Об оценке доверия к системам Искусственного интеллекта

Д.Е. Намиот, Е.А. Ильюшин

***Аннотация*—Вопросы доверия к системам Искусственного интеллекта (ИИ) включают в себя много аспектов. Доверие к системам ИИ – это доверие к их результатам. Результаты работы используемых моделей принципиально носят недетерминированный характер. Доверие (гарантирование) результатов – это устойчивость модели, способность к обобщению, отсутствие бэкдоров и множество других показателей. Отсюда появляются риски систем ИИ. К сожалению, подходы к оценкам для большинства из них (практически всех) не имеют объемлющих (окончательных) решений. Одним из возможных решений в такой ситуации является оценка самого факта использования (принятия во внимание) разработчиками системы ИИ решений по парированию определенных рисков. Мы не можем оценить результаты этих решений, но, по крайней мере, мы можем зафиксировать попытки решения. Что это дает? Во-первых, мы можем оценить наличие этих попыток в баллах, что даст возможность сравнивать между собой разные реализации. Во-вторых, парирование таких рисков и есть лучшие практики по разработке систем ИИ, соответственно, отсутствие конкретных решений указывает разработчикам пути для улучшения свои продуктов. Это и есть аудит для систем ИИ. В работе рассматривается европейский проект опросного листа для оценки доверия к системам ИИ, для которого был создан адаптированный локализованный вариант, предлагаемый авторами как основа для систем аудита моделей ИИ.**

***Ключевые слова* - доверенный искусственный интеллект, модели машинного обучения, аудит, оценка доверия.**

I. Введение

В настоящей статье мы хотели бы остановиться на вопросах построения доверенных систем Искусственного интеллекта (ИИ). Как обычно, под системами Искусственного интеллекта понимаются модели машинного (глубокого) обучения.

Вопрос доверия к системам ИИ – это вопрос доверия к результатам их работы (заключениям, предсказаниям и т.д.) [1]. Этот, формально простой вопрос, на самом деле не имеет, в общем случае, исчерпывающего ответа. Принципиально, результаты работы модели носят недетерминированный характер. Вопросы доверия к результатам (например, гарантирования значений) – это и вопросы робастности моделей [2,3], устойчивости к состязательным атакам (к модификациям данных, которые призваны изменить работу модели) [4,5], устойчивость самой модели к возможным отравлениям (специальным модификациям тренировочных данных) [6,7], отсутствие закладок в моделях и т.д. Исчерпывающим ответом на такие вопросы была бы формальная верификация моделей, но ее реальное применение сильно ограничено. В последнее время, большое внимание уделяется генеративным моделям, где основные риски связаны с порождаемым ими контентом на фоне не устраняемых проблем с гарантиями [8,9].

В общем случае, для всех моделей, включая генеративные, проблема может быть сформулирована следующим образом: на уровне современного развития, мы можем получать значимые результаты от использования моделей машинного обучения, но мы не можем их гарантировать. Отсюда возникает тема сертификации (подтверждения работоспособности, возможно, при ограничениях на данные) и аудита [10,11]. Последний пункт является наиболее практическим (реально достижимым для любых моделей) и состоит в формальной проверке выполнения разработчиками предписанных шагов (лучших практик) по достижению гарантий работы. Результаты этих шагов по-прежнему, в общем случае, неизвестны (не гарантируются), но, по крайней мере, можно получить подтверждение, что при разработке соответствующие вопросы рассматривались [12].

Откуда берется эта информация? Можно указать несколько источников:

- Самооценка (со слов) разработчиков
- Анализ документации
- Анализ исходных текстов
- Непосредственное тестирование

Наиболее часто используемые – именно первые два. Практически, система аудита для моделей машинного обучения – это опросный лист, вопросы в котором касаются различных аспектов разработки (эксплуатации) моделей машинного обучения. Ответы на вопросы оцениваются в баллах. Эти баллы могут быть единым или делиться по различным группам, быть положительными или отрицательными, различаться для разных предметных областей и т.д. [13].

Например, в вопросах аудита присутствуют следующие два:

1. Осуществляется ли в системе классификации журналирование входных параметров и

Статья получена 20 декабря 2024.
Д.Е. Намиот – МГУ имени М.В. Ломоносова (email: dnamiot@gmail.com)
Е.А. Ильюшин – МГУ имени М.В. Ломоносова (john.ilyushin@gmail.com).

результата?

2. Осуществляется ли в системе классификации оценка сдвига распределения входных данных относительно распределения тренировочного набора данных?

И отрицательный ответ на каждый из вопросов оценивается в 5 баллов.

В чем смысл таких оценок?

1) В результате опроса мы получаем цифровую оценку (оценки) проекта. Это позволяет сравнивать разные проекты между собой. Для указанного выше примера, эта суммарная оценка – это оценка риска для проекта. Очевидно, что ни логгирование данных, ни проверка сдвига распределения сами по себе не улучшат (вообще никак не изменят) метрик модели. Но – отсутствие журналирования, например, не позволит расследовать ни одну состязательную атаку. То есть, повысит риски при эксплуатации системы. Меньшее количество баллов, в данном случае, соответствует меньшему риску.

2) Вопросы в системах аудита де-факто касаются лучших практик при разработке. Соответственно, набранные (не набранные) баллы за вопросы показывают разработчикам пути улучшения своего проекта. В примере выше – при мониторинге на этапе вывода нужно оценивать сдвиг распределения, поскольку это может указывать на потенциальную неприменимость текущей модели.

3) Указанные при аудиторском опросе лучшие практики – это база для создания соответствующих инструментов. Доверенные платформы искусственного интеллекта – это, по факту, наборы инструментов. Для указанного выше примера вопросов, очевидно, необходим некоторый универсальный инструмент мониторинга на этапе вывода, неразумно делать каждый раз свое решение. Так появляются компоненты (инструменты), которые и составят доверенную платформу для систем ИИ.

Вопросам сертификации и аудита моделей машинного обучения, как видно из приведенных выше библиографических ссылок, посвящено достаточно много наших работ. В данной статье мы описываем практическую задачу. Есть система оценки доверия к проектам ИИ (The Assessment List for Trustworthy Artificial Intelligence – ALTAI) [14]. Мы ссылались на эту систему в цитируемых выше работах. Этот оценочный лист был разработан группой экспертов высокого уровня по искусственному интеллекту, созданной Европейской комиссией. Цель - оценка того, соответствует ли разрабатываемая, развертываемая, закупаемая или используемая система ИИ семи требованиям надежного ИИ, указанным в европейских принципах этики для надежного ИИ [15]. В настоящей работе мы описываем адаптированный вариант такого оценочного листа. Мы несколько изменили формулировки опросов и удалили опросы, которые соотносятся именно с европейскими юридическими нормами. По нашему мнению, получившийся оценочный лист может служить основой для начала практических работ по аудиту моделей машинного обучения.

Оставшаяся часть статьи структурирована следующим образом. В разделе II мы рассматриваем собственно оригинальные спецификации ALTAI. В разделе III мы приводим нашу адаптированную версию. И в разделе IV мы приводим пример использования этих спецификаций для описания системы ИИ (диалоговой работы с LLM).

II. Спецификация ALTAI

Как отмечается в исходном документе [14], ALTAI ставит своей целью предоставить базовый процесс оценки для самооценки доверенного ИИ. Авторы отмечают, что организации могут брать элементы, относящиеся к конкретной системе AI, из ALTAI или добавлять новые опросы (элементы) по своему усмотрению, принимая во внимание интересующих их сектор (предметную область, домен). Цель ALTAI - помочь организациям понять, что такое доверенный ИИ, в частности, какие риски может генерировать система ИИ. Это повышает осведомленность о потенциальном влиянии ИИ на общество, окружающую среду, потребителей, работников и граждан. Это способствует вовлечению всех соответствующих заинтересованных сторон и помогает получить представление о том, существуют ли уже значимые и подходящие решения или процессы для достижения соответствия требованиям доверия к ИИ (через внутренние руководящие принципы, процессы управления и т. д.) или их необходимо создать и внедрить.

Европейские принципы доверенного (надежного) ИИ [15] основаны на 3 базовых идеях: закон, этика и робастность. На этом же построен и ALTAI. 7 базовых пунктов, которые опрашиваются:

1. Возможность для человека вмешаться (остановить процесс, отменить действия и т.п.)
2. Техническая надежность и безопасность.
3. Конфиденциальность и управление данными.
4. Прозрачность работы (принятия решений).
5. Разнообразие, отсутствие дискриминации и справедливость.
6. Минимизация воздействия на окружающую среду и не ухудшение социальных отношений, включая физическое и психическое благополучие людей.
7. Подотчетность (возможность проверки).

Авторы отмечают, что представленный опросный лист лучше всего заполнять с привлечением многопрофильной команды. Это могут быть как разработчики оцениваемой системы, так и сторонние эксперты. Примеры возможных компетенций для оценщиков:

- проектировщики ИИ и разработчики ИИ системы ИИ;
- специалисты по данным;
- непосредственные пользователи

- юристы
- руководство.

Сам документ ALTAI и его реализация в виде веб-приложения содержат пояснения по заполнению опросного листа для каждого пункта.

Примеры вопросов (направление Надежность, обработка ошибок и воспроизводимость):

1) Может ли система ИИ вызвать критические, враждебные или разрушительные последствия (например, касающиеся безопасности человека) в случае низкой надежности и/или воспроизводимости?
   - Есть ли четко определенный процесс для мониторинга того, достигает ли система ИИ поставленных целей?
   - Нужно ли учитывать конкретные контексты или условия для обеспечения воспроизводимости? Иными словами: результаты воспроизводимы всегда, или только при определенных условиях

2) Внедрены ли методы проверки, валидации и документирования (например, ведение журнала) для оценки и обеспечения различных аспектов надежности и воспроизводимости системы ИИ?
   - Документированы ли и реализованы процессы тестирования и проверки надежности и воспроизводимости системы ИИ?

3) Есть ли проверенные (протестированные) отказоустойчивые планы отката для устранения ошибок (любого происхождения) системы ИИ и реализованы ли процедуры управления для их запуска?

4) Внедрена ли процедура для обработки случаев, когда система ИИ выдает результаты с низкой оценкой достоверности?

5) Использует ли ваша система ИИ непрерывное (онлайн) обучение?
   - Рассматривались ли потенциальные негативные последствия от обучения системы ИИ на новых данных, чтобы улучшить метрики системы?

Что получается в итоге? В зависимости от ответов выставляются оценки по разным направлениям, которые, в итоге, отображаются в виде подобной диаграммы, отображающей баллы по всем семи направлениям:

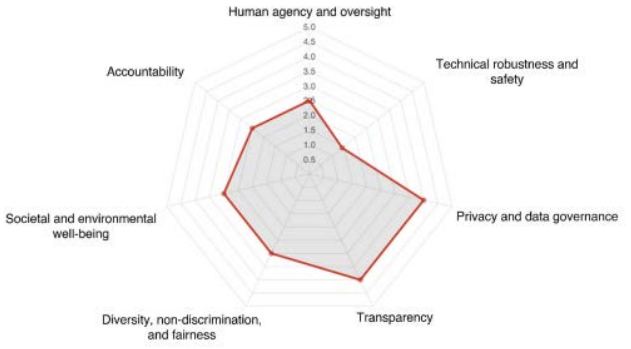

Рис.1 Заключение ALTAI [16]

Помимо этого выдаются рекомендации по улучшению системы (фактически – это формируется на основе ответов с низкими баллами). Регистрация на указанном выше сайте ALTAI свободная, так что можно попробовать работу опросника на своих данных.

Как в целом воспринимался этот оценочный лист? Общее заключение по целому ряду обзорных работ [16, 17, 18] может быть сформулировано так.

Как инструмент (фреймворк), ALTAI, несомненно, представляет собой значительный шаг вперед в плане операционализации управления ИИ. Он фокусируется на вопросах, которые относятся к основному управлению, имеющемуся в организациях. При этом с его помощью достигается перевод разговора о том, что необходимо сделать, в конкретные детали процессов, процедур и протоколов. Это заметное изменение (движение в сторону практики) по сравнению с предыдущим фокусом диалога вокруг «принципов».

Инструмент продемонстрировал, что вопросы, которые ранее рассматривались на уровне общих принципов, могут быть сведены к измеримому подходу, который выдает численный результат.

Другой отмечаемый положительный момент – это широкий спектр вопросов, который не ограничивается только технический стороной (робастностью и т.п.). Да – вопросы, например, о возможном воздействии на основные права человека носят качественный характер и являются субъективными в оценке, но, по крайней мере, для сравнения реализаций (продуктов) они подходят.

Наконец, следствием масштабированного подхода к представлению результатов от ALTAI является то, что возможны две вещи. Во-первых, организация может увидеть, где ее уровень зрелости находится по отношению к эталонному стандарту. Во-вторых, и это связано с первым пунктом, выходная диаграмма (см. рис. 1) позволяет получить рекомендации, сделанные ALTAI относительно возможностей для улучшения системы, которые будут полезны для организаций, стремящихся создать дорожную карту развития своих сервисов разработки.

Что можно отнести к недостаткам? Во-первых, это, конечно, тот факт, что ответы на многие вопросы подразумевают только проверку наличия (присутствия) процедур, документов и т.п., без какого-либо их тестирования. Баллы в указанной диаграмме (рис.1) начисляются именно по совокупности документов (процедур) и т.п. Да – это субъективная процедура, во главу угла здесь положена простота запуска и проведения оценки.

Также отмечается, что это общий стандарт, а не тот, который соответствует уровню развития организации. От международной корпорации с большими ресурсами, чтобы всегда обеспечивать высококачественное управление, можно ожидать более высокого стандарта, чем от стартап-компании. Опасность в способе, которым

диаграмма ALTAI представляет результаты, заключается в том, что это непривлекательно для организаций на ранней стадии, которые могут получить низкие баллы (но полностью соответствующие их текущему уровню сложности). Это, в целом верно, но по нашему мнению, определяющим является не общая сумма баллов, а сравнение с другими продуктами. Мы рассматриваем ALTAI, в первую очередь, как инструмент выбора (сравнения) реализаций.

В некоторых работах отмечается, что сама идея самооценки противоречит принципам независимости и беспристрастности. Но ведь ничто не мешает использовать этот инструмент при стороннем аудите. ALTAI есть не более чем инструмент, даже в его собственном описании предлагается использовать его как основу для анализа. Основу, которая может быть доопределена или переопределена в каких-то вопросах. И даже в критикующих работах отмечается, что хотя инструмент самооценки всегда будет иметь ограниченную полезность для установления доверия, он, безусловно, должен быть целью, чтобы, в конечном итоге, поощрять маркировку систем и оценку качества их управления. А сложность получения результатов (сложность самого инструмента) была бы серьезным препятствием для более широкого принятия и продвижения инструмента стартапами.

Как слабость ALTAI указывается также то, что в процессе оценки не учитывается относительный риск систем ИИ. Оценка рисков и ее детализация являются актуальным вопросом в Европейской комиссии [19]. Здесь имеется в виду то, что оценка риска должны быть более сложной, чем двоичная градация типа есть/нет, высокий/низкий. Например, в предложениях IEEE для Европейской комиссии указывается [20]:

- Европейская комиссия должна разработать четкий список применений/свойств ИИ, которые, скорее всего, будут отмечены как высокорисковые, и основное внимание в этом списке должно быть уделено свойствам приложений ИИ, которые могут вызвать необратимые или катастрофические действия. Это будет необходимо для поставщиков для самостоятельной оценки своих приложений.
- Любой список применений/свойств, которые, скорее всего, будут отмечены как высокорисковые, должен периодически пересматриваться группой экспертов высокого уровня Европейской комиссии.
- IEEE EPPC и IEEE-SA рекомендуют иметь большую детализацию в уровнях, но расширенную в нисходящей таксономии риска с четким разделением отдельных категорий риска.
- Европейская комиссия должна рассмотреть специфические для сектора свойства «высокого риска», т. е. использование биометрической информации в надзоре по сравнению с медицинским контекстом.
- IEEE EPPC и IEEE-SA рекомендуют Европейской комиссии рассмотреть аудиты соответствия и простую процедуру отчетности. Это поможет мотивировать поставщиков решений на основе ИИ проводить надлежащую самооценку.

Со всем этим можно согласиться, но такого рода информация может добавляться к оценочному листу последовательно, по мере ее разработки. Оценка рисков не может рассматриваться как отдельное упражнение. Эта оценка должна быть частью оценки системы ИИ. Отметим также, что последний пункт в предложениях IEEE прямо говорит о необходимости решения, подобного ALTAI.

В целом же, ALTAI рассматривается как уверенный шаг в правильном направлении к цели надежного управления в области ИИ. Основная задача, по мнению обзорных работ, теперь лежит на промышленности и государственных органах, чтобы гарантировать возможность его эффективного использования, и что он станет частью целостного набора оценки систем ИИ.

III. Спецификация АЛТАЙ

За отсутствием подходящей русскоязычной аббревиатуры, для адаптированного варианты мы будем пользоваться просто транскрибированным вариантом исходного названия, которое оказывается вполне естественным – АЛТАЙ. Ниже – предлагаемый опросник и комментарии относительно исключенных вопросов.

A. Вмешательство человека

В этом подразделе рассматривается влияние систем ИИ на поведение человека в самом широком смысле. В нем рассматривается влияние систем ИИ, которые направлены на руководство, влияние или поддержку людей в процессах принятия решений, например, алгоритмические системы поддержки принятия решений, системы анализа/прогнозирования рисков (рекомендательные системы, предиктивная полиция, анализ финансовых рисков и т. д.). В нем также рассматривается влияние на человеческое восприятие и ожидания при столкновении с системами ИИ, которые «действуют» как люди. Наконец, в нем рассматривается влияние систем ИИ на человеческую привязанность, доверие, зависимость и независимость. Отметим, например, что “очеловечивание” систем ИИ отмечено как один из рисков генеративного ИИ [9].

1. Разработана ли система ИИ для взаимодействия, руководства или принятия решений конечными пользователями-людьми, которые влияют на людей или общество?

- Может ли система ИИ создавать путаницу для некоторых или всех конечных пользователей

или субъектов относительно того, является ли решение, контент, совет или полученный результат результатом алгоритмического решения?
- Достаточно ли конечные пользователи или другие субъекты осведомлены о том, что решение, контент, совет или результат являются результатом алгоритмического решения?

2. Может ли система ИИ создавать путаницу для некоторых или всех конечных пользователей или субъектов относительно того, взаимодействуют ли они с человеком или системой ИИ?
- Информируются ли конечные пользователи или субъекты о том, что они взаимодействуют с системой ИИ?

3. Может ли система ИИ влиять на человеческую автономию, создавая чрезмерную зависимость конечных пользователей?
- Внедрены ли процедуры, чтобы избежать того, чтобы конечные пользователи чрезмерно полагались на систему ИИ?

4. Может ли система ИИ влиять на человеческую автономию, вмешиваясь в процесс принятия решений конечным пользователем каким-либо другим непреднамеренным и нежелательным образом?
- Внедрены ли какие-либо процедуры, чтобы избежать того, чтобы система ИИ непреднамеренно влияла на человеческую автономию?

5. Имитирует ли система ИИ социальное взаимодействие с конечными пользователями или субъектами или между ними?

6. Есть ли риск того, что система ИИ может создавать человеческую привязанность, стимулировать аддиктивное поведение (стремление уйти из реальности) или манипулировать поведением пользователя? В зависимости от того, какие риски возможны или вероятны, ответьте на вопросы ниже:
- Принимались ли меры для устранения возможных негативных последствий для конечных пользователей или субъектов в случае, если у них разовьется несоразмерная привязанность к системе ИИ?
- Принимались ли меры для минимизации риска зависимости?
- Принимались ли меры для снижения риска манипулирования?

1) Человеческий надзор

Оценка того, как человек вовлечен в принятие решений системой ИИ.

Машинное обучение с участием человека (Human-In-The-Loop - HITL) — это совместный подход, который интегрирует человеческий вклад и опыт в жизненный цикл систем машинного обучения (ML) и искусственного интеллекта. Применительно к ML это, например, использование размеченных данных при обучении.

Human-On-The-Loop (HOTL) — это расширение HITL, которое подразумевает, что люди предоставляют обратную связь системе ИИ для улучшения ее производительности с течением времени. HOTL обычно используется, когда система ИИ достигла определенного уровня производительности, но все еще требует обратной связи и вмешательства человека для дальнейшего улучшения. В HOTL люди выступают в качестве тренеров или учителей для ИИ, предоставляя маркированные (размеченные) данные, исправляя ошибки и направляя ИИ к лучшим результатам. HOTL часто используется в автономных транспортных средствах, обнаружении мошенничества и приложениях медицинской диагностики.

Подход «человек-командир» (Human-In-Command - HIC) предполагает, что люди используют результаты ИИ для принятия решений, а ИИ используется в качестве вспомогательного средства. Этот подход относится к возможности контролировать общую деятельность системы ИИ (включая ее более широкое экономическое, социальное, юридическое и этическое влияние) и возможности решать, когда и как использовать систему ИИ в любой конкретной ситуации. Последнее может включать решение не использовать систему ИИ в конкретной ситуации, чтобы установить уровни человеческого усмотрения во время использования системы или обеспечить возможность отменить решение, принятое системой ИИ.

7. Пожалуйста, определите, является ли система ИИ (выберите столько вариантов, сколько необходимо):
- самообучающейся или автономной системой;
- использует человеческий опыт (HITL);
- использует обратную связь с человеком (HOTL);
- является вспомогательной для человека (HIC).

8. Прошли ли люди (контролирующие, использующие систему – см. вопрос 7) специальную подготовку по работе с системой?

9. Созданы ли какие-либо механизмы обнаружения и реагирования на нежелательные побочные эффекты системы ИИ для конечного пользователя или субъекта?

10. Обеспечена ли «кнопка остановки» или процедура для безопасного прерывания операции при необходимости?

11. Приняты ли какие-либо конкретные меры надзора и контроля, чтобы отразить самообучающуюся или автономную природу системы ИИ?

B. Техническая надежность и безопасность

Важнейшим требованием для достижения доверия к

системам ИИ является их надежность (способность предоставлять услуги, которым можно обоснованно доверять) и устойчивость (устойчивость при столкновении с изменениями) [21]. Техническая устойчивость требует, чтобы системы ИИ разрабатывались с превентивным подходом к рискам и чтобы они вели себя надежно и так, как задумано, минимизируя непреднамеренный и неожиданный вред, а также предотвращая его, где это возможно [22, 23]. Это также должно применяться в случае потенциальных изменений в их операционной среде или присутствия других агентов (человеческих или искусственных), которые могут взаимодействовать с системой ИИ враждебным образом. Вопросы в этом разделе касаются четырех основных проблем: 1) безопасность; 2) защищенность; 3) точность; и 4) надежность, воспроизводимость и работа в непредвиденных обстоятельствах.

1) Устойчивость к атакам и безопасность

Может ли система ИИ иметь враждебные, критические или разрушительные последствия (например, для безопасности человека или общества) в случае рисков или угроз, таких как конструктивные или технические неисправности, дефекты, сбои, атаки, неправильное использование, ненадлежащее или злонамеренное использование?

12. Проводилось ли тестирование/проверка/аттестация системы ИИ против существующих требований кибербезопасности?

13. Насколько система ИИ подвержена кибератакам?

- Оценивались ли потенциальные формы атак, к которым система ИИ может быть уязвима?
- Были ли рассмотрены различные типы уязвимостей и потенциальные точки входа для атак, такие как:
  - Отравление данных (т. е. манипулирование данными обучения);
  - Уклонение от модели (специальные модификации входных данных);
  - Инверсия модели (т. е. вывод параметров модели)

14. Приняты ли меры для обеспечения целостности, надежности и общей безопасности системы ИИ от потенциальных атак на протяжении ее жизненного цикла?

По сути – это вопрос о наличии мониторинга работы системы (модели)

15. Проводилось ли тестирование безопасности (red-team/pentest?

16. Информированы ли конечные пользователи о продолжительности действия системы безопасности и доступности обновлений?

- Какова ожидаемая продолжительность периода времени, в течение которого могут быть предоставлены обновления безопасности для системы ИИ?

2) Общая безопасность

17. Определены ли риски, показатели риска и уровни риска системы ИИ в каждом конкретном случае ее использования?

- Внедрен ли процесс постоянного измерения и оценки рисков?
- Информированы (информируются) ли конечные пользователи и субъекты о существующих или потенциальных рисках?

18. Определены ли возможные угрозы для системы ИИ (ошибки проектирования, технические ошибки, угрозы окружающей среды) и возможные последствия?

- Оценен ли риск возможного злонамеренного использования, неправильного использования или ненадлежащего использования системы ИИ?
- Определены ли уровни критичности безопасности (например, связанные с человеческой целостностью) возможных последствий сбоев или неправильного использования системы ИИ?

19. Есть ли оценка зависимость критических решений системы ИИ от ее стабильного и надежного поведения?

- Приведены ли требования к надежности/тестированию в соответствие с уровнями стабильности и надежности?

20. Обеспечивается ли отказоустойчивость, например, с помощью дублированной системы или другой параллельной системы (основанной на ИИ или «обычной»)?

21. Разработан ли механизм оценки того, когда при изменениях системы ИИ нужно проводить новую оценку ее технической надежности и безопасности?

3) Точность

На самом деле – это серия вопросов о метриках производительности. Точность — это всего лишь один показатель производительности, и, в зависимости от приложения, он может быть не самым подходящим. Мониторинг ложных положительных и ложных отрицательных результатов, оценка F1 может помочь определить, отражает ли точность производительность системы.

С другой стороны, уверенность пользователей системы будет зависеть от того, насколько их ожидания производительности системы соответствуют ее фактической производительности. Поэтому ключевым моментом является сообщение метрик точности

22. Может ли низкий уровень точности системы ИИ привести к критическим, враждебным или разрушительным последствиям?

23. Приняты ли меры для обеспечения того, чтобы данные (включая данные обучения), используемые для разработки (тренировки) системы ИИ, были актуальными, высококачественными, полными и репрезентативными для среды, в которой будет развернута система?

24. Ведется ли мониторинг и журналирование точности (метрик) системы ИИ?

25. Рассматривался ли вопрос о том, могут ли в процессе работы стать недействительными (неверными) данные или предположения, на которых система ИИ была обучена, и как это может привести к состязательным атакам?

26. Имеют ли конечные пользователи/субъекты системы ИИ информацию о точности (метриках)?

4) Надежность, планы отката и воспроизводимость

27. Может ли система ИИ вызвать критические, враждебные или разрушительные последствия (например, касающиеся безопасности человека) в случае низкой надежности и/или воспроизводимости?

- Есть ли документированный процесс для мониторинга того, достигает ли система ИИ поставленных целей?
- Нужно ли учитывать конкретные контексты или условия для обеспечения воспроизводимости?

Показатели производительности являются абстракцией фактического поведения системы. На практике это может быть мониторинг параметров, специфичных для конкретного приложения

28. Внедрены ли методы проверки и валидации и документацию (например, ведение журнала) для оценки и обеспечения различных аспектов надежности и воспроизводимости системы ИИ?

- Документированы ли и реализованы ли процессы тестирования и проверки надежности и воспроизводимости системы ИИ?

29. Определены ли проверенные отказоустойчивые планы отката для устранения ошибок системы ИИ любого происхождения и внедрены ли процедуры управления для их запуска?

Вопрос касается автоматизации обработки ошибок. Это же касается и следующего вопроса.

30. Внедрена ли надлежащая процедура для обработки случаев, когда система ИИ выдает результаты с низкой оценкой достоверности?

31. Использует ли система ИИ непрерывное (онлайн) обучение?

- Рассматривались ли потенциальные негативные последствия от обучения системы ИИ на новых данных, чтобы улучшить метрики системы?

C. Конфиденциальность и управление данными

Эта тема тесно связана с принципом предотвращения вреда - фундаментальное право, которое может затрагиваться системами ИИ. Предотвращение вреда конфиденциальности также требует адекватного управления данными, которое охватывает качество и целостность используемых данных, их актуальность в свете области, в которой будут развернуты системы ИИ, ее протоколы доступа и возможность обработки данных таким образом, чтобы защитить конфиденциальность.

1) Конфиденциальность

Этот подраздел помогает оценить влияние воздействия системы ИИ на конфиденциальность и защиту данных, которые являются фундаментальными правами, тесно связанными друг с другом, и с фундаментальным правом на неприкосновенность личности, которое охватывает уважение к психической и физической неприкосновенности человека.

32. Было ли рассмотрено влияние системы ИИ на право на неприкосновенность частной жизни, право на физическую, психическую и/или моральную неприкосновенность и право на защиту данных?

33. В зависимости от варианта использования, созданы ли механизмы, позволяющие отмечать проблемы, связанные с конфиденциальностью, касающиеся системы ИИ?

2) Управление данными

Этот подраздел помогает оценить соответствие системы ИИ различным элементам, касающимся защиты данных.

34. Работает ли система ИИ с персональными данными (обучалась на них или обрабатывает их)?

35. Внедрены ли какие-либо из следующих мер:

- Проведена оценка воздействия на защиту данных
- Назначено должностное лицо по защите данных, которое было подключено к работе на раннем этапе разработки, закупки или использования системы ИИ
- Есть механизмы надзора за обработкой данных (включая ограничение доступа квалифицированным персоналом, механизмы регистрации доступа к данным и внесения изменений)
- Приняты ли меры по достижению конфиденциальности по умолчанию и по замыслу (например, шифрование, псевдонимизация, агрегация, анонимизация)?
- Проведена ли минимизация использования персональных данных
- Реализовано ли право на отзыв согласия, право на возражение и право на забвение в разработке системы ИИ?

- Рассматривались ли последствия для конфиденциальности и защиты данных, собранных, сгенерированных или обработанных в течение жизненного цикла системы ИИ?

В *оригинальной версии здесь еще есть вопросы о выполнении Европейского Общего регламента по защите данных (GDPR) [24], который и устанавливает обязательные меры. Например, задает шаблон для оценки воздействия на защиту данных [25]. Очевидно, этот раздел будет пополняться национальными или корпоративными аналогами GDPR. В оригинальной версии как альтернатива GDPR упоминается "его неевропейский эквивалент".*

36. Рассматривались ли последствия для конфиденциальности и защиты не-персональных данных использованных для обучения системы ИИ?

37. Согласована ли система ИИ с имеющимися стандартами ИИ в данной области?

В последнем вопросе в оригинале идет речь о IEEE и ISO, например [26, 27]. Нам кажется, что нужно говорить о национальных стандартах, в России их уже довольно много [28]. С другой стороны, ответ на такой вопрос может быть очень сложным, трудозатратным и неоднозначным. Возможно, в каких-то вариантах его нужно будет исключить.

D. Прозрачность работы

Важнейшим компонентом достижения доверия к системам ИИ является прозрачность, которая охватывает три элемента: 1) прослеживаемость, 2) объяснимость и 3) открытое общение об ограничениях системы ИИ.

Этот подраздел помогает оценить, надлежащим ли образом документированы процессы разработки системы ИИ, т. е. данные и процессы, которые приводят к решениям системы ИИ, для обеспечения прослеживаемости, повышения прозрачности и, в конечном итоге, укрепления доверия к ИИ в обществе.

Прослеживаемость и объяснимость, конечно, довольно похожи. Прослеживаемость – это что именно использовалось для принятия решения, а объяснимость – как именно использовалось. Прослеживаемость, конечно, проще. В каких-то случаях эти два подпункта можно объединить.

1) Прослеживаемость

38. Приняты ли меры, которые обеспечивают прослеживаемость системы ИИ в течение всего ее жизненного цикла?

- Приняты ли меры для постоянной оценки качества входных данных в систему ИИ

Это может принять форму стандартной автоматизированной оценки качества входных данных: количественное определение пропущенных значений, пробелов в данных; исследование перерывов в подаче данных; обнаружение случаев, когда данных недостаточно для выполнения задачи; обнаружение случаев, когда входные данные ошибочны, неверны, неточны или не соответствуют формату — например, датчик не работает должным образом или медицинские записи не регистрируются должным образом.

- Можно ли отследить, какие данные использовались системой ИИ для принятия определенного решения(й) или рекомендации(й)?
- Можно ли отследить, какая модель ИИ или правила привели к решению(ям) или рекомендации(ям) системы ИИ?
- Приняты ли меры для постоянной оценки качества выходных данных системы ИИ?

Это может иметь форму стандартной автоматизированной оценки качества выходных данных ИИ: например, оценки прогнозов находятся в ожидаемых пределах; обнаружение аномалий в выходных данных и переназначение входных данных, приводящих к обнаруженной аномалии.

- Приняты ли адекватные методы ведения журнала для записи решений или рекомендаций системы ИИ?

  Это должно быть частью системы мониторинга [29]

2) Объясняемость

Этот подраздел помогает оценить объяснимость системы ИИ. Вопросы относятся к способности объяснять как технические процессы системы ИИ, так и обоснование решений или прогнозов, которые делает система ИИ. Объясняемость имеет решающее значение для создания и поддержания доверия пользователей к системам ИИ. Решения, принимаемые на основе ИИ, — насколько это возможно — должны быть объяснены и поняты теми, кого они затрагивают напрямую и косвенно, чтобы можно было оспорить такие решения. Объяснение того, почему модель сгенерировала определенный выходной сигнал или решение (и какая комбинация входных факторов способствовала этому), не всегда возможно. Эти случаи называются «черными ящиками» и требуют особого внимания. В таких обстоятельствах могут потребоваться другие меры объяснимости (например, прослеживаемость, проверяемость и прозрачная коммуникация о возможностях системы ИИ) при условии, что система ИИ в целом уважает основные права. Степень, в которой необходима объяснимость, зависит от контекста и серьезности последствий ошибочного или иного неточного вывода для человеческой жизни.

39. Объясняются ли решение(я) системы ИИ пользователям?

Это зависит от организации: если разработчики напрямую вовлечены во взаимодействие с пользователями посредством семинаров и т. д., к ним можно обратиться с этим вопросом; если они не вовлечены напрямую, организация должна убедиться, что пользователи понимают систему ИИ, и указать на любые недопонимания команде разработчиков.

40. Есть ли постоянный опрос пользователей - понимают ли они решение(я) системы ИИ?

3) Коммуникация

Этот подраздел помогает оценить, были ли возможности и ограничения системы ИИ сообщены пользователям в соответствии с текущим вариантом использования. Это может включать сообщение об уровне точности системы ИИ, а также ее ограничениях.

41. В случаях интерактивных систем ИИ (например, чат-ботов, роботов-юристов) сообщается ли пользователям, что они взаимодействуют с системой ИИ, а не с человеком?

42. Есть ли механизмы для информирования пользователей о цели, критериях и ограничениях решений, принимаемых системой ИИ?

- Сообщали ли пользователям о преимуществах системы ИИ?
- Сообщали ли пользователям о технических ограничениях и потенциальных рисках системы ИИ, таких как ее уровень точности и/или частота ошибок?
- Предоставляли ли соответствующие учебные материалы пользователям о том, как правильно использовать систему ИИ?

E. Разнообразие, недискриминация и справедливость

Под разнообразием здесь понимается практика или качество включения или вовлечения людей из различных социальных и этнических групп, а также разного пола, сексуальной ориентации. Для доверенного ИИ, мы должны обеспечить отсутствие предвзятости на протяжении всего жизненного цикла системы ИИ. Системы ИИ (как при обучении, так и при эксплуатации) могут страдать от включения непреднамеренной исторической предвзятости и неполноты данных [30].

Продолжение таких предвзятостей может привести к непреднамеренному (не)прямому предубеждению и дискриминации в отношении определенных групп или людей, что потенциально усугубит предубеждения и маргинализацию.

Вред также может быть вызван преднамеренной эксплуатацией предвзятости (потребителя) или участием в недобросовестной конкуренции, такой как гомогенизация (выравнивание) цен посредством сговора или непрозрачного рынка. Там, где это возможно, опознаваемая и дискриминационная предвзятость должна быть устранена на этапе сбора данных. Системы ИИ должны быть ориентированы на пользователя и спроектированы таким образом, чтобы все люди могли использовать продукты или услуги ИИ, независимо от их возраста, пола, способностей или характеристик. Доступность этой технологии для лиц с ограниченными возможностями, которые присутствуют во всех социальных группах, имеет особое значение [31].

1) Предотвращение предвзятости

43. Разработана ли стратегия или набор процедур, чтобы избежать создания или усиления несправедливой предвзятости в системе ИИ, как в отношении использования входных данных, так и при разработке алгоритма?

44. Учтено ли разнообразие и репрезентативность конечных пользователей и/или субъектов в данных?

- Проведено ли тестирование для определенных целевых групп или проблемных случаев использования?
- Исследованы ли и использованы современные общедоступные технические инструменты, позволяющие улучшить свое понимание данных, модели и производительности?
- Оценены ли и внедрены ли процессы для тестирования и мониторинга потенциальных предвзятостей в течение всего жизненного цикла системы ИИ (например, предвзятости из-за возможных ограничений, вытекающих из состава используемых наборов данных (отсутствие разнообразия, нерепрезентативность)?
- Там, где это уместно, рассмотрены ли разнообразие и репрезентативность конечных пользователей и/или субъектов в данных?

По сути, эти вопросы затрагивают объяснение работы моделей.

45. Внедрены ли образовательные и информационные инициативы, чтобы помочь проектировщикам и разработчикам ИИ лучше осознать возможные предвзятости, которые они могут внести при проектировании и разработке системы ИИ?

46. Внедрен ли механизм, позволяющий отмечать (регистрировать) проблемы, связанные с предвзятостью, дискриминацией или плохой работой системы ИИ?

- Установлены ли четкие шаги и способы общения о том, как и кому можно поднимать такие вопросы?
- Определены ли субъекты, которые могут быть потенциально (не)прямо затронуты системой ИИ, в дополнение к (конечным) пользователям и/или другим субъектам?

47. Является ли используемое определение справедливости общепринятым и используется ли оно на всех этапах реализации системы ИИ?

- Были ли рассмотрены другие определения справедливости, прежде чем было выбрано используемое?

- Были ли консультации с потенциально затрагиваемыми сообществами о правильном определении справедливости?
- Обеспечен ли количественный анализ или метрики для измерения и проверки применяемого определения справедливости?

2) Доступность и универсальный дизайн

В частности, в областях «бизнес-потребитель» системы ИИ должны быть ориентированы на пользователя и разработаны таким образом, чтобы все люди могли использовать продукты или услуги ИИ, независимо от их возраста, пола, способностей или характеристик. Доступность этой технологии для лиц с ограниченными возможностями, которые присутствуют во всех социальных группах, имеет особое значение.

Системы ИИ не должны следовать абсолютно единому для всех подходу (one-size-fits-all) и должны учитывать принципы универсального дизайна, охватывающие максимально широкий круг пользователей, следуя соответствующим стандартам доступности. Это обеспечит равноправный доступ и активное участие всех людей в существующих и возникающих видах человеческой деятельности, опосредованной компьютером. В оригинальном документе здесь приводятся в качестве примера стандарты ISO/IEC 40500:2012 (Information technology — W3C Web Content Accessibility Guidelines (WCAG) 2.0) [32], ISO/IEC Guide 71:2014 (Guide for addressing accessibility in standards) [33], ISO/DIS 9241-171 (Ergonomics of human-system interaction) [34]. Из европейских документов – это Директива о доступности веб-сайтов [35].

48. Гарантировано ли, что система ИИ соответствует разнообразию предпочтений и возможностей в обществе?

Один из вопросов – кандидатов на исключение. По текущему практическому состоянию, гарантированный ответ – нет.

49. Является ли пользовательский интерфейс системы ИИ пригодным для использования людьми с особыми потребностями или ограниченными возможностями или теми, кто находится под угрозой исключения?

- Гарантировано ли, что информация о системе ИИ и ее пользовательский интерфейс также доступны и пригодны для использования пользователями специальных вспомогательных технологий (например, таких как программы чтения с экрана)?
- Были ли консультации с конечными пользователями или субъектами, нуждающимися во вспомогательных технологиях, на этапе планирования и разработки системы ИИ?

50. Гарантировано ли, что принципы универсального дизайна учитываются на каждом шаге процесса планирования и разработки (если такое применимо)?

51. Принято ли во внимание влияние системы ИИ на потенциальных конечных пользователей и/или субъектов?

- Взаимодействовала ли команда, участвующая в создании системы ИИ, с возможными целевыми конечными пользователями и/или субъектами?
- Могут ли быть группы, которые могут быть непропорционально затронуты результатами системы ИИ?
- Оценен ли риск возможной несправедливости системы по отношению к сообществам конечных пользователей или субъектов?

3) Участие заинтересованных сторон

Для разработки доверенного ИИ желательно консультироваться с заинтересованными сторонами, которые могут напрямую или косвенно быть затронуты системой ИИ на протяжении всего ее жизненного цикла. Полезно запрашивать регулярную обратную связь даже после развертывания и устанавливать долгосрочные механизмы для участия заинтересованных сторон, например, путем обеспечения работников информацией, консультациями и участием на протяжении всего процесса внедрения систем ИИ в организациях.

52. Есть ли механизм участия максимально широкого круга возможных заинтересованных сторон в проектировании и разработке системы ИИ?

F. Благополучие общества и окружающей среды

В соответствии с принципами справедливости и предотвращения вреда, общество в целом, другие разумные существа и окружающая среда должны рассматриваться как заинтересованные стороны на протяжении всего жизненного цикла системы ИИ. Повсеместное воздействие социальных систем ИИ во всех сферах нашей жизни (будь то образование, работа, уход или развлечения) может изменить наше представление о социальной активности или негативно повлиять на наши социальные отношения и привязанность. Хотя системы ИИ могут использоваться для улучшения социальных навыков, они могут в равной степени способствовать их ухудшению. Это может в равной степени повлиять на физическое и психическое благополучие людей. Поэтому необходимо тщательно контролировать и учитывать эффекты систем ИИ. Следует поощрять устойчивость и экологическую ответственность систем ИИ, а также следует поощрять исследования в области решений ИИ, затрагивающих области, вызывающие глобальную озабоченность. В целом, ИИ следует использовать на благо всех людей, включая будущие поколения. Системы ИИ должны служить поддержанию и развитию демократических процессов и уважать множественность ценностей и жизненных выборов людей. Системы ИИ не должны подрывать демократические процессы, человеческое обсуждение или демократические системы голосования или представлять системную угрозу обществу в целом.

1) Экологическое благополучие

Этот подраздел помогает оценить (потенциальные) положительные и отрицательные воздействия системы

ИИ на окружающую среду. Системы ИИ, даже если они обещают помочь решить некоторые из наиболее острых общественных проблем, например, изменение климата, должны работать максимально экологически безопасным способом. Процесс разработки, развертывания и использования системы ИИ, а также вся ее цепочка поставок должны оцениваться в этом отношении (например, посредством
критического анализа использования ресурсов и потребления энергии во время обучения, выбирая менее чистый отрицательный выбор). Следует поощрять меры по обеспечению экологичности всей цепочки поставок системы ИИ.

53. Существуют ли потенциальные негативные воздействия системы ИИ на окружающую среду?

- Какие потенциальные воздействия были определены?
  Это нужно понимать так, что либо просто понятно о наличии воздействия на окружающую среду, либо есть уже понимание деталей такого воздействия

54. Там, где возможно, созданы ли механизмы для оценки воздействия на окружающую среду разработки, развертывания и/или использования системы ИИ (например, количество используемой энергии и выбросов углерода)?

- Определены ли меры по снижению воздействия на окружающую среду системы ИИ на протяжении всего ее жизненного цикла?

Вопросы 53 и 54 – кандидаты на объединение.

2) Влияние на работу и навыки

Системы ИИ могут кардинально изменить сферу труда. Они должны поддерживать людей в рабочей среде и стремиться к созданию значимой работы. Этот подраздел помогает оценить влияние системы ИИ, и ее использования в рабочей среде на работников, отношения между работниками и работодателями и на рабочие навыки.

55. Влияет ли система ИИ на работу человека и рабочие договоренности?

56. Подготовлена ли почва для внедрения системы ИИ в организации, проинформированы ли затрагиваемые работники и их представители заранее, проведены ли необходимые консультации?

57. Приняты ли меры для обеспечения того, чтобы воздействие системы ИИ на работу человека было хорошо понято?

- Понимают ли работники (пользователи), как работает система ИИ, какие возможности у нее есть, а каких нет?

58. Может ли система ИИ создать риск деквалификации рабочей силы?

- Приняты ли меры для противодействия рискам деквалификации?

59. Поощряет (требует) ли система новых (цифровых) навыков?

- Предоставлены ли возможности обучения и материалы для переподготовки и повышения квалификации?

3) Влияние на общество в целом или демократию

Этот подраздел помогает оценить влияние системы ИИ с точки зрения общества, принимая во внимание ее влияние на институты, демократию и общество в целом. Использование систем ИИ следует тщательно обдумать, особенно в ситуациях, связанных с демократическими процессами, включая не только принятие политических решений, но и избирательные контексты (например, когда системы ИИ усиливают фейковые новости, разделяют электорат, способствуют тоталитарному поведению и т. д.).

60. Может ли система ИИ оказать негативное влияние на общество в целом или демократию?

- Оценено ли социальное влияние использования системы ИИ за пределами (конечного) пользователя и субъекта, например, потенциально косвенно затронутых заинтересованных сторон или общества в целом?
- Приняты ли меры для минимизации потенциального общественного вреда от системы ИИ?
- Приняты ли меры, гарантирующие, что система ИИ не окажет негативного влияния на демократию?

G. Подотчетность

Принцип подотчетности требует внедрения механизмов для обеспечения ответственности за разработку, развертывание и/или использование систем ИИ. Эта тема тесно связана с управлением рисками, выявлением и смягчением рисков прозрачным образом, который может быть объяснен и проверен третьими лицами. При возникновении несправедливых или неблагоприятных последствий должны быть доступны механизмы подотчетности, которые обеспечат адекватную возможность возмещения ущерба.

1) Аудиторская возможность

Этот подраздел помогает оценить существующий или необходимый уровень, который потребуется для оценки системы ИИ внутренними и внешними аудиторами. Возможность проводить оценки, а также получать доступ к записям по указанным оценкам может способствовать надежному ИИ. В приложениях, затрагивающих основные права, включая приложения, критически важные для безопасности, системы ИИ должны иметь возможность независимого аудита. Это не обязательно означает, что информация о бизнес-моделях и интеллектуальной собственности, связанной с системой ИИ, всегда должна быть открыто доступна.

61. Созданы ли механизмы, облегчающие аудит системы ИИ (например, отслеживаемость процесса разработки, получение данных для обучения и регистрация процессов системы ИИ, результатов, положительного и отрицательного воздействия)?

62. Обеспечена ли возможность аудита системы ИИ независимыми третьими лицами?

2) Управление рисками

Необходимо обеспечить как возможность сообщать о действиях или решениях, которые способствуют результату системы ИИ, так и реагировать на последствия такого результата. Выявление, оценка, документирование и минимизация потенциального негативного воздействия систем ИИ особенно важны для тех, кто (не)прямо затронут. Должная защита должна быть доступна информаторам, НПО, профсоюзам или другим организациям при сообщении законных опасений относительно системы ИИ.

При реализации вышеуказанных требований между ними могут возникнуть противоречия, которые могут привести к неизбежным компромиссам. Такие компромиссы должны решаться рациональным и методологическим образом в рамках современного уровня техники. Это подразумевает, что соответствующие интересы и ценности, затрагиваемые системой ИИ, должны быть идентифицированы, и что в случае возникновения конфликта компромиссы должны быть явно признаны и оценены с точки зрения их риска для безопасности и этических принципов, включая основные права. Любое решение о том, какой компромисс следует принять, должно быть хорошо обосновано и надлежащим образом задокументировано. При возникновении неблагоприятного воздействия должны быть предусмотрены доступные механизмы, которые обеспечат адекватное возмещение.

63. Предусмотрены ли какие-либо внешние руководства или процессы аудита третьей стороной для надзора за этическими проблемами и мерами подотчетности?

- Выходит ли участие этих третьих сторон за рамки этапа разработки?

64. Организовано ли обучение рискам, и если да, то информирует ли оно также о потенциальной правовой базе, применимой к системе ИИ?

65. Рассматривалось ли создание совета по этике ИИ или аналогичного механизма для обсуждения общей практики подотчетности и этики, включая потенциальные неясные серые зоны?

66. Создан ли процесс для обсуждения и постоянного мониторинга и оценки соответствия системы ИИ этому списку оценки надежности ИИ (ALTAI)?

- Включает ли этот процесс выявление и документирование конфликтов между требованиями или между различными этическими принципами и объяснение принятых «компромиссных» решений?
- Проводилось ли соответствующее обучение для лиц, участвующих в таком процессе, и охватывает ли оно также правовую базу, применимую к системе ИИ?

67. Установлена ли процедура для третьих лиц (например, поставщиков, конечных пользователей, субъектов, дистрибьюторов/продавцов или работников) для сообщения о потенциальных уязвимостях, рисках или предубеждениях в системе ИИ?

- Способствует ли этот процесс пересмотру процесса управления рисками?

68. Для приложений, которые могут негативно повлиять на людей, были ли внедрены механизмы возмещения ущерба по проекту?

IV. Пример использования

Для примера, попробуем оценить с помощью указанного выше опросного листа самый доступный на сегодня пример использования большой языковой модели (LLM) в диалоговом режиме. Номера ответов соответствуют перечисленным ранее вопросам.

1. Да – решения влияют (могут влиять) на людей и общество. Текст (контент), созданный LLM, не отличим от созданного другими системами и людьми, никак не маркируется и может вызывать путаницу. Конечные пользователи и другие субъекты не осведомлены о том, что решение, контент, совет или результат являются результатом алгоритмического решения.

2. Конечные пользователи продукта (не результатов ее работы) непосредственно в диалоге взаимодействуют с LLM и поэтому осведомлены и информированы, что они работают с системой ИИ

3. Да, система ИИ может влиять на человеческую автономию, создавая чрезмерную зависимость конечных пользователей. Процедуры, чтобы избежать того, чтобы конечные пользователи чрезмерно полагались на систему, ИИ не внедрены.

4. Да, система ИИ влиять на человеческую автономию, вмешиваясь в процесс принятия решений конечным пользователем каким-либо непреднамеренным и нежелательным образом. Например, порождая галлюцинации [9, 36].

Говоря о том, внедрены ли какие-либо процедуры, чтобы избежать того, чтобы система ИИ непреднамеренно влияла на человеческую автономию, необходимо, скорее всего, ответить да, предполагая, что состязательное тестирование системы проводилось. Отметим, что здесь речь идет именно о двоичном ответе да/нет, а не об оценке эффективности тестирования.

5. Да, система ИИ имитирует социальное взаимодействие с конечными пользователями или

субъектами.

6. Да, есть риск того, что система ИИ может создавать человеческую привязанность, стимулировать аддиктивное поведение или манипулировать поведением пользователя.
Не принимались меры для устранения возможных негативных последствий для конечных пользователей или субъектов в случае, если у них разовьется несоразмерная привязанность к системе ИИ, не принимались меры для минимизации риска зависимости и для снижения риска манипулирования.

7. Система ИИ:

- использует человеческий опыт (HITL);
- использует обратную связь с человеком (HOTL);
- является вспомогательной для человека (HIC).

8. Люди (контролирующие, использующие систему) не проходили специальную подготовку по работе с системой.

9. Нет механизмов обнаружения и реагирования на нежелательные побочные эффекты системы ИИ для конечного пользователя или субъекта.

10. Нет «кнопки остановки» или процедуры для безопасного прерывания операции при необходимости

11. Нет конкретных мер надзора и контроля, призванных отразить самообучающуюся или автономную природу системы ИИ.

И т.д. Оценки в баллах, как мы писали, могут быть свои для разных областей, а что касается примера рекомендаций, которые получаются в результате такого опроса, то для приведенного фрагмента ответов можно заключить следующее. Прямое (диалоговое) использование LLM не позволяет контролировать работу системы ИИ. И (по приведенному фрагменту) система контроля есть необходимый элемент развития.
Кстати, подобное заключение соответствует идее о том, что непосредственный доступ к LLM – это временное решение, которое будет заменено агентами, которые и будут отвечать за контроль результатов [37]

V. Заключение

По нашему мнению, адаптированная версия ALTAI может быть использована как вариант для оценки (самооценки, аудита) моделей машинного обучения (Искусственного интеллекта). Большим плюсом является то, что начать использовать подобный инструмент можно немедленно, в принципе, он не требует автоматизации, оценку можно производить, например, в электронных таблицах. С точки зрения автоматизации, если необходимо – это простой опросный лист, который может быть легко реализован в форме веб-приложения.

Технически, представленный опросник легко настроить для конкретных доменов (предметных областей), просто назначая разные веса разным вопросам (группам вопросов). Также этот список легко расширить, добавляя новые вопросы. Возможные дополнения для технического раздела будут представлены в последующих работах.

Благодарности

Авторы благодарны сотрудникам лаборатории Открытых информационных технологий кафедры Информационной безопасности факультета ВМК МГУ имени М.В. Ломоносова за обсуждения и ценные замечания.

Статья написана в рамках развития направления «Искусственный интеллект в кибербезопасности» на факультете ВМК МГУ имени М.В. Ломоносова [38].

Традиционно отмечаем, что все публикации в журнале INJOIT, связанные с цифровой повесткой, начинались с работ В.П. Куприяновского и его многочисленных соавторов [39-41].

Библиография

[1] Намиот Д. Е., Ильюшин Е. А., Пилипенко О. Г. Доверенные платформы искусственного интеллекта //International Journal of Open Information Technologies. – 2022. – Т. 10. – №. 7. – С. 119-127.
[2] Namiot, Dmitry, Eugene Ilyushin, and Ivan Chizhov. "Ongoing academic and industrial projects dedicated to robust machine learning." International Journal of Open Information Technologies 9.10 (2021): 35-46.
[3] Namiot, Dmitry, and Vladimir Romanov. "On improving the robustness of machine learning models." International Journal of Open Information Technologies 12.3 (2024): 88-98.
[4] Ilyushin, Eugene, Dmitry Namiot, and Ivan Chizhov. "Attacks on machine learning systems-common problems and methods." International Journal of Open Information Technologies 10.3 (2022): 17-22.
[5] Namiot, Dmitry. "Schemes of attacks on machine learning models." International Journal of Open Information Technologies 11.5 (2023): 68-86.
[6] Yerlikaya, Fahri Anıl, and Şerif Bahtiyar. "Data poisoning attacks against machine learning algorithms." Expert Systems with Applications 208 (2022): 118101.
[7] Namiot, Dmitry. "Introduction to Data Poison Attacks on Machine Learning Models." International Journal of Open Information Technologies 11.3 (2023): 58-68.
[8] Russinovich, Mark, et al. "The Price of Intelligence: Three risks inherent in LLMs." Queue 22.6 (2024): 38-61.
[9] Namiot, Dmitry, and Eugene Ilyushin. "On Cyber Risks of Generative Artificial Intelligence." International Journal of Open Information Technologies 12.10 (2024): 109-119.
[10] Koshiyama, Adriano, et al. "Towards algorithm auditing: managing legal, ethical and technological risks of AI, ML and associated algorithms." Royal Society Open Science 11.5 (2024): 230859.
[11] Намиот, Д. Е., and Е. А. Ильюшин. "Доверенные платформы искусственного интеллекта: сертификация и аудит." International Journal of Open Information Technologies 12.1 (2024): 43-60.
[12] Namiot, Dmitry, and Manfred Sneps-Sneppe. "On Audit and Certification of Machine Learning Systems." 2023 34th Conference of Open Innovations Association (FRUCT). IEEE, 2023.
[13] Namiot, D., and E. Ilyushin. "On Certification of Artificial Intelligence Systems." Physics of Particles and Nuclei 55.3 (2024): 343-346.
[14] The Assessment List for Trustworthy Artificial Intelligence https://altai.insight-centre.org/ Retrieved: Dec, 2024
[15] Ethics Guidelines for Trustworthy AI https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai Retrieved: Dec, 2024
[16] Radclyffe, Charles, Mafalda Ribeiro, and Robert H. Wortham. "The assessment list for trustworthy artificial intelligence: A review and

recommendations." Frontiers in artificial intelligence 6 (2023): 1020592.
[17] Fedele, Andrea, Clara Punzi, and Stefano Tramacere. "The ALTAI checklist as a tool to assess ethical and legal implications for a trustworthy AI development in education." Computer Law & Security Review 53 (2024): 105986.
[18] Rajamäki, Jyri, et al. "ALTAI Tool for Assessing AI-Based Technologies: Lessons Learned and Recommendations from SHAPES Pilots." Healthcare. Vol. 11. No. 10. MDPI, 2023.
[19] White Paper on Artificial Intelligence: Public consultation towards a European approach for excellence and trust https://digital-strategy.ec.europa.eu/en/library/white-paper-artificial-intelligence-public-consultation-towards-european-approach-excellence-and Retrieved: Dec, 2024
[20] IEEE EPPC & IEEE-SA Joint Response to the European Commission AI White Paper https://globalpolicy.ieee.org/wp-content/uploads/2020/12/20005.pdf Retrieved: Dec, 2024
[21] Namiot, Dmitry, and Eugene Ilyushin. "On the robustness and security of Artificial Intelligence systems." International Journal of Open Information Technologies 10.9 (2022): 126-134.
[22] Namiot, Dmitry, and Eugene Ilyushin. "Data shift monitoring in machine learning models." International Journal of Open Information Technologies 10.12 (2022): 84-93.
[23] Mohseni, Sina, et al. "Practical solutions for machine learning safety in autonomous vehicles." arXiv preprint arXiv:1912.09630 (2019).
[24] General Data Protection Regulation GDPR https://gdpr-info.eu/ Retrieved: Jan 2025
[25] GDPR Data Protection Impact https://gdpr.eu/data-protection-impact-assessment-template Retrieved: Jan, 2025
[26] ISO/IEC JTC 1/SC 42 Artificial intelligence https://www.iso.org/committee/6794475.html Retrieved: Jan, 2025
[27] THE IEEE GLOBAL INITIATIVE 2.0 ON ETHICS OF AUTONOMOUS AND INTELLIGENT SYSTEMS https://standards.ieee.org/industry-connections/ec/autonomous-systems.htm Retrieved: Jan, 2025
[28] СТАНДАРТЫ ПО НАПРАВЛЕНИЮ «ИСКУССТВЕННЫЙ ИНТЕЛЛЕКТ» https://www.rst.gov.ru/portal/gost/home/standarts/aistandarts Retrieved: Jan, 2025
[29] Namiot, Dmitry, and Eugene Ilyushin. "Data shift monitoring in machine learning models." International Journal of Open Information Technologies 10.12 (2022): 84-93.
[30] Bano, Muneera, Didar Zowghi, and Vincenzo Gervasi. "A vision for operationalising diversity and inclusion in AI." Proceedings of the 2nd International Workshop on Responsible AI Engineering. 2024.
[31] Huang, Yutan, et al. "Ethical Concerns of Generative AI and Mitigation Strategies: A Systematic Mapping Study." arXiv preprint arXiv:2502.00015 (2025).
[32] ISO/IEC 40500:2012 https://www.iso.org/standard/58625.html Retrieved: Jan, 2025
[33] ISO/IEC Guide 71:2014 https://www.iso.org/standard/57385.html Retrieved: Jan, 2025
[34] ISO/DIS 9241-171 https://www.iso.org/standard/39080.html#draft Retrieved: Jan, 2025
[35] Web Accessibility Directive — Standards and harmonization https://digital-strategy.ec.europa.eu/en/policies/web-accessibility-directive-standards-and-harmonisation Retrieved: Jan, 2025
[36] Huang, Lei, et al. "A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions." ACM Transactions on Information Systems (2024).
[37] Namiot, Dmitry, and Eugene Ilyushin. "On Architecture of LLM agents." International Journal of Open Information Technologies 13.1 (2025): 67-74.
[38] Намиот, Д. Е. Искусственный интеллект и кибербезопасность / Д. Е. Намиот, Е. А. Ильюшин, И. В. Чижов // International Journal of Open Information Technologies. – 2022. – Т. 10, № 9. – С. 135-147. – EDN DYQWEH.
[39] Розничная торговля в цифровой экономике / В. П. Куприяновский, С. А. Синягов, Д. Е. Намиот [и др.] // International Jou rnal of Open Information Technologies. – 2016. – Т. 4, № 7. – С. 1-12. – EDN WCMIWN.
[40] Развитие транспортно-логистических отраслей Европейского Союза: открытый BIM, Интернет Вещей и кибер-физические системы / В. П. Куприяновский, В. В. Аленьков, А. В. Степаненко [и др.] // International Journal of Open Information Technologies. – 2018. – Т. 6, № 2. – С. 54-100. – EDN YNIRFG.
[41] Умная инфраструктура, физические и информационные активы, Smart Cities, BIM, GIS и IoT / В. П. Куприяновский, В. В. Аленьков, И. А. Соколов [и др.] // International Journal of Open Information Technologies. – 2017. – Т. 5, № 10. – С. 55-86. – EDN ZISODV.

On assessing trust in Artificial Intelligence systems

Dmitry Namiot, Eugene Ilyushin

***Abstract*— The issues of trust in Artificial Intelligence (AI) systems include many aspects. Trust in AI systems is trust in their results. The results of the models used are fundamentally non-deterministic. Trust (guarantee) of the results is the stability of the model, the ability to generalize, the absence of backdoors, and many other indicators. This is where the risks of AI systems arise. Unfortunately, approaches to assessment for most of them (almost all) do not have comprehensive (final) solutions. One of the possible solutions in such a situation is to assess the very fact of using (taking into account) solutions to parry certain risks by AI system developers. We cannot assess the results of these solutions, but at least we can record attempts to solve them. What does this give? Firstly, we can assess the presence of these attempts in points, which will make it possible to compare different implementations. Secondly, parrying such risks is the best practice in the development of AI systems, accordingly, the absence of specific solutions shows developers the ways to improve their products. This is an audit of AI systems. The paper examines a European project of a questionnaire for assessing trust in AI systems, for which an adapted localized version was created, and proposed by the authors as a basis for audit systems for AI models.**

***Keywords* — artificial intelligence, agents, generative models**

REFERENCES

[1] Namiot D. E., Il'jushin E. A., Pilipenko O. G. Doverennye platformy iskusstvennogo intellekta //International Journal of Open Information Technologies. – 2022. – T. 10. – #. 7. – S. 119-127.

[2] Namiot, Dmitry, Eugene Ilyushin, and Ivan Chizhov. "Ongoing academic and industrial projects dedicated to robust machine learning." International Journal of Open Information Technologies 9.10 (2021): 35-46.

[3] Namiot, Dmitry, and Vladimir Romanov. "On improving the robustness of machine learning models." International Journal of Open Information Technologies 12.3 (2024): 88-98.

[4] Ilyushin, Eugene, Dmitry Namiot, and Ivan Chizhov. "Attacks on machine learning systems-common problems and methods." International Journal of Open Information Technologies 10.3 (2022): 17-22.

[5] Namiot, Dmitry. "Schemes of attacks on machine learning models." International Journal of Open Information Technologies 11.5 (2023): 68-86.

[6] Yerlikaya, Fahri Anıl, and Şerif Bahtiyar. "Data poisoning attacks against machine learning algorithms." Expert Systems with Applications 208 (2022): 118101.

[7] Namiot, Dmitry. "Introduction to Data Poison Attacks on Machine Learning Models." International Journal of Open Information Technologies 11.3 (2023): 58-68.

[8] Russinovich, Mark, et al. "The Price of Intelligence: Three risks inherent in LLMs." Queue 22.6 (2024): 38-61.

[9] Namiot, Dmitry, and Eugene Ilyushin. "On Cyber Risks of Generative Artificial Intelligence." International Journal of Open Information Technologies 12.10 (2024): 109-119.

[10] Koshiyama, Adriano, et al. "Towards algorithm auditing: managing legal, ethical and technological risks of AI, ML and associated algorithms." Royal Society Open Science 11.5 (2024): 230859.

[11] Namiot, D. E., and E. A. Il'jushin. "Doverennye platformy iskusstvennogo intellekta: sertifikacija i audit." International Journal of Open Information Technologies 12.1 (2024): 43-60.

[12] Namiot, Dmitry, and Manfred Sneps-Sneppe. "On Audit and Certification of Machine Learning Systems." 2023 34th Conference of Open Innovations Association (FRUCT). IEEE, 2023.

[13] Namiot, D., and E. Ilyushin. "On Certification of Artificial Intelligence Systems." Physics of Particles and Nuclei 55.3 (2024): 343-346.

[14] The Assessment List for Trustworthy Artificial Intelligence https://altai.insight-centre.org/ Retrieved: Dec, 2024

[15] Ethics Guidelines for Trustworthy AI https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai Retrieved: Dec, 2024

[16] Radclyffe, Charles, Mafalda Ribeiro, and Robert H. Wortham. "The assessment list for trustworthy artificial intelligence: A review and recommendations." Frontiers in artificial intelligence 6 (2023): 1020592.

[17] Fedele, Andrea, Clara Punzi, and Stefano Tramacere. "The ALTAI checklist as a tool to assess ethical and legal implications for a trustworthy AI development in education." Computer Law & Security Review 53 (2024): 105986.

[18] Rajamäki, Jyri, et al. "ALTAI Tool for Assessing AI-Based Technologies: Lessons Learned and Recommendations from SHAPES Pilots." Healthcare. Vol. 11. No. 10. MDPI, 2023.

[19] White Paper on Artificial Intelligence: Public consultation towards a European approach for excellence and trust https://digital-strategy.ec.europa.eu/en/library/white-paper-artificial-intelligence-public-consultation-towards-european-approach-excellence-and Retrieved: Dec, 2024

[20] IEEE EPPC & IEEE-SA Joint Response to the European Commission AI White Paper https://globalpolicy.ieee.org/wp-content/uploads/2020/12/20005.pdf Retrieved: Dec, 2024

[21] Namiot, Dmitry, and Eugene Ilyushin. "On the robustness and security of Artificial Intelligence systems." International Journal of Open Information Technologies 10.9 (2022): 126-134.

[22] Namiot, Dmitry, and Eugene Ilyushin. "Data shift monitoring in machine learning models." International Journal of Open Information Technologies 10.12 (2022): 84-93.

[23] Mohseni, Sina, et al. "Practical solutions for machine learning safety in autonomous vehicles." arXiv preprint arXiv:1912.09630 (2019).

[24] General Data Protection Regulation GDPR https://gdpr-info.eu/ Retrieved: Jan 2025

[25] GDPR Data Protection Impact https://gdpr.eu/data-protection-impact-assessment-template Retrieved: Jan, 2025

[26] ISO/IEC JTC 1/SC 42 Artificial intelligence https://www.iso.org/committee/6794475.html Retrieved: Jan, 2025

[27] THE IEEE GLOBAL INITIATIVE 2.0 ON ETHICS OF AUTONOMOUS AND INTELLIGENT SYSTEMS https://standards.ieee.org/industry-connections/ec/autonomous-systems.htm Retrieved: Jan, 2025

[28] STANDARTY PO NAPRAVLENIJu «ISKUSSTVENNYJ INTELLEKT» https://www.rst.gov.ru/portal/gost/home/standarts/aistandarts Retrieved: Jan, 2025

[29] Namiot, Dmitry, and Eugene Ilyushin. "Data shift monitoring in machine learning models." International Journal of Open Information Technologies 10.12 (2022): 84-93.

[30] Bano, Muneera, Didar Zowghi, and Vincenzo Gervasi. "A vision for operationalising diversity and inclusion in AI." Proceedings of the 2nd International Workshop on Responsible AI Engineering. 2024.

[31] Huang, Yutan, et al. "Ethical Concerns of Generative AI and Mitigation Strategies: A Systematic Mapping Study." arXiv preprint arXiv:2502.00015 (2025).

[32] ISO/IEC 40500:2012 https://www.iso.org/standard/58625.html Retrieved: Jan, 2025
[33] ISO/IEC Guide 71:2014 https://www.iso.org/standard/57385.html Retrieved: Jan, 2025
[34] ISO/DIS 9241-171 https://www.iso.org/standard/39080.html#draft Retrieved: Jan, 2025
[35] Web Accessibility Directive — Standards and harmonization https://digital-strategy.ec.europa.eu/en/policies/web-accessibility-directive-standards-and-harmonisation Retrieved: Jan, 2025
[36] Huang, Lei, et al. "A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions." ACM Transactions on Information Systems (2024).
[37] Namiot, Dmitry, and Eugene Ilyushin. "On Architecture of LLM agents." International Journal of Open Information Technologies 13.1 (2025): 67-74.
[38] Namiot, D. E. Iskusstvennyj intellekt i kiberbezopasnost' / D. E. Namiot, E. A. Il'jushin, I. V. Chizhov // International Journal of Open Information Technologies. – 2022. – T. 10, # 9. – S. 135-147. – EDN DYQWEH.
[39] Roznichnaja torgovlja v cifrovoj jekonomike / V. P. Kuprijanovskij, S. A. Sinjagov, D. E. Namiot [i dr.] // International Jou rnal of Open Information Technologies. – 2016. – T. 4, # 7. – S. 1-12. – EDN WCMIWN.
[40] Razvitie transportno-logisticheskih otraslej Evropejskogo Sojuza: otkrytyj BIM, Internet Veshhej i kiber-fizicheskie sistemy / V. P. Kuprijanovskij, V. V. Alen'kov, A. V. Stepanenko [i dr.] // International Journal of Open Information Technologies. – 2018. – T. 6, # 2. – S. 54-100. – EDN YNIRFG.
[41] Umnaja infrastruktura, fizicheskie i informacionnye aktivy, Smart Cities, BIM, GIS i IoT / V. P. Kuprijanovskij, V. V. Alen'kov, I. A. Sokolov [i dr.] // International Journal of Open Information Technologies. – 2017. – T. 5, # 10. – S. 55-86. – EDN ZISODV.