях необходимо помнить, что мы предполагаем, что у нас есть учебные данные, которые соответствуют нашему целевому состоянию ИТ-безопасности.

Использование алгоритмов или искусственного интеллекта в цифровой экономике создает новые проблемы. Даже установленные процедуры выбора и внедрения решений по ИТ-безопасности работают лишь в ограниченной степени. Наша нынешняя ИТ-безопасность основана на том, что информационные потоки известны и относительно стабильны. По-прежнему необходимо уметь понимать ИТ-функциональность всех ИТ-объектов, поскольку многие решения по безопасности изменяют или ограничивают ее.

Более того, в сегодняшней ИТ-безопасности обычно происходит так, что атакующий сталкивается с защитниками в том смысле, что он пытается преодолеть защиту. В целом, анализ показал, что

вышеупомянутые предпосылки больше не обязательно применимы при использовании искусственного интеллекта. Искусственный интеллект сам управляет информационными потоками, причем не предсказуемым образом; принципы его работы непрозрачны, и злоумышленники, возможно, косвенно компрометируют ИТ-безопасность.

При более внимательном рассмотрении основ использования искусственного интеллекта становятся очевидными конкретные риски или возможности для атаки, которых в рамках цифровой экономики великое множество. Они рассматривались как в виде фундаментальных проблем, так и на основе конкретных примеров. Это привело к не очень обнадеживающему предложению ограничить ИТ-безопасность защитой интерфейсов системы ИИ с ее окружением или уменьшить воздействие за счет усиления других периферийных мер. Ни защита периметра, ни смягчение последствий (типичные для управления инцидентами) не являются особенно изобретательными творениями или даже новинками.

Поэтому можно подумать о том, стоит ли оснащать решения SIEM алгоритмами и искусственным интеллектом в ответ. В конце концов, в ограниченной степени это происходит уже сегодня, например, когда их функционирование основано на анализе Больших Данных. Однако трудно представить, что «интеллект» решения безопасности настолько превосходит «интеллект» бизнес-контроля, что он может «понимать» или «контролировать» действия других.

То же самое можно сказать, если злоумышленники начнут использовать «искусственный интеллект» для осуществления своих атак. Это может позволить им маскировать и скрывать их таким образом, что они не могут быть обнаружены самыми современными решениями SIEM. Все стратегии современной защиты кибербезопасности, которые в основном основаны на SIEM (и анализе угроз), потеряют свое превосходство.

Возникает вопрос будем ли мы тогда использовать искусственный интеллект для создания фальшивых документов и корпоративных активов, чтобы смешивать их с нашими настоящими в надежде защитить последние от доступа атакующего искусственного интеллекта в рамках сегодняшней цифровизации экономики. Тогда атакующему «интеллекту» придется бороться с двумя «интеллектами»: с тем, который он хочет обнаружить, и с тем, который скрывает ценности в терминах «безопасности через неизвестность».

Результаты

Таким образом, в статье определены основные методы и подходы к обеспечению безопасности в системах с искусственным интеллектом:

1. Защита от атак внутри системы. Для защиты от атак на систему необходимо укреплять меры безопасности на всех уровнях системы, начиная с аппаратного (криптографических модулей, компьютеров и т.д.) и заканчивая уровнем пользователей (безопасность паролей и т.д.). Децентрализация системы позволяет снизить риски и повысить ее устойчивость к атакам.
2. Защита данных. Важно обеспечить защиту данных, используемых в системе. Для этого следует использовать криптографические протоколы, алгоритмы шифрования, фильтры и т.д. Кроме того, можно использовать механизмы контроля целостности данных, а также проверять корректность источника данных.
3. Мониторинг и обнаружение аномалий. Одним из важных аспектов информационной безопасности является мониторинг системы и поиск аномальных состояний, которые могут свидетельствовать о наличии угрозы. Для этого применяются методы машинного обучения, например, обнаружение выбросов, кластеризация, анализ временных рядов и т.д.
4. Система аутентификации и авторизации. Эти механизмы помогают контролировать доступ пользователей к системе. Они могут быть реализованы в виде системы учетных записей, двухфакторной аутентификации или биометрической идентификации.
5. Резервное копирование и восстановление системы. Данный подход обеспечивает возможность быстрого восстановления системы после атаки или сбоя. Важно понимать, что этот подход является пассивным и не должен заменять активный мониторинг и защиту системы.

Кроме того, для обеспечения информационной безопасности в системах с искусственным интеллектом необходимо учитывать социальные и этические аспекты, связанные с использованием подобных систем. Это может быть особенно важно в отраслях, где высокая степень автоматизации может повлиять на жизнь и здоровье людей.

Заключение

В данной работе указаны причины необходимости обеспечения безопасности искусственного интеллекта, критические и фундаментальные трудности, а также определены основные методы и подходы к обеспечению безопасности в системах с искусственным интеллектом. Выявлены и проанализированы предпосылки использования ИИ в информационной безопасности и возможные угрозы в рамках цифровой экономики.

Список литературы

1. Афанасьев А. А., Проворова И. П., Файзуллин Р. В. СПРОС ПРОМЫШЛЕННОГО ПРОИЗВОДСТВА НА ЦИФРОВЫЕ ТЕХНОЛОГИИ: ГЛОБАЛЬНЫЕ ТРЕНДЫ И РОССИЙСКАЯ РЕАЛЬНОСТЬ //Московский экономический журнал. – 2022. – №. 10. – С. 447-467.
2. Афанасьева Д. В. Применение искусственного интеллекта в обеспечении безопасности данных //Известия Тульского государственного университета. Технические науки. – 2020. – №. 2. – С. 151-154.
3. Скрыпников А. В. и др. Решение задач информационной безопасности с использованием искусственного интеллекта //Современные наукоемкие технологии. – 2021. – №. 6-2. – С. 277-281.
4. Уровни модели OSI // URL: https://www.securitylab.ru/analytics/533599.php (дата обращения: 05.03.2023).

5. SIEM (Security information and event management) // URL: https://encyclopedia.kaspersky.ru/glossary/siem/ (дата обращения: 10.03.2023).
6. Amazon scraps secret AI recruiting tool that showed bias against women // URL: https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G (дата обращения: 11.03.2023).
7. Файзуллин Р. В., Херинг Ш. Тенденции внедрения концепции "интернет вещей" для автоматизации производства // Социально-экономическое управление: теория и практика. – 2018. – №. 4. – С. 154-157.
8. Что такое программа-вымогатель WannaCry // URL: https://www.kaspersky.ru/resource-center/threats/ransomware-wannacry (дата обращения: 06.04.2023).

Гумерова Л.Д.,
Ефимова Ю.А..,
Файзуллин Р.В.
МИРЭА – Российский технологический университет, Москва, Россия

Ключевые слова:
Информационная безопасность, цифровые технологии, цифровая экономика, искусственный интеллект, системы.

Information security in systems with artificial intelligence

Keywords:
Information security, digital technology, digital economy, artificial intelligence, systems.

DOI:

JEL

Abstract:
The paper analyzes current issues in the field of information technology (IT) security and describes the problems that may arise from the use of algorithms or artificial intelligence in business applications and their use by intruders. The article points to possible critical and fundamental difficulties in the development of digital technology in terms of information security.