Доверенные платформы искусственного интеллекта: сертификация и аудит

Д.Е. Намиот, Е.А. Ильюшин

***Аннотация*—Под системами Искусственного интеллекта в данной работе понимаются системы машинного обучения. Именно системы машинного обучения (глубокого обучения) являются, на сегодняшний день, основными примерами использования Искусственного интеллекта в самых разнообразных областях. С практической точки зрения, можно говорить о том, что машинное обучение является синонимом понятия Искусственный интеллект. При этом системы машинного обучения, по своей природе, зависят от данных, на которых они обучаются и, принципиально, выдают недетерминированные результаты. Доверенные платформы, в соответствии со своим названием, представляют собой набор инструментов, призванных повысить доверие (уверенность пользователя) к результатам работы моделей машинного обучения. Применение систем машинного обучения в так называемых критических областях (авионика, автоматическое вождение и т.п.) требует гарантий работоспособности программного обеспечения, что подтверждается процедурой (процессом) сертификации. А под аудитом понимается идентификация возможных проблем с работоспособностью и безопасностью систем машинного обучения.**

***Ключевые слова*—машинное обучение, робастность, сертификация, аудит.**

I. Введение

Сама идея доверенных платформ в компьютерных науках не нова [1]. Термин “доверенные платформы” существует достаточно давно и всегда означал наличие подлинной вычислительной среды (системы), гарантированно исключающей ее сторонние изменения. Основной смысл доверенных вычислений состоит в том, чтобы дать производителям оборудования контроль над тем, какое программное обеспечение работает (не работает) в системе, отказываясь запускать неподписанное программное обеспечение. Благодаря доверенным вычислениям компьютер будет постоянно вести себя ожидаемым образом, и это поведение будет обеспечиваться компьютерным оборудованием и программным обеспечением. Обеспечение такого поведения достигается за счет загрузки аппаратного обеспечения с уникальным ключом шифрования, который недоступен для остальной части системы и ее владельца. Эта концепция необходима и для систем машинного обучения в критических применениях, поскольку есть, например, атаки, которые ориентированы на фреймворки машинного обучения. Изменение, например, функции вычисления потерь в конкретном фреймворке будет затрагивать все модели машинного обучения на такой платформе [4]. Соответственно, требования по достоверности программного обеспечения важны и для моделей машинного обучение (“чистые” реализации фреймворков, библиотек и т.д.), Но для систем машинного обучения это лишь самая малая из проблем. В этой части инфраструктура для моделей машинного обучения ни чем не отличается от других программ, и принципы построения доверенных сред абсолютно одинаковы.

Главная причина недоверия к результатам работы систем машинного обучения проистекает из самой модели машинного обучения [1]. Основная проблема недоверия происходит именно из-за отсутствия доверия к обработке данных. Мы вырабатываем какие-то заключения (обобщения) по выделенному набору данных, а затем распространяем свои заключения на всю, чаще всего неизвестную нам, генеральную совокупность. Тренировочные данные, на которых вырабатывались обобщения, могут быть просто неверными или даже специально модифицированными (состязательные атаки отравления [2]), данные на этапе выполнения могут отличаться по своим статистическим характеристикам от тренировочных (различные сдвиги данных [3]), или также быть специальным образом изменены (состязательные атаки [4]), выбранная модель может не обладать свойством робастности [5] и т.д. Именно это есть основные причины недоверия к результатам работы систем машинного обучения. Многие из указанных проблем не могут быть устранены полностью, мы можем говорить только о смягчении возможных последствий [6]. Соответственно, доверенные платформы для машинного обучения – это коллекции инструментов, предназначенных для работы с указанными выше проблемами. При этом инструменты доверенных платформ касаются не только процесса разработки. Они должны охватывать все элементы конвейера машинного обучения. Сдвиг данных, например, необходимо определять уже в процессе использования (эксплуатации).

Доверенные платформы – это платформы,

Статья получена 9 ноября 2023.
Д.Е. Намиот – МГУ имени М.В. Ломоносова (e-mail: dnamiot@gmail.com).
Е.А. Ильюшин - МГУ имени М.В. Ломоносова (e-mail: john.ilyushin@gmail.com).

инструменты которых позволяют повысить доверие к моделям машинного обучения, платформы, которые позволяют анализировать тренировочные данных, противостоять состязательным атакам, определять сдвиги данных при работе системы и т.д. Примерами таких платформ являются Datarobot [42] или IBM Trustworthy [43].

Системы машинного обучения на этапе эксплуатации (вывода, inference), используемые в критических областях (авионика, специальные системы и т.п.), вообще говоря, должны сертифицироваться, как и любое другое программное обеспечение в этих областях. Сертификация, в частности, означает подтверждение правильной работы при всех допустимых входных данных. А аудитом, как обычно, является процесс идентификации проблемных мест (решений).

Именно этим компонентам доверенных платформ машинного обучения (искусственного интеллекта) - сертификации и аудиту посвящена настоящая работа.

Наши работы в области аудита и сертификации систем машинного обучения представлялись на конференциях GRID-2023 [7] и FRUCT-2023 [8].

Оставшаяся часть статьи структурирована следующим образом. В разделе III содержатся общие положения. Раздел III посвящен правовым регуляциям. Раздел IV посвящен аудиту и сертификации. В разделе V рассматривается сертификация для систем машинного обучения. В разделе VI кратко описывается сертификация моделей машинного обучения.

II. О безопасности систем машинного обучения

Модели машинного обучения зависят от данных. Изменение данных на этапе обучения, например, ведет к изменению параметров модели. Изменение входных данных (по отношению к данным, на которых модель обучалась) ведут к изменению результатов работы. Такие изменения могут быть весьма существенными и качественными (например, изменение классификации объектов и т.п.) или же просто вести к снижению точности работы системы. Соответственно, исходя из этого и возникают так называемые состязательные атаки на модели машинного обучения – сознательные модификации данных на разных этапах конвейера, которые призваны либо помешать работе системы машинного обучения, либо, наоборот, добиться желаемого для атакующего результата работы.

Google (Deepmind) в обзорной публикации своей исследовательской группы Robust and Verified Deep Learning group отмечает, что “системы машинного обучения по умолчанию не являются надежными. Даже системы, которые превосходят людей в определенной области, могут потерпеть неудачу в решении простых проблем, если будут внесены различия в исходные данные” [9].

Презентация Madry-lab (MIT) представила три заповеди Secure / Safe ML (рис. 1)

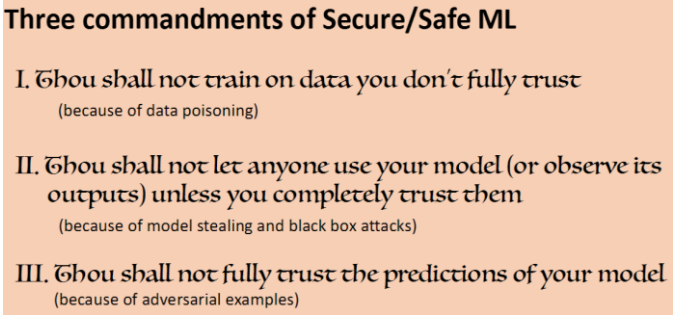

Рис. 1. Secure/Safe ML [10]

I. Вы не должны тренироваться на данных, которым не полностью доверяете (из-за возможного отравления данных – изменения данных с целью обмана модели)

II. Вы не должны позволять никому использовать вашу модель (или наблюдать за ее работой), если вы полностью им не доверяете (из-за кражи модели и атак черного ящика). Это можно представить как аналогию декомпилирования или reverse engineering в программных системах – работа (поведение) модели изучается с целью построения состязательного примера.

III. Вы не должны полностью доверять предсказаниям вашей модели (из-за возможных состязательных примеров).

Особую значимость такие состязательные примеры имеют, естественно, для критических приложений (авионика, автоматическое вождение, ядерная энергетика и т.п.). Последствия ошибок здесь всегда серьезные, и для подобного рода систем могут найтись заинтересованные лица в подобных атаках.

NIST, согласно последним рекомендациям [4], выделяет три базовых типа атак в отношении систем машинного обучения: отравления [2], уклонения [11] и атаки на интеллектуальную собственность [12]. Последние представляют собой специальный опрос моделей с целью извлечения непубличной информации и не оказывают воздействия на результаты работы (исключая сознательные искажения вывода для противодействия таким атакам). Термин отравление используется для того, чтобы подчеркнуть долговременный характер воздействия на модели и включает в себя отравление данных (специальные модификации данных на этапе тренировки) и отравление моделей (непосредственная модификация готовых моделей [13]). Для осуществления таких атак нужен доступ к тренировочным данным (или загрузка отравленных данных), либо загрузка модифицированных (отравленных) моделей. В первом приближении можно сказать, что требования по защите от таких атак похожи на обычные требования кибербезопасности (цифровой гигиены), с запретом загрузки чего либо из неизвестных источников (по крайней мере, для критических приложений это точно должно быть исключено). Остаются атаки уклонения, которые заключаются в модификации (в цифровом или физическом доменах) входных данных. В классической

форме, на момент своего появления, это и были минимальные модификации входных данных, которые вызывали неверную работу системы. Такая атака изображена на рисунке 2, когда с помощью градиента функции потерь вычисляется шум, добавление которого к исходному изображению меняет классификацию.

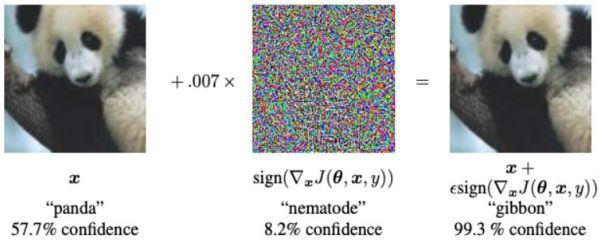

Рис. 2. Атака уклонения [14].

Проблемой является то, что для достижения указанного эффекта с изменением классификации вовсе не обязательны состязательные атаки. Обученная модель может быть такова, что небольшие изменения в данных вызывают значительные изменения в результате. И здесь мы приходим к гораздо более важному понятию – это устойчивые модели машинного обучения. Классически, для систем машинного обучения устойчивость определяется как независимость результата системы от небольших изменений входных данных. Наличие такой зависимости исключает, естественным образом, использование таких систем в приложениях, где результаты работы должны гарантироваться. Устойчивые модели машинного обучения являются востребованной темой исследований, основанием для запуска которых являлись как раз потребности использования систем машинного обучения в критических системах [15].

Формально, устойчивость определяется, примерно, в следующей форме. При заданных входных данных x и интересующей модели f мы хотим, чтобы предсказание модели оставалось одинаковым для всех входных данных x' в окрестности x, где окрестности определяются некоторой функцией расстояния δ и некоторым максимальным расстоянием Δ:

$$\forall x'.\ \delta(x, x') \leq \Delta \Rightarrow f(x) = f(x') \qquad (1)$$

Например, результаты работы классификатора не менялись при небольшом изменении данных. Фундаментальная основа исследований в области устойчивости совершенно понятна. Принципиально, любая модель обучается на некотором подмножестве данных, а затем обобщается на всю генеральную совокупность данных. Которая, в общем случае, неизвестна на момент обучения. И к машинному обучению (искусственным нейронным сетям) мы обратились именно потому, что связи внутри данных нам неизвестны. Именно их мы хотим восстановить (смоделировать), обучая нейронную сеть. Эта неопределенность заставляет предполагать, что данные во время эксплуатации могут отличаться от тех, на которых модель обучалась. Поскольку данные при эксплуатации изменились, то вполне может оказаться так, что обобщения, сделанные на этапе обучения уже неверны. Если данные меняются специальным образом, то это называют атаками на системы машинного обучения. Именно вокруг описанной выше формулы (1) и строятся все исследования в области устойчивости. Как подобрать минимально отличающиеся данные, которые, тем не менее, классифицируются по-иному? Поскольку в большинстве случаев речь идет об изображениях, то говорят именно о незаметных человеческому взгляду изменениях, формально выражаемых в одной из L-метрик, которые приводят к изменению классификации. Или, в противоположную сторону, проверить, что при заданных небольших изменениях классификация не изменится. Практически сразу, при такой постановке, возникает вопрос – а как такая постановка соотносится именно с безопасностью? Допустим, мы доказали, что в малой окрестности известных данных работа системы остается стабильной. А что происходит вне этой окрестности? Насколько вообще важна "незаметность" изменений, если в критических применениях (авионика и т.п.) мы имеем дело с автоматическими системами, там попросту нет человека в контуре принятия решения, и размах изменений, вообще говоря, ничего не решает. Все выглядит так, что малые изменения выбраны потому, что это позволяет формально описать происходящие процессы и использовать известные ранее подходы. Но это вовсе не продиктовано именно задачами безопасности. Представляется, что на самом деле, по крайней мере, для критических приложений, устойчивость трактуется (воспринимается) в иной форме. А именно – сохранение показателей работы модели, достигнутых на этапе тренировки, во время ее практического использования [5].

Тут прослеживается полная параллель с традиционным внедрением программного обеспечения. На этапе тестирования мы проверили работоспособность системы, и ожидаем, что эта работоспособность сохранится на этапе эксплуатации. Отметим, что для критических применений программное обеспечение еще и подлежит сертификации. Смысл этой сертификации как раз и состоит во всеобъемлющем тестировании (доказательстве правильности работы). Сообразно такому же принципу и воспринимается устойчивость. На этапе тренировки мы достигли определенных выбранных показателей работы (аккуратность, ROC и т.д.) и ожидаем сохранения этих же параметров при тестировании (эксплуатации) модели. Для критических применений показатели натренированной модели ниже некоторого определенного уровня просто будут останавливающим фактором при переходе к эксплуатации. То есть устойчивость становится синонимом работоспособности. Это не сохранение показателей при малых возмущениях тренировочных данных, а сохранение достигнутых на этапе тренировки

показателей уже на всей генеральной совокупности. А это уже совсем не то, что, обычно, исследуется в работах по устойчивости систем машинного обучения.

Этот же факт отмечается в работе [16]. Устойчивость - это термин, который практикующие специалисты часто используют, но он обычно обобщенно относится к правильности или достоверности прогнозов модели, а не к формальному понятию устойчивости (1), изучаемому в академической литературе.

Что не так, и в чем же тогда вообще смысл работ по устойчивости? Уверенность в правомерности таких вопросов укрепилась после прочтения работы [17], где Christian Kästner из Carnegie Mellon написал ровно о том же.

Отметим, что в формуле (1) ничего не говорится о правильности работы системы (например, о результатах классификации). То есть, вполне может существовать устойчивая система, которая выдает неверные результаты. И эти неверные результаты остаются таковыми при малых возмущениях исходных данных. Отсюда, устойчивость сама по себе не может свидетельствовать о безопасности программного обеспечения. Безопасность – это свойство системы, которая включает в себя модель машинного обучения. Применительно к системам машинного обучения, безопасность используется как синоним доверия к результатам работы [1]. Устойчивость с этой практической точки зрения неотделима от объяснения работы модели. В самом деле – по определению понятия “черный ящик” мы не можем гарантировать никакие свойства и характеристики для него, в силу того, что они неизвестны и не предсказуемы.

Общее состояние дел с дискриминантными моделями машинного обучения можно описать просто. Мы, очевидно, можем получать важные результаты с помощью машинного обучения (этим и объясняется всеобщий интерес), но, в общем случае, нельзя гарантировать эти результаты. Естественно, что, в первую очередь, это проблема именно для критических систем. Программное обеспечение, например, в авионике, сертифицируется для подтверждений гарантий работоспособности на всем диапазоне возможных входных данных. Модель машинного обучения на этапе применения (inference), в той же авионике, есть также не что иное, как некоторое программное обеспечение. И для такой программы также нужна сертификация. Отсутствие таковой для систем искусственного интеллекта приведет к делению специального программного обеспечения на сорта – сертифицированные (проверенные) программы и несертифицированные.

Отметим, что порождающие модели (если говорить о LLM, которые вызывают сейчас наибольший интерес) в этой части не отличаются от дискриминантных. Атаки отравления данных существуют и для LLM [18]. Отсутствие верификации отмечено и в списке OWASP - “Чрезмерная зависимость от контента, генерируемого LLM, без контроля со стороны человека может привести к пагубным последствиям“ [19]. Хотя в целом, основные риски для больших языковых моделей в настоящее время видятся в плане доступа к аккумулированной информации [20]. Поэтому в данной статье мы останавливаемся на дискриминантных моделях и системах классификации, которые как раз и характерны для критических применений.

III. Правовое регулирование

Обеспечение гарантий результатов работы, естественным образом, входит в различные регулирующие акты для систем ИИ (ML). Как отмечает издание MIT Technology Review в своем сборнике The Algorithm: “Suddenly, everyone wants to talk about how to regulate AI”

Более того, что необычно для отрасли, руководители крупнейших компаний высказываются за регулирование ИИ. Руководители OpenAI, Microsoft и Google публично высказываются в пользу регулирования и проводят встречи с мировыми лидерами. А национальные правительства предлагают новые ограничения для генеративного ИИ. Генеральный директор OpenAI (автор ChatGPT) Сэм Альтман отправился в мировое турне, чтобы выразить поддержку новым законам, включая предстоящий Закон Европейского Союза об искусственном интеллекте. Руководители OpenAI призвали глобальный регулирующий орган контролировать сверхразумные машины и свидетельствовали в пользу регулирования ИИ перед Конгрессом США. Компания OpenAI выделяет гранты разработку сред управления ИИ [21]. Президент Microsoft Брэд Смит повторил призывы OpenAI к агентству США по регулированию ИИ. Отдельно генеральный директор Google Сундар Пичаи согласился сотрудничать с европейскими законодателями для разработки «пакта об ИИ» — набора добровольных правил, которым разработчики должны следовать до вступления в силу правил ЕС. Несомненно, этот процесс ускорился именно из-за успехов больших языковых моделей. На ежегодном собрании в Японии в 2023 году G7, неформальный блок промышленно развитых демократических правительств, объявила о Хиросимском процессе. Это межправительственная целевая группа для исследования рисков генеративного ИИ. Члены G7, пообещали разработать взаимно совместимые законы, которые позволят регулировать ИИ в соответствии с демократическими ценностями. К ним относятся справедливость, подотчетность, прозрачность, безопасность, конфиденциальность данных, защита от злоупотреблений и соблюдение прав человека.

Президент США Джо Байден опубликовал стратегический план развития ИИ. Инициатива призывает регулирующие органы США разработать общедоступные наборы данных, контрольные показатели и стандарты для обучения, измерения и оценки систем ИИ. Регулятор конфиденциальности данных Франции объявил о структуре регулирования

генеративного ИИ. На сегодняшний день Китай уже прямо регулирует генеративный ИИ. В марте официальные лица ЕС переписали закон Союза об искусственном интеллекте, чтобы классифицировать генеративные модели искусственного интеллекта как «высокорисковые», из-за чего они подлежат бюрократическому надзору и регулярным проверкам [22].

В США в 2022 году в конгресс и сенат был внесен Закон об алгоритмической ответственности (Algorithmic Accountability Act). Законопроект потребует от компаний проведения алгоритмической оценки воздействия и рисков для устранения ощутимого вреда автоматизированных систем принятия решений, таких как те, например, которые отказывают людям в их заявках на ипотеку.

Американский закон о защите конфиденциальности данных (American Data Privacy Protection Act) – это попытка регулировать сбор и обработку данных компаниями. Дебаты вокруг рисков генеративного ИИ могут придать ему дополнительную срочность. Закон запретит компаниям, занимающимся генеративным искусственным интеллектом, собирать, обрабатывать или передавать данные дискриминационным образом. Это также даст пользователям больше контроля над тем, как компании используют их данные. Например, от компаний может потребоваться разрешить внешним экспертам проверять свои технологии до их выпуска, а также предоставлять пользователям и правительству дополнительную информацию об их системах искусственного интеллекта.

Текущие дебаты американских законодателей свидетельствуют о том, что, видимо, появится еще одно агентство для регулирования ИИ. Также возможно, по данным издания Algorithm, появление нового регулятора специально для задач ИИ [23].

Новый регулирующий орган, созданный Европейским союзом - Европейский центр алгоритмической прозрачности (ECAT) будет изучать алгоритмы, которые идентифицируют, классифицируют и ранжируют информацию на сайтах социальных сетей и поисковых системах. ECAT уполномочен определять, соответствуют ли алгоритмы (ИИ и другие) Закону о цифровых услугах Европейского Союза, который призван блокировать разжигание ненависти в Интернете. Закон должен блокировать определенный контент. Для агентства определены три основные задачи:

1. Расследование. Это оценка функционирования алгоритмов «черного ящика». Включает в себя анализ отчетов и аудиты, проведенные компаниями, которые по закону обязаны представлять отчеты регуляторам. Он установит процедуры для независимых исследователей и регуляторов для получения доступа к данным, связанным с алгоритмами.
2. Исследование. Это анализ возможностей алгоритмов рекомендаций для распространения незаконного контента, нарушения прав человека, нанесения ущерба демократии или вреда здоровью пользователей, оценка рисков и мероприятия по их снижению, повышение прозрачности алгоритмов (то есть – объяснения их работы).
3. Создание центра обмена информацией и передовым опытом между исследователями в академических кругах, промышленности и государственной службе.

И, конечно, здесь нужно упомянуть принятый Закон об искусственном интеллекте [22]. Европарламент определил, что разработанный (равно как используемый) в Европе безопасный ИИ должен полностью соответствовать правам и ценностям ЕС, включая права человека, безопасность, конфиденциальность, прозрачность, отсутствие дискриминации, а также социальное и экологическое благополучие.

Системы ИИ с неприемлемым уровнем риска для безопасности людей будут запрещены. В эту категорию попадают, например, те, которые используются для классификации людей на основе их социального поведения или каких-либо иных личных характеристик (так называемые социальной оценки).

Закон расширяет ограничительный список запретов на навязчивое и дискриминационное использование ИИ. Сюда относится, например, распознавание лиц (в тексте - удаленная биометрическая идентификация в реальном времени), а также биометрическая категоризация (то есть категоризация по полу, расе и т.п.). Также отмечаются предиктивные полицейские системы, systems распознавания эмоций в правоохранительных органах, на рабочих местах и в учебных заведениях. Закон запрещает нецелевое извлечение изображений лиц из интернета или видеозаписей с камер наблюдения для создания баз данных распознавания лиц. Для генеративного ИИ вводится запрет использования любых материалов, защищенных авторским правом, в обучающем наборе больших языковых моделей, таких как OpenAI GPT-4.

К числу приложений с высоким риском относятся системы ИИ, которые наносят значительный вред здоровью людей, безопасности, основным правам или окружающей среде, системы, используемые для влияния на избирателей и исход выборов. Также отметим, что рекомендательные системы, используемые платформами социальных сетей, классифицируются как приложения с высоким риском.

Все перечисленные акты определяют требования к законченному продукту. Они не определяют практических шагов по достижению требуемых характеристик, равно как и не определяют метрик, которыми должны измеряться эти характеристики.

Системы, относящиеся к аудиту – перечисляют требования по проверке систем ИИ (систем машинного обучения).

IV. Аудит и сертификация

Эти два понятия уже непосредственно относятся к практической области. Классически: аудит представляет собой процесс инспекции (проверки), а сертификация – это уже подтверждение (гарантия) данных (результатов работы).

Аудит систем машинного обучения – новая и достаточно быстро развивающаяся область. Причины – указанные выше проблемы с гарантированием результатов работы. Отчет Game Changers среди 9 технологий, которые изменят каждую индустрию, на первом месте называет именно AI аудит [25]. Вот из самых свежих областей: агентство Bloomberg сообщает, что Администрация киберпространства Китая объявила о проекте руководящих принципов, которые потребуют проверки безопасности сервисов генеративного ИИ, прежде чем им будет разрешено работать. В предлагаемых правилах говорится, что операторы ИИ должны обеспечивать точность контента, уважать интеллектуальную собственность, не подвергать опасности безопасность и не допускать дискриминации. Кроме того, контент, созданный ИИ, должен быть четко помечен. Этот шаг является частью растущих усилий Китая по регулированию быстрого распространения генеративного ИИ с момента дебюта ChatGPT OpenAI в прошлом году.

По факту, аудит для систем машинного обучения – это набор лучших практик о том, что и как проверять для готовых систем. Проактивно, это также должны быть практики того, как разрабатывать безопасные системы. На сегодняшний день можно сказать, что понимание разработчиками того факта, что такие практики нужны явно превалирует над пониманием того, что же конкретно нужно делать [24]. Для примера приведем первые 10 практик из этой работы:

- Оценка рисков перед развертыванием системы
- Оценки опасных возможностей
- Аудит сторонних моделей
- Тестирование безопасности (red team)
- Ограничения безопасности
- Техники верификации модели
- План реагирования на инциденты безопасности
- Пред-тренировочная оценка риска
- Системы мониторинга и их использование
- Оценки модели после развертывания

По факту – это достаточно общие пункты плана работ. По большинству из них нет каких-то объемлющих (закрывающих) решений.

Аудит для системы машинного обучения (ИИ) – это оценка своих алгоритмов, моделей, данных и процессов проектирования. Такая оценка приложений ИИ внутренними и внешними аудиторами помогает обосновать надежность системы ИИ, продемонстрировать ответственность проектировщиков и повысить обоснованность прогнозов, сделанных моделями. Аудит ИИ охватывает [26]:

- Оценку моделей, алгоритмов и потоков данных
- Анализ операций, результатов и обнаруженных аномалий
- Технические аспекты систем ИИ для оценки точности результатов
- Этические аспекты систем ИИ для справедливости, законности и конфиденциальности

Это соответствует общепринятым определениям того, что аудит — это инструмент для опроса сложных процессов, для определения того, соответствуют ли они политике компании, отраслевым стандартам или правилам [27]. Стандарт IEEE для разработки программного обеспечения определяет аудит как «независимую оценку соответствия программных продуктов и процессов применимым нормам, стандартам, руководствам, планам, спецификациям и процедурам» [28].

Рис. 3 из указанной работы иллюстрирует этапы аудита. Опять таки – это именно чеклист для проверки наличия необходимых активностей. Например, возьмем раздел ML-мониторинга. Здесь предлагается выяснить ответы на следующие вопросы:

- Есть ли в системе ИИ соответствующий процесс мониторинга для отслеживания производительности модели, отклонений и действий модели?
- Какие действия предприняты при выполнении конвейера машинного обучения для обеспечения соответствия приложений ИИ законам и нормативным стандартам, соответствия целям организации и демонстрации этической и социальной ответственности?

Вопрос мониторинга моделей машинного обучения достаточно сложен сам по себе. В зависимости от типа возможного сдвига данных [3], заключения по мониторингу будут разные. И в зависимости от характера системы решения так же могут быть разными.

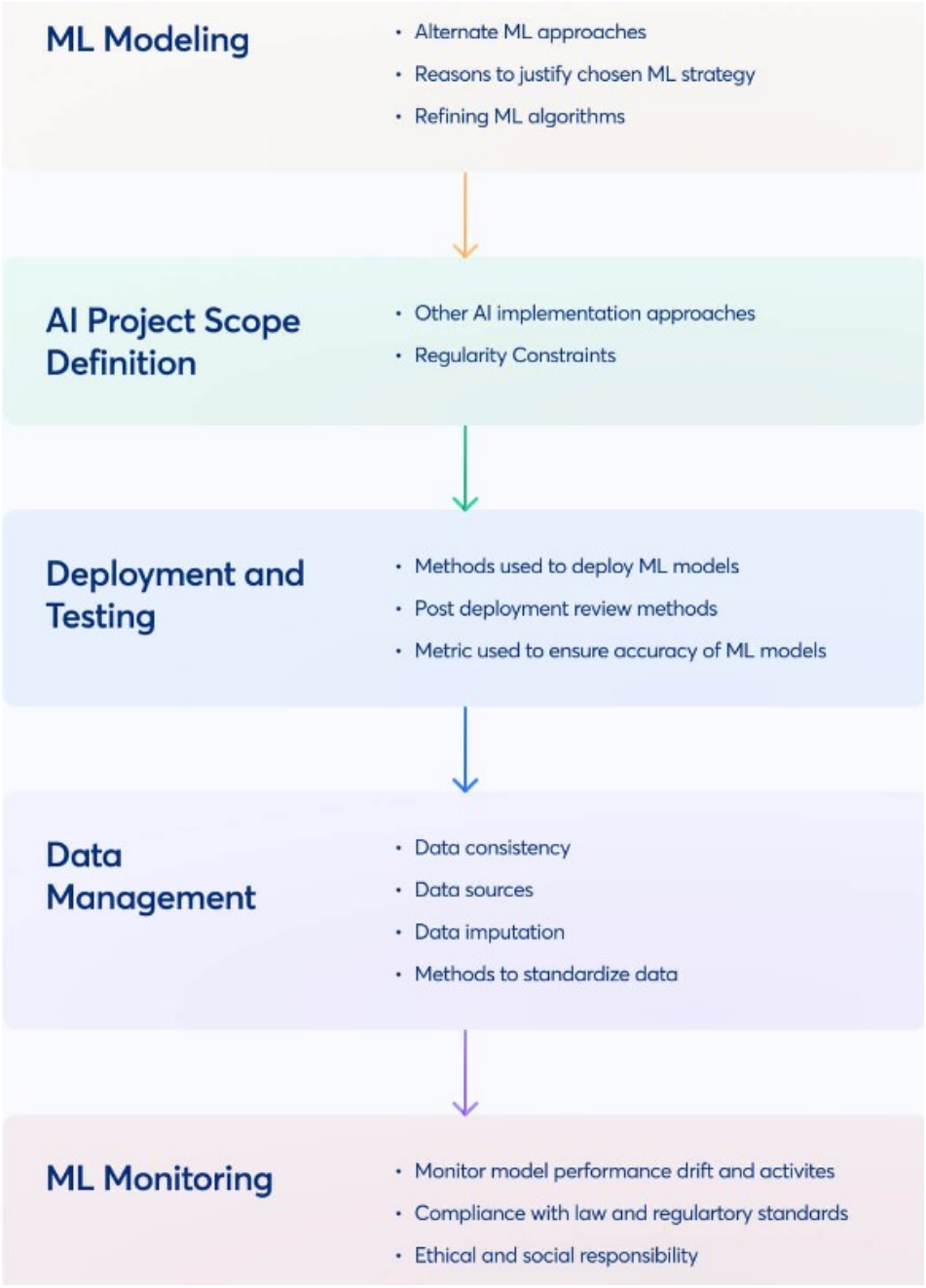

Рис.3. Элементы ИИ аудита [26]

Для приложения, которое работает 24x7, например, так называемый сдвиг концепции является катастрофой, поскольку такие приложения нельзя останавливать для переобучения. Вместе с тем именно как чеклист, который определяет обязательные шаги в разработке (эксплуатации) – это вполне работающие рекомендации. Оценка соответствия заданным позициям ведется “вручную”. Пример – работа Стенфордского университета [32] по оценке соответствия черновому варианту европейского закона об ИИ (Рис. 4). Оценки соответствия требованиям проставлялись экспертами вручную

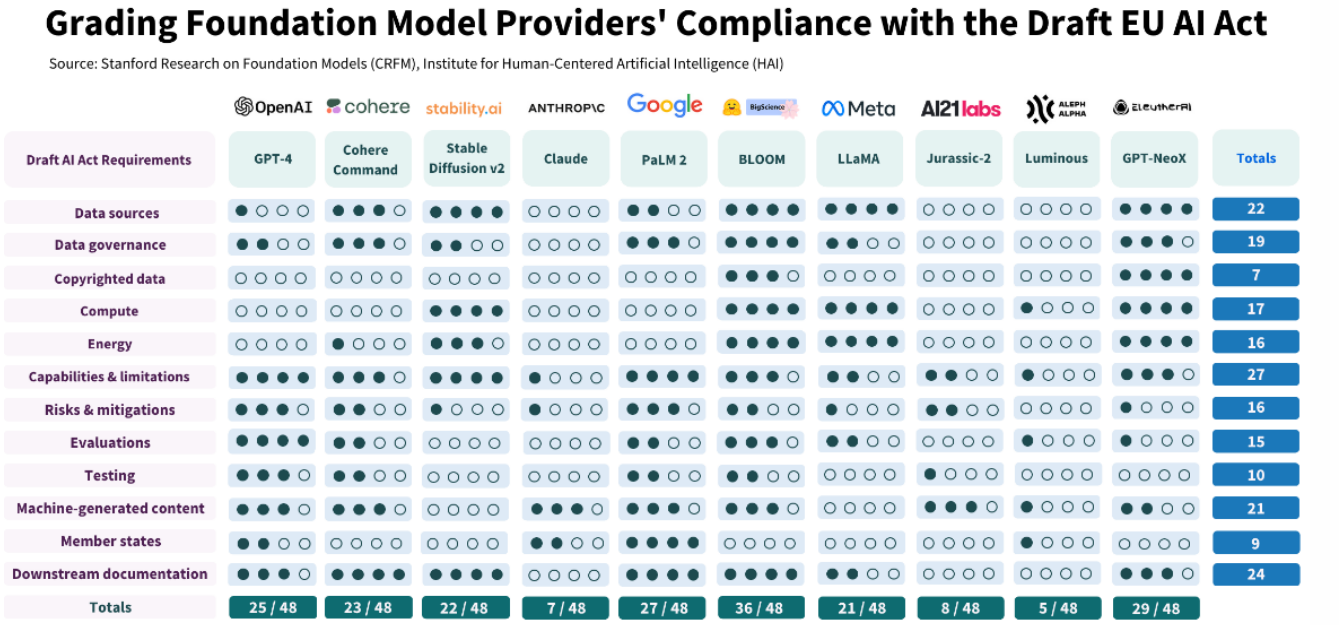

Рис. 4. Языковые модели и Европейский закон об ИИ [32].

Есть готовые фреймворки, которые помогают ориентироваться в этих задачах. Некоторые из них есть известные фреймворки, используемые для описания IT-активов. Например, COBIT [29]. Есть фреймворки, ориентированные прямо на задачи ИИ, например, IIA Artificial Intelligence Auditing Framework [30] или Deloitte's Trustworthy AI Framework [31] (рис. 5).

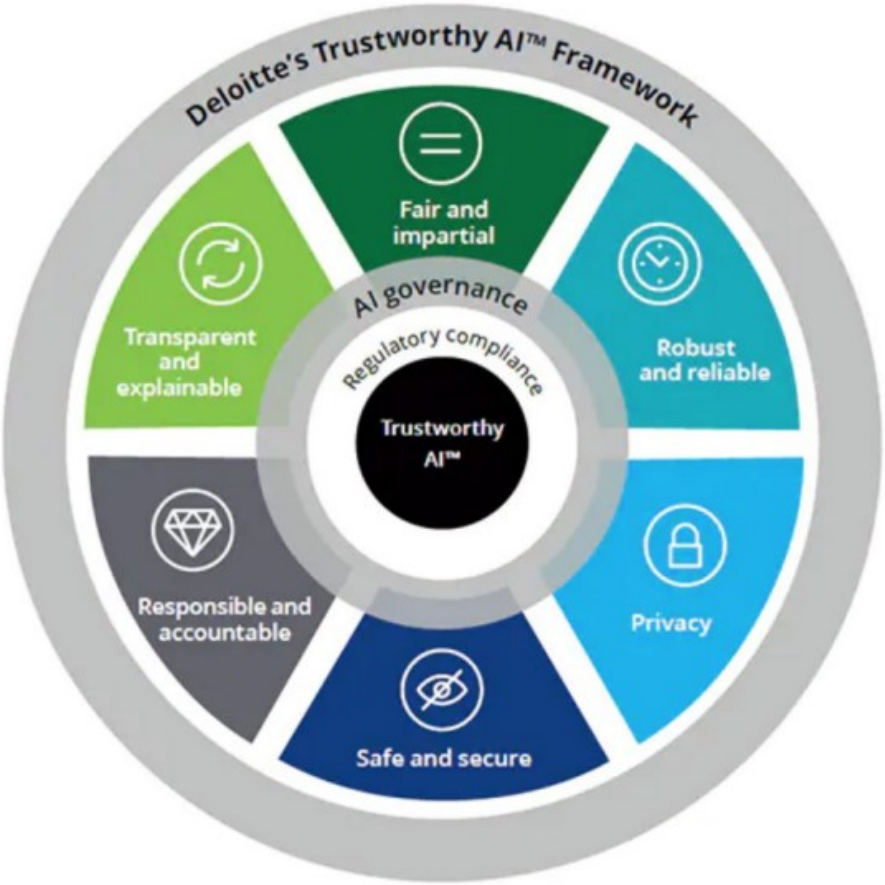

Рис. 5. Deloitte's Trustworthy AI Framework [31].

Говоря о возможной стандартизации, можно отметить NIST AI RISC Management framework [33] и ISO/IEC 23894 [34]. Работы по созданию фреймворков для аудита продолжаются. В работе сотрудников Google [35] представлен еще один такой проект. На рисунке 6 из этой статьи серый цвет указывает на процесс, а цветные секции представляют документы. Документы, выделенные оранжевым цветом, составляются аудиторами, а документы синего цвета — командой разработчиков и специалистов по продукту. «Зеленые продукты» разрабатываются совместно.

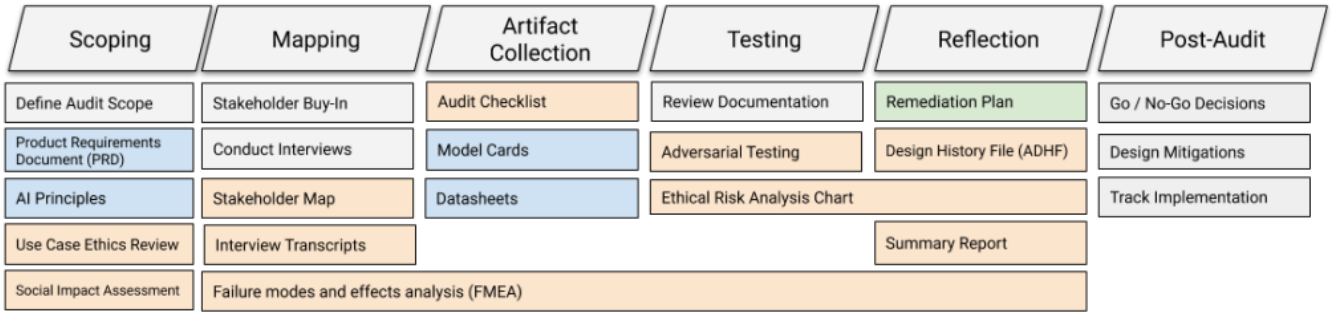

Рис. 6. Фреймворк для аудита алгоритмов ИИ [35].

Другой пример также относится к Google (Deepmind) [36] – это работа, посвященная фреймворку для оценки экстремальных рисков многоцелевых систем ИИ [37]. Работа инспирирована практическим внедрением ChatGPT и других больших языковых моделей. Несмотря на уже достаточную историю фреймворков для аудита, эта работа 2023 года позиционируется авторами как первая. По факту – это набор утверждений о необходимости ответственной тренировки модели и соответствующего внедрения. Универсальный фреймворк для аудита систем ИИ предлагается в работе [38] немецкой промышленной компании TUV. Его основное содержание – учет текстовых требований типа «Модель должна быть устойчива к атаке PGD с бюджетом $\epsilon = 0{,}5$» и «Модель ML должна использовать не более 10% фоновой информации для классификации».

Исследования по аудиту систем ИИ (ML) ведутся в организации Fraunhofer [39]. Выпущенный манифест об аудируемых ИИ-системах [40] предлагает матрицу для оценки (рис. 7), похожую на использованную в упомянутой выше работе Стенфорда [32].

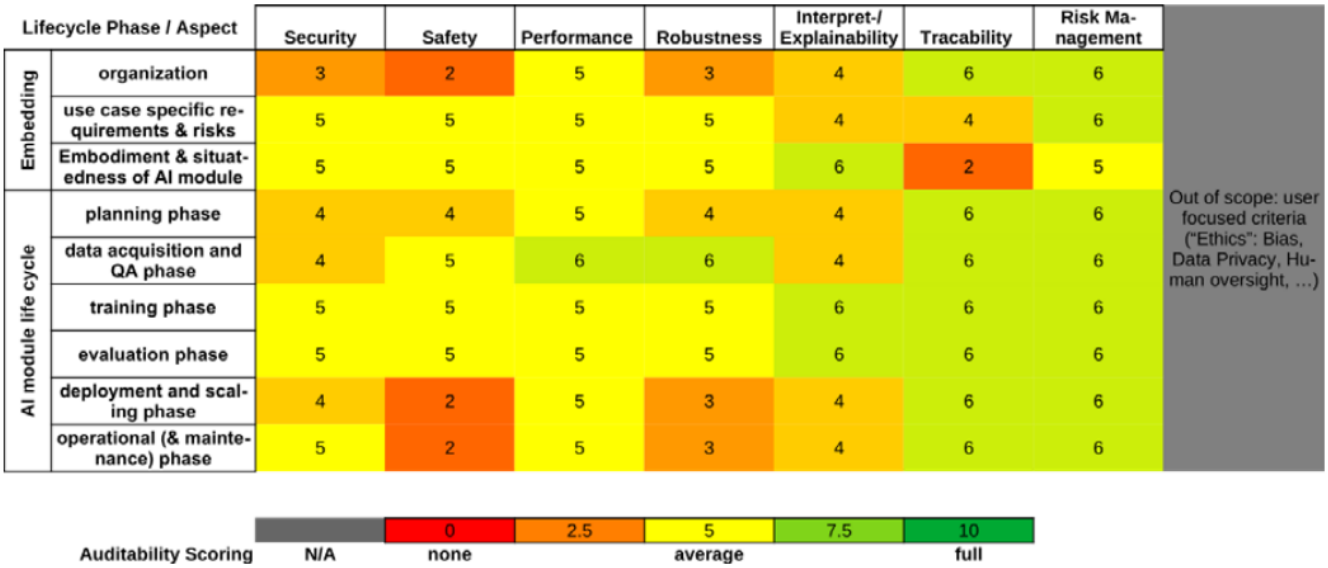

Рис. 7. Оценки аудируемости [40].

Gartner предложил AI TRiSM (Artificial Intelligence (AI) Trust, Risk, and Security Management - управление доверием, рисками и безопасностью искусственного интеллекта) как фреймворк, который обеспечивает управление, надежность, справедливость, эффективность и конфиденциальность ИИ [41]. AI TRiSM фокусируется на:

- Доверии к системам ИИ
- Рисках систем ИИ
- Управлении безопасностью ИИ

Кроме того, Gartner определяет 5 базовых элементов AI TRiSM, на которых можно строить эффективные решения искусственного интеллекта:

- Объяснимость
- ModelOps – по Gartner это руководство и управление жизненным циклом моделей искусственного интеллекта (ИИ) и моделей принятия решений, включая машинное обучение, графы знаний, правила, оптимизацию, лингвистические и агентные модели. Основные возможности включают непрерывную интеграцию, среды разработки моделей, тестирование, управление версиями моделей, хранилище моделей.
- Обнаружение аномалий данных
- Противодействие состязательным атакам
- Защита данных

Сертификация и аудит должны быть обязательными компонентами доверенных платформ для разработки

ИИ приложений [1].

V. Сертификация систем ИИ

Как отмечалось ранее сертификация – это уже гарантия результатов работы системы. Здесь необходимо остановиться на существующей, по факту, разной интерпретацией понятия сертификации для ML систем и программных систем.

Для ML систем (моделей) сертификация – это получение оценок выбранных метрик (в том числе – и вероятностных оценок). Для программ – это гарантия работоспособности. Дословно, например, для систем авионики и сертификации по DO-178: “avionic systems should safely perform their intended function under all foreseeable operating and environmental conditions”. Но - модель машинного обеспечения на этапе вывода – это программа. И, соответственно, она должна сертифицироваться, как и любая другая программа для критических применений. И вероятностные оценки, например, здесь уже вообще не работают. При этом тестируют модели опять-таки на некотором подмножестве данных, а вовсе не на всех возможных и т.д.

Каким же образом может гарантироваться работа систем машинного обучения как программ? Гарантии для программного обеспечения (Software Assurance или SwA) — это критический процесс разработки программного обеспечения, который обеспечивает надежность, безопасность и защищенность программных продуктов. Он включает в себя множество действий: анализ требований, анализ проекта, проверку кода, тестирование и формальную проверку. Одним из важнейших компонентов обеспечения безопасности программного обеспечения являются методы безопасного кодирования, которые соответствуют принятым в отрасли стандартам и передовым методам.

Есть классическая V-модель разработки программного обеспечения [44]. Два направления проверки (рис. 8)

- Верификация – правильно ли мы строим продукт?
- Валидация – построен ли правильный продукт,

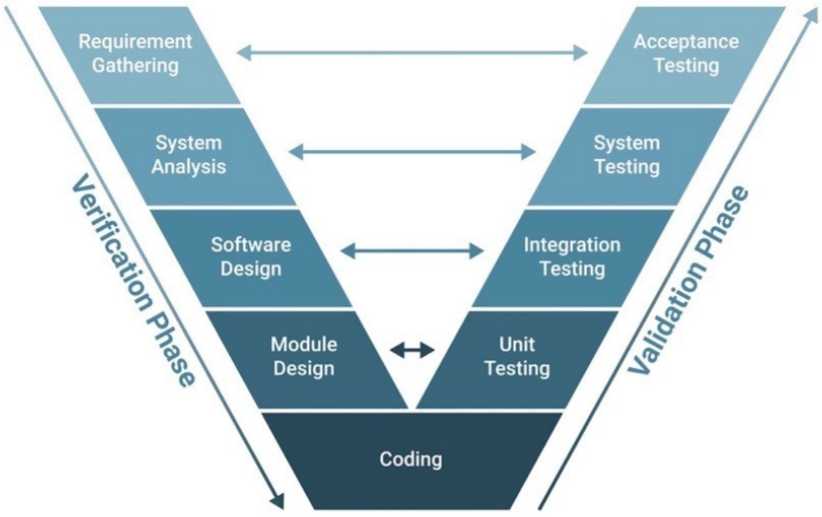

Рис. 8. V-модель жизненного цикла [45]

На каждом уровне есть соответствующий набор тестов. Во время верификации проверяется, соответствует ли продукт требованиям: у него есть все функции для использования по назначению, как описано на этапе планирования после проверки с его потенциальными пользователями, и эти функции работают по назначению. Это подразумевает, во-первых, установление требований, а затем создание на их основе спецификации проекта системы. Затем разработка движется вглубь, уточняя данные предыдущего шага. Во время валидации проверяется, описывают ли требования то, что действительно необходимо, правильно ли они учитывают цели заинтересованных сторон, соответствует ли полученное программное обеспечение модели применения.

Для систем машинного обучения (нейронных сетей) есть, очевидно, компоненты, которые могут быть проверены подобным же образом. Например, анализ входных данных, мониторинг работы системы и т.п. Но ключевая функция (вывод) таким образом (построчно) проверена быть не может. Компания Daedalean и EASA (European Union Aviation Safety Agency) предложили термин Learning Assurance (гарантии обучения) вместо Software Assurance [45] и соответствующую W-модель (рис. 9). Это было опубликовано как концепция обеспечения проектирования для нейронных сетей (Concepts of Design Assurance for Neural Networks или CoDANN) [46]. Эта концепция может стать основой для будущих нормативных требований. Концепция EASA по сертификации ML приложений изложена в их отчетах [53, 54, 55]. По нашему мнению, элемент этой дорожной карты “EASA concept paper: First usable

guidance for level 1 machine learning applications" есть лучший на сегодняшний день документ, который представляет собой аудит для систем машинного обучения.

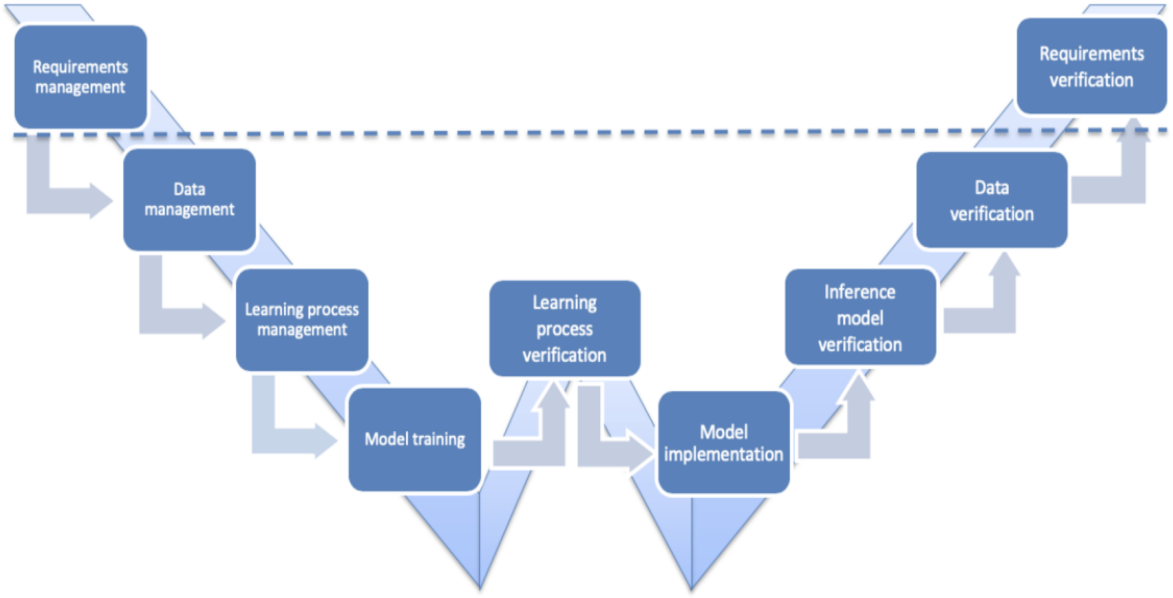

Рис. 9. W-модель [45]

Чем принципиально W-модель отличается от V-модели? Для систем машинного обучения каждый шаг на рисунке 9 существует сам по себе. Например, предполагается, что мы исследуем датасеты и зафиксируем "правильный" вариант. Далее для фиксированного датасета мы отлаживаем модель. После получения устраивающих метрик модель фиксируется и т.д. То есть задача сводится к последовательности детерминированных шагов (на каждом шаге получается некоторый детерминированный результат). Получается, что вопрос возможного сдвига данных вообще выпадает из рассмотрения. И как это процесс будет работать в случае, например, сдвига концепций [3] – совершенно не ясно.

EASA опубликовало дорожную карту для своих проектов сертификации [47]. Последняя версия датирована Май 2023. Приложения ИИ для авиации разбиты на 3 уровня (рис. 10)

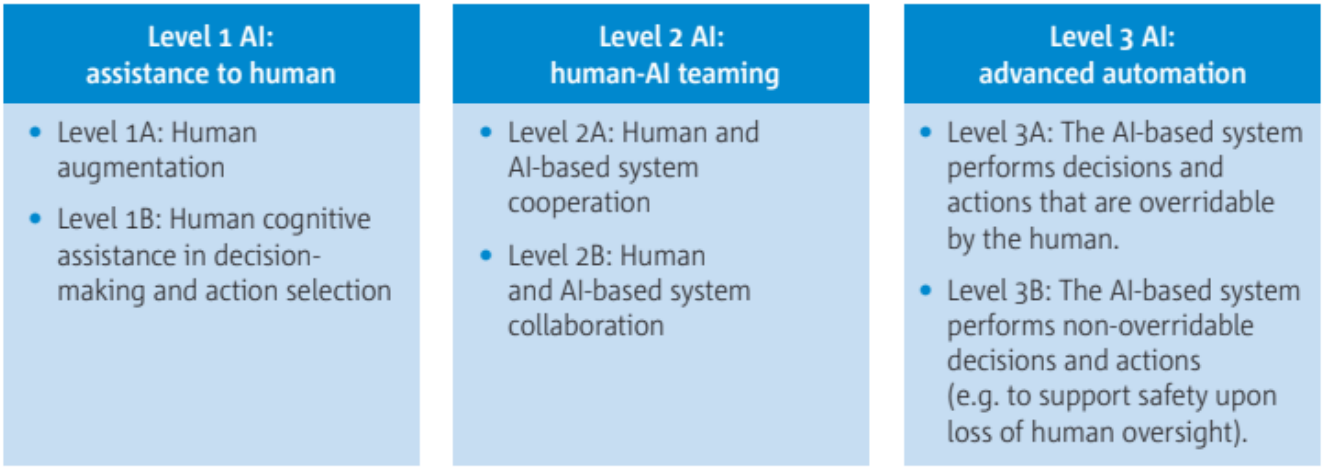

Рис. 10. Классификации приложений ИИ [47]

Сертификация приложений первого уровня (ассистенты для человека) относится к 2025 году, последнего третьего уровня (неотменяемые действия) – к 2035-2050 годам.

Есть стандарты SAE (Society of Automotive Engineers), посвященные искусственному интеллекту [48]. Комитет G-34 по искусственному интеллекту в авиации отвечает за создание и ведение технических отчетов SAE (т. е. отчетов об аэрокосмической информации, аэрокосмической рекомендуемой практики и аэрокосмических стандартов) по аспектам внедрения и сертификации, связанным с технологиями ИИ, включая любые бортовые системы для безопасной эксплуатации аэрокосмических систем и аэрокосмических аппаратов. Рабочие группы включают все необходимые комитеты:

- SG1 - Airborne & Ground Applications
- SG2 - ML Data Management & Validation
- SG3 - ML Design & Verification
- SG4 - ML Implementation & Verification
- SG5 - System & Safety Considerations for ML
- SG7 - Process Considerations (Planning, Config. Mgmt., Quality, Levelling, and Certification/Approval)

Но есть только один публично опубликованный

документ 2021 года, который озаглавлен Artificial Intelligence in Aeronautical Systems: Statement of Concerns. В принципе, название этого документа точно описывает текущее состояние процесса сертификации.

В работе [49] отмечается, количественная оценка безопасности ИИ в авиации пока еще находится в разработке. Дословно: «Процесс машинного обучения по своей природе очень недетерминирован — он должен быть таковым, по крайней мере, на этапе обучения. Однако на этапе развертывания механизмы вывода, которые запускают, например, сверточные нейронные сети и используют эту «обученную информацию», могут быть настроены так, чтобы удовлетворить требования органов сертификации. Не все осознают это или имеют дело с этим, но в какой-то момент детерминизм механизма логического вывода должен решаться каждой системой, которая хочет достичь высокого уровня критичности безопасности». Требование детерминизма понятно, но оно вступает в противоречие с основной концепцией машинного обучений – мы считаем, что выработанные на этапе обучения на тренировочном наборе данных обобщения останутся таковыми для всей генеральной совокупности. В общем случае, без каких-то внешних ограничений на эту генеральную совокупность (то есть на допустимые данные) гарантировать такое нельзя.

В работе [50] авторы, сотрудники Airbus отмечают, что сертификация систем машинного обучения это комплексная проблема (рис. 11)

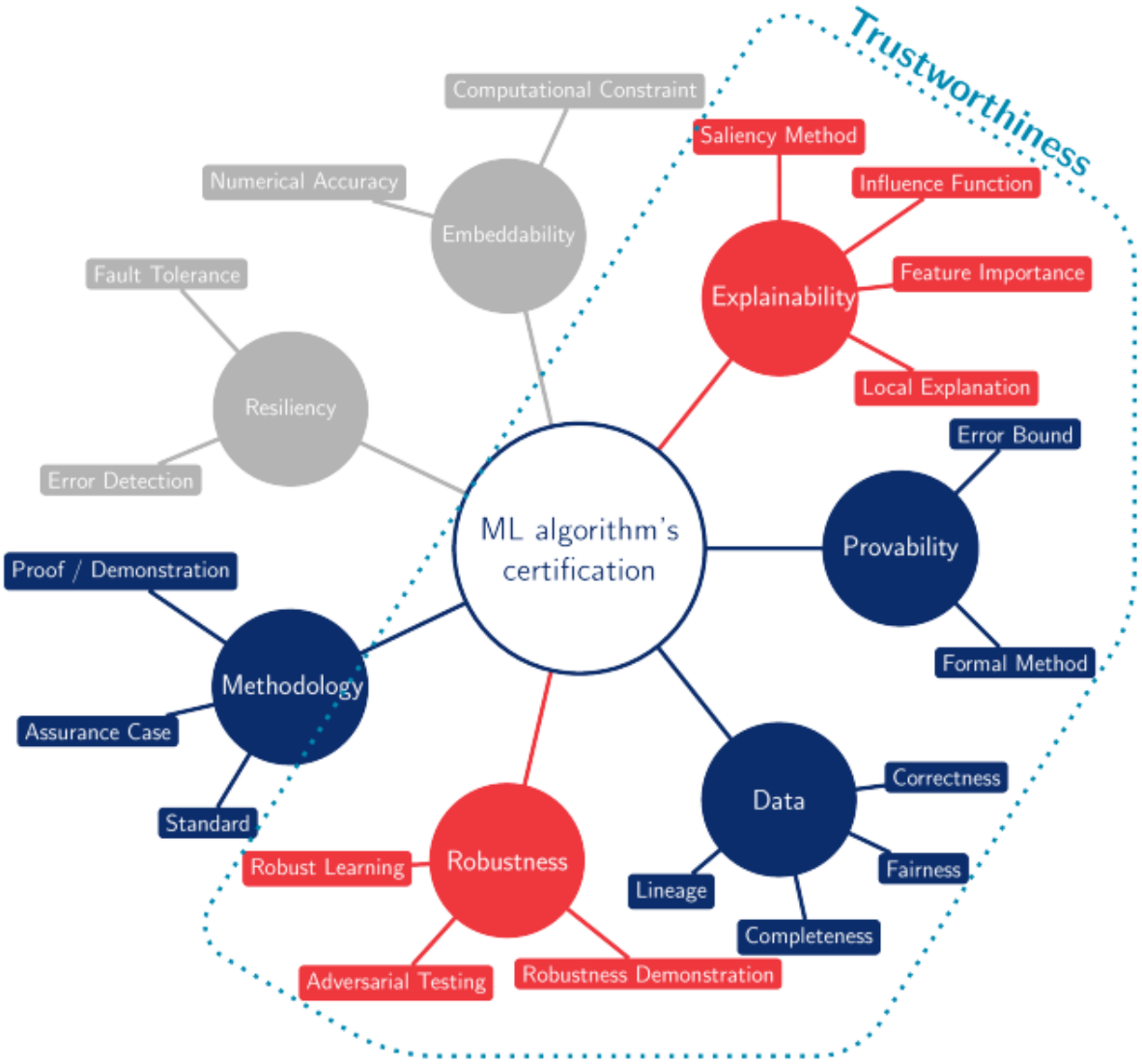

Рис. 11. Сертификация алгоритмов машинного обучения [50]

В этой схеме анализ данных, например, это не только статический анализ тренировочных наборов данных. Безусловно, их необходимо анализировать, поскольку, например, бэкдоры в модели могут оказаться именно из-за специально подготовленных данных на этапе обучения. Но анализ данных необходим и на этапе работы модели. В результате сложности такого процесса, он, во многих работах упрощается и сводится к анализу робастности [50].

Ключевая роль робастности двояка: с одной стороны, и согласно ED-12C/DO-178C, это степень, в которой программное обеспечение может продолжать работать правильно, несмотря на ненормальные входные данные и условия. С другой стороны, и более конкретно для приложения ML, EASA [51] определяет, что система ML

является робастной, когда она выдает одни и те же выходные данные для входных данных, варьирующихся в области пространства состояний. Вариации (возмущения) могут быть естественными (например, шум датчика, смещение измерений и т.п.), вариациями из-за сбоев (например, неверные данные от испорченных датчиков) или преднамеренно вставленными (например, измененными пикселями в изображениях), чтобы обмануть прогнозы модели. Когда возмущенные примеры обманывают алгоритм ML, мы говорим о состязательных примерах. Обычно это определяется как шумы на входах, которые незаметны или не превышают порогового значения.

В работе [52] авторы достаточно подробно останавливаются на принципиальной несовместимости процесса разработки ML приложений и положений DO-178. Основные расхождения могут быть представлены так:

- Детерминированный подход к сертификации программного обеспечения и недетерминированный вывод моделей ML.
- Прослеживание кода – V-модель позволяет определить, для удовлетворения каких требований присутствует определенная строка кода. Это невозможно для моделей ML, поскольку модуль вывода носит общий характер, а логика работы определяется данными.
- Охват данных. Стандартный подход для моделей ML – так называемая точечная робастность. Сертификация моделей ML оперирует ограниченными модификациями корректных данных. Типичное описание – “сертифицированная точность 35.42% на MNIST при бюджете изменений ϵ= 8/255”. О проверке на всем диапазоне данных речи не идет.

VI Сертификация моделей машинного обучения

Эта тема, в принципе, заслуживает отдельного рассмотрения. Суммарно, проблема состоит в том, что понятие “сертификация” трактуется совершенно по-разному для моделей машинного обучения и программных систем (которыми, на самом деле, являются те же модели машинного обучения на этапе вывода). Работа [56] представляет собой наиболее полный известный нам обзор этой темы. Речь идет о сертифицированной робастности для моделей. Робастность, как уже упоминалось выше, не содержит гарантий работоспособности. Но для моделей сертифицируется именно робастность.

Сертифицированная робастность - это установление теоретически подтвержденной нижней границы устойчивости при определенных ограничениях на возмущения. Это еще одно (также упомянутое выше) отличие от сертификации программ – проверяются ограниченные отклонения для корректных данных. Соответственно, подходы к робастному обучению направлены на улучшение этой нижней границы.

Мы можем разделить подходы к проверке на полную и неполную проверку. Когда метод проверки выдает «не верифицировано» для данного x_0, то если гарантировано существование состязательного примера x в заданной окрестности x_0, мы называем это полной проверкой. Если существование состязательного примера не гарантировано, то это - неполная проверка.

Мы также можем разделить подходы к проверке на детерминированную проверку и вероятностную проверку. Когда данные входные данные неустойчивы к атаке (есть состязательные примеры), детерминированная проверка гарантированно выдает «не верифицировано». Вероятностная проверка выдает «не верифицировано» с определенной вероятностью (например, 99,9%), которая не зависит от входных данных. Рисунок 12 представляет существующие подходы.

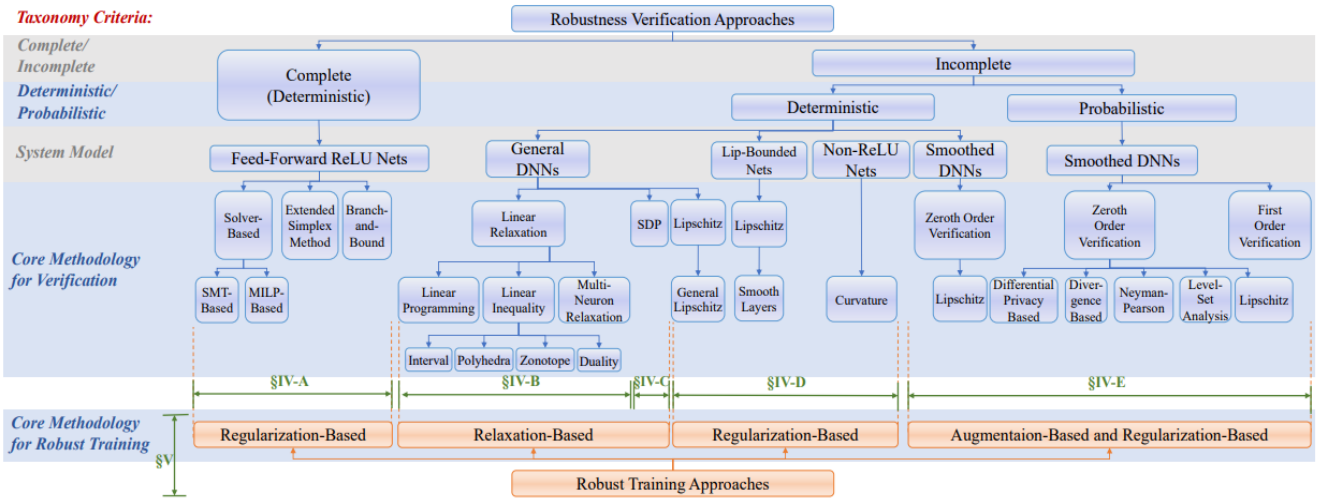

Рис. 12 Верификация робастности [56]

Очевидно, что программная сертификация в существующих требованиях – это полные и детерминированные подходы. Принципы сертификации программ соответствуют тому, что в указанной работе [56] названо глобальной атакой уклонения (global evasion attack). Глобальные атаки уклонения могут

модифицировать любой допустимый входной пример для введения модели в заблуждение, тогда как локальные атаки уклонения могут исказить только данные в распределении входных параметров. Таким образом, устойчивость к глобальным атакам уклонения означает, что свойство устойчивости сохраняется для всей входной области.

Получается, что из всего дерева на рисунке 12, для программной сертификации подходит только самая левая ветвь, которая представляет собой задачи формальной верификации. Такие подходы существуют [57, 58], но имеют очевидные проблемы с масштабированием. Если с поддержкой нейронных сетей с большим количеством параметров какие-то продвижения есть, то с практическим применением все пока остается в зачаточном состоянии. Результаты ежегодных соревнований по верификации [59], показывают, например, что примеров использования за пределами задач классификации нет.

Есть еще одна проблема с сертификацией робастности. В работе [60] приведены конкретные результаты сертификации моделей на выбранных датасетах при заданных ограниченных модификациях. Последний столбец на рисунке 13 – это доля правильных классификаций (accuracy).

Dataset	Method	FLOPs	Test	Robust	Certified
MNIST ($\epsilon = 0.3$)	Group Sort (Anil et al., 2019)	2.9M	97.0	34.0	2.0
	COLT (Balunovic & Vechev, 2020)	4.9M	97.3	-	85.7
	IBP (Gowal et al., 2018)	114M	97.88	93.22	91.79
	CROWN-IBP (Zhang et al., 2020b)	114M	98.18	93.95	92.98
	ℓ_∞-dist Net	82.7M	98.54	94.71	92.64
	ℓ_∞-dist Net+MLP	85.3M	**98.56**	**95.28**	**93.09**
Fashion MNIST ($\epsilon = 0.1$)	CAP (Wong & Kolter, 2018)	0.41M	78.27	68.37	65.47
	IBP (Gowal et al., 2018)	114M	84.12	80.58	77.67
	CROWN-IBP (Zhang et al., 2020b)	114M	84.31	80.22	78.01
	ℓ_∞-dist Net	82.7M	**87.91**	79.64	77.48
	ℓ_∞-dist Net+MLP	85.3M	**87.91**	**80.89**	**79.23**
CIFAR-10 ($\epsilon = 8/255$)	PVT (Dvijotham et al., 2018a)	2.4M	48.64	32.72	26.67
	DiffAI (Mirman et al., 2019)	96.3M	40.2	-	23.2
	COLT (Balunovic & Vechev, 2020)	6.9M	51.7	-	27.5
	IBP (Gowal et al., 2018)	151M	50.99	31.27	29.19
	CROWN-IBP (Zhang et al., 2020b)	151M	45.98	34.58	33.06
	CROWN-IBP (loss fusion) (Xu et al., 2020a)	151M	46.29	35.69	33.38
	ℓ_∞-dist Net	121M	**56.80**	**37.46**	33.30
	ℓ_∞-dist Net+MLP	123M	50.80	37.06	**35.42**

Рис. 13 Сертифицированная робастность [60].

Такого рода значения в критических технических системах принято измерять в так называемых “девятках” (99.999 – количество цифр 9 в дробной части). Как видно из представленных данных, реальные результаты сертификации много ниже того, что вообще может использоваться на практике.

VII Заключение

Суммируя существующие возможности гарантирования результатов работы систем машинного обучения, можно отметить следующее.

Во-первых, следует отметить, что правовые нормы лишь описывают конечные требования к продукции и, соответственно, не имеют ничего общего ни с процессом достижения (удовлетворения) этих требований, ни с самой процедурой проверки соответствия. Регламент в его нынешнем виде содержит некоторые рекомендации, но их следует считать достаточно очевидными.

Сертификация систем машинного обучения, как это понимают для традиционного программного обеспечения, сегодня вообще невозможна. Работающий на сегодняшний день подход — это формальная проверка моделей машинного обучения, но у него есть проблемы как с масштабируемостью, так и с практическим применением. Возможно, решением проблемы сертификации систем машинного обучения будет изменение существующих стандартов сертификации. Но это может быть проблемным, поскольку тогда появится программное обеспечение разных сортов: сертифицированное по старым и новым правилам.

С практической точки зрения сертификация моделей машинного обучения сейчас — это сертификация робастности, когда метрики гарантируются для заданного бюджета (размера) модификации обучающих данных. Другой подход, используемый на практике, заключается в моделировании возможных проблем с измерительными устройствами (например, камерами). Это можно назвать семантически обусловленными изменениями. Перспективно также изучение проблемы глобальных атак уклонения.

Следующий этап — аудит систем машинного обучения. С практической точки зрения аудит — это, прежде всего, чек-лист, в котором указан список необходимых действий (процедур) на разных этапах стандартного конвейера моделей машинного обучения. Деятельность влечет за собой создание документов, описывающих характеристики проверяемых моделей. Сегодня это практические и абсолютно осуществимые процедуры, которые следует применять на практике для всех промышленных моделей машинного обучения. В отдельной работе мы изложим наши предложения по

использованию концептуального документа EASA для построения корпоративных (отраслевых или даже национальных) систем аудита.

Благодарности

Авторы благодарны сотрудникам кафедры Информационной безопасности факультета ВМК МГУ имени М.В. Ломоносова за ценные обсуждения. Также хотелось бы отметить юбилей публикаций В.П. Куприяновского, которые положили начало всей цифровой повестке в журнале INJOIT [61, 62].

Библиография

[1] Namiot, Dmitry, Eugene Ilyushin, and Oleg Pilipenko. "On Trusted AI Platforms." International Journal of Open Information Technologies 10.7 (2022): 119-127.
[2] Namiot, Dmitry. "Introduction to Data Poison Attacks on Machine Learning Models." International Journal of Open Information Technologies 11.3 (2023): 58-68.
[3] Намиот, Д. Е., and Е. А. Ильюшин. "Мониторинг сдвига данных в моделях машинного обучения." International Journal of Open Information Technologies 10.12: 2022.
[4] Namiot, Dmitry. "Schemes of attacks on machine learning models." International Journal of Open Information Technologies 11.5 (2023): 68-86.
[5] 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.
[6] MITRE Atlas mitigations https://atlas.mitre.org/mitigations/ Retrieved: Dec, 2023
[7] GRID 2023 https://indico.jinr.ru/event/3505/ Retrieved: Dec, 2023
[8] Namiot, Dmitry, and Manfred Sneps-Sneppe. "On Audit and Certification of Machine Learning Systems." 2023 34th Conference of Open Innovations Association (FRUCT). IEEE, 2023.
[9] Robust and Verified Deep Learning group https://deepmindsafetyresearch.medium.com/towards-robust-and-verified-ai-specification-testing-robust-training-and-formal-verification-69bd1bc48bda Retrieved: Dec, 2023
[10] Madry Lab https://people.csail.mit.edu/madry/6.S979/files/lecture_4.pdf Retrieved: Dec, 2023
[11] Kostyumov, Vasily. "A survey and systematization of evasion attacks in computer vision." International Journal of Open Information Technologies 10.10 (2022): 11-20. (in Russian)
[12] Song, Junzhe, and Dmitry Namiot. "A Survey of the Implementations of Model Inversion Attacks." Distributed Computer and Communication Networks: 25th International Conference, DCCN 2022, Moscow, Russia, September 26–29, 2022, Revised Selected Papers. Cham: Springer Nature Switzerland, 2023.
[13] Bidzhiev, Temirlan, and Dmitry Namiot. "Research of existing approaches to embedding malicious software in artificial neural networks." International Journal of Open Information Technologies 10.9 (2022): 21-31. (in Russian)
[14] Goodfellow, I.J.; Shlens, J.; Szegedy, C. Explaining and harnessing adversarial examples. arXiv, 2014; arXiv:1412.6572.
[15] 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. (in Russian)
[16] Borg, Markus, et al. "Safely entering the deep: A review of verification and validation for machine learning and a challenge elicitation in the automotive industry." arXiv preprint arXiv:1812.05389 (2018)
[17] Why Robustness is not Enough for Safety and Security in Machine Learning https://towardsdatascience.com/why-robustness-is-not-enoughfor-safety-and-security-in-machine-learning-1a35f6706601 Retrieved: Jun, 2023
[18] Gu, Kang, et al. "Towards Sentence Level Inference Attack Against Pre-trained Language Models." Proceedings on Privacy Enhancing Technologies 3 (2023): 62-78.
[19] OWASP Top 10 List for Large Language Models version 0.1 https://owasp.org/www-project-top-10-for-large-language-model-applications/descriptions/
[20] Derner, Erik, and Kristina Batistič. "Beyond the Safeguards: Exploring the Security Risks of ChatGPT." arXiv preprint arXiv:2305.08005 (2023).
[21] Democratic inputs to AI https://openai.com/blog/democratic-inputs-to-ai Retrieved: Dec, 2023
[22] The AI Act https://artificialintelligenceact.eu/ Retrieved: Dec, 2023
[23] AI regulation https://www.technologyreview.com/2023/05/23/1073526/suddenly-everyone-wants-to-talk-about-how-to-regulate-ai/ Retrieved: Dec, 2023
[24] Schuett, Jonas, et al. "Towards best practices in AGI safety and governance: A survey of expert opinion." arXiv preprint arXiv:2305.07153 (2023).
[25] Game Changers https://www.cbinsights.com/research/report/game-changing-technologies-2022/ Retrieved: Dec, 2023
[26] An In-Depth Guide To Help You Start Auditing Your AI Models https://censius.ai/blogs/ai-audit-guide Retrieved: Dec, 2023
[27] Jie Liu. 2012. The enterprise risk management and the risk oriented internal audit. Ibusiness 4, 03 (2012), 287.
[28] IEEE. 2008. IEEE Standard for Software Reviews and Audits. IEEE Std 1028-2008 (Aug 2008), 1–53. https://doi.org/10.1109/IEEESTD.2008.4601584
[29] van Wyk, Jana, and Riaan Rudman. "COBIT 5 compliance: best practices cognitive computing risk assessment and control checklist." Meditari Accountancy Research (2019).
[30] The IIA's Artificial Intelligence Auditing Framework https://www.theiia.org/en/content/articles/global-perspectives-and-insights/2017/the-iias-artificial-intelligence-auditing-framework-practical-applications-part-ii/ Retrieved: Dec, 2023
[31] REALIZE THE FULL POTENTIAL OF ARTIFICIAL INTELLIGENCE https://www.coso.org/Shared%20Documents/Realize-the-Full-Potential-of-Artificial-Intelligence.pdf Retrieved: Dec, 2023
[32] Do Foundation Model Providers Comply with the Draft EU AI Act? https://crfm.stanford.edu/2023/06/15/eu-ai-act.html Retrieved: Dec, 2023
[33] AI Risk Management Framework https://www.nist.gov/itl/ai-risk-management-framework Retrieved: Dec, 2023
[34] ISO/IEC 23894 – A new standard for risk management of AI https://aistandardshub.org/a-new-standard-for-ai-risk-management Retrieved: Dec, 2023
[35] Raji, Inioluwa Deborah, et al. "Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing." Proceedings of the 2020 conference on fairness, accountability, and transparency. 2020.
[36] New research proposes a framework for evaluating general-purpose models against novel threats https://www.deepmind.com/blog/an-early-warning-system-for-novel-ai-risks Retrieved: Jun, 2023
[37] Shevlane, Toby, et al. "Model evaluation for extreme risks." arXiv preprint arXiv:2305.15324 (2023).
[38] Markert, Thora, Fabian Langer, and Vasilios Danos. "GAFAI: Proposal of a Generalized Audit Framework for AI." INFORMATIK 2022 (2022).
[39] Auditing and Certification of AI Systems https://www.hhi.fraunhofer.de/en/departments/ai/technologies-and-solutions/auditing-and-certification-of-ai-systems.html Retrieved: Dec, 2023
[40] Towards Auditable AI Systems Retrieved: Jun, 2023 https://www.hhi.fraunhofer.de/fileadmin/Departments/AI/TechnologiesAndSolutions/AuditingAndCertificationOfAiSystems/2022-05-23-whitepaper-tuev-bsi-hhi-towards-auditable-ai-systems.pdf Retrieved: Dec 2023
[41] AI TRiSM https://www.gartner.com/en/information-technology/glossary/ai-trism Retrieved: Dec, 2023
[42] Datarobot https://www.datarobot.com/platform/trusted-ai/ Retrieved: Dec, 2023
[43] IBM Trustworthy https://research.ibm.com/topics/trustworthy-ai Retrieved: Dec, 2023
[44] Ruparelia, Nayan B. "Software development lifecycle models." ACM SIGSOFT Software Engineering Notes 35.3 (2010): 8-13.
[45] Explaining W-shaped Learning Assurance https://daedalean.ai/tpost/pxl6ih0yc1-explaining-w-shaped-learning-assurance Retrieved: Dec, 2023
[46] Force, DA EASA AI Task, and A. G. Daedalean. "Concepts of Design Assurance for Neural Networks (CoDANN)." Concepts of Design Assurance for Neural Networks (CoDANN). EASA, Daedalean (2020).
[47] EASA roadmap https://www.easa.europa.eu/en/domains/research-innovation/ai Retrieved: Dec, 2023
[48] G-34 Artificial Intelligence in Aviation https://standardsworks.sae.org/standards-committees/g-34-artificial-intelligence-aviation Retrieved: Dec, 2023
[49] DO-178 continues to adapt to emerging digital technologies https://militaryembedded.com/avionics/safety-certification/do-178-continues-to-adapt-to-emerging-digital-technologies Retrieved: Dec, 2023

[50] Vidot, Guillaume, et al. "Certification of embedded systems based on Machine Learning: A survey." arXiv preprint arXiv:2106.07221 (2021).
[51] EASA Artificial Intelligence Roadmap 1.0. https://www.easa.europa.eu/sites/default/files/dfu/EASA-AIRoadmap-v1.0.pdf Retrieved: Dec, 2023
[52] Dmitriev, Konstantin, Johann Schumann, and Florian Holzapfel. "Towards Design Assurance Level C for Machine-Learning Airborne Applications." 2022 IEEE/AIAA 41st Digital Avionics Systems Conference (DASC). IEEE, 2022.
[53] "Concepts of design assurance for neural networks (CoDANN)," European Aviation Safety Agency, Tech. Rep., 2020.
[54] "Report. concepts of design assurance for neural networks (CoDANN) II," European Aviation Safety Agency, Tech. Rep., 2021.
[55] "EASA concept paper: First usable guidance for level 1 machine learning applications," European Aviation Safety Agency, Tech. Rep., 2021.
[56] Li, Linyi, Tao Xie, and Bo Li. "Sok: Certified robustness for deep neural networks." 2023 IEEE symposium on security and privacy (SP). IEEE, 2023.
[57] Namiot, Dmitry, Eugene Ilyushin, and Ivan Chizhov. "On a formal verification of machine learning systems." International Journal of Open Information Technologies 10.5 (2022): 30-34.
[58] Stroeva, Ekaterina, and Aleksey Tonkikh. "Methods for Formal Verification of Artificial Neural Networks: A Review of Existing Approaches." International Journal of Open Information Technologies 10.10 (2022): 21-29.
[59] Brix, Christopher, et al. "The Fourth International Verification of Neural Networks Competition (VNN-COMP 2023): Summary and Results." arXiv preprint arXiv:2312.16760 (2023).
[60] Zhang, Bohang, et al. "Towards certifying robustness using neural networks with l-dist neurons." arXiv preprint arXiv:2102.05363 (2021).
[61] Куприяновский, В. П., Д. Е. Намиот, and С. А. Синягов. "Демистификация цифровой экономики." International Journal of Open Information Technologies 4.11 (2016): 59-63.
[62] Куприяновский, В. П., et al. "Розничная торговля в цифровой экономике." International Journal of Open Information Technologies 4.7 (2016): 1-12.

Trusted Artificial Intelligence Platforms: Certification and Audit

Dmitry Namiot, Eugene Ilyushin

Abstract— **Artificial intelligence systems in this work refer to machine learning systems. It is machine learning (deep learning) systems that are, today, the main examples of the use of Artificial Intelligence in a wide variety of areas. From a practical point of view, we can say that machine learning is synonymous with the concept of Artificial Intelligence. At the same time, machine learning systems, by their nature, depend on the data on which they are trained and, in principle, produce non-deterministic results. Trusted platforms, as their name suggests, are a set of tools designed to increase trust (user confidence) in the output of machine learning models. The use of machine learning systems in so-called critical areas (avionics, automatic driving, etc.) requires guarantees of software functionality, which is confirmed by a certification procedure (process). And by audit, we mean the identification of possible problems with the performance and security of machine learning systems.**

Keywords— **machine learning, robustness, certification, audit.**

REFERENCES

[1] Namiot, Dmitry, Eugene Ilyushin, and Oleg Pilipenko. "On Trusted AI Platforms." International Journal of Open Information Technologies 10.7 (2022): 119-127.
[2] Namiot, Dmitry. "Introduction to Data Poison Attacks on Machine Learning Models." International Journal of Open Information Technologies 11.3 (2023): 58-68.
[3] Namiot, D. E., and E. A. Il'jushin. "Monitoring sdviga dannyh v modeljah mashinnogo obuchenija." International Journal of Open Information Technologies 10.12: 2022.
[4] Namiot, Dmitry. "Schemes of attacks on machine learning models." International Journal of Open Information Technologies 11.5 (2023): 68-86.
[5] 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.
[6] MITRE Atlas mitigations https://atlas.mitre.org/mitigations/ Retrieved: Dec, 2023
[7] GRID 2023 https://indico.jinr.ru/event/3505/ Retrieved: Dec, 2023
[8] Namiot, Dmitry, and Manfred Sneps-Sneppe. "On Audit and Certification of Machine Learning Systems." 2023 34th Conference of Open Innovations Association (FRUCT). IEEE, 2023.
[9] Robust and Verified Deep Learning group https://deepmindsafetyresearch.medium.com/towards-robust-and-verified-ai-specification-testing-robust-training-and-formal-verification-69bd1bc48bda Retrieved: Dec, 2023
[10] Madry Lab https://people.csail.mit.edu/madry/6.S979/files/lecture_4.pdf Retrieved: Dec, 2023
[11] Kostyumov, Vasily. "A survey and systematization of evasion attacks in computer vision." International Journal of Open Information Technologies 10.10 (2022): 11-20. (in Russian)
[12] Song, Junzhe, and Dmitry Namiot. "A Survey of the Implementations of Model Inversion Attacks." Distributed Computer and Communication Networks: 25th International Conference, DCCN 2022, Moscow, Russia, September 26–29, 2022, Revised Selected Papers. Cham: Springer Nature Switzerland, 2023.
[13] Bidzhiev, Temirlan, and Dmitry Namiot. "Research of existing approaches to embedding malicious software in artificial neural networks." International Journal of Open Information Technologies 10.9 (2022): 21-31. (in Russian)
[14] Goodfellow, I.J.; Shlens, J.; Szegedy, C. Explaining and harnessing adversarial examples. arXiv, 2014; arXiv:1412.6572.
[15] 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. (in Russian)
[16] Borg, Markus, et al. "Safely entering the deep: A review of verification and validation for machine learning and a challenge elicitation in the automotive industry." arXiv preprint arXiv:1812.05389 (2018)
[17] Why Robustness is not Enough for Safety and Security in Machine Learning https://towardsdatascience.com/why-robustness-is-not-enoughfor-safety-and-security-in-machine-learning-1a35f6706601 Retrieved: Jun, 2023
[18] Gu, Kang, et al. "Towards Sentence Level Inference Attack Against Pre-trained Language Models." Proceedings on Privacy Enhancing Technologies 3 (2023): 62-78.
[19] OWASP Top 10 List for Large Language Models version 0.1 https://owasp.org/www-project-top-10-for-large-language-model-applications/descriptions/
[20] Derner, Erik, and Kristina Batistič. "Beyond the Safeguards: Exploring the Security Risks of ChatGPT." arXiv preprint arXiv:2305.08005 (2023).
[21] Democratic inputs to AI https://openai.com/blog/democratic-inputs-to-ai Retrieved: Dec, 2023
[22] The AI Act https://artificialintelligenceact.eu/ Retrieved: Dec, 2023
[23] AI regulation https://www.technologyreview.com/2023/05/23/1073526/suddenly-everyone-wants-to-talk-about-how-to-regulate-ai/ Retrieved: Dec, 2023
[24] Schuett, Jonas, et al. "Towards best practices in AGI safety and governance: A survey of expert opinion." arXiv preprint arXiv:2305.07153 (2023).
[25] Game Changers https://www.cbinsights.com/research/report/game-changing-technologies-2022/ Retrieved: Dec, 2023
[26] An In-Depth Guide To Help You Start Auditing Your AI Models https://censius.ai/blogs/ai-audit-guide Retrieved: Dec, 2023
[27] Jie Liu. 2012. The enterprise risk management and the risk oriented internal audit. Ibusiness 4, 03 (2012), 287.
[28] IEEE. 2008. IEEE Standard for Software Reviews and Audits. IEEE Std 1028-2008 (Aug 2008), 1–53. https://doi.org/10.1109/IEEESTD.2008.4601584
[29] van Wyk, Jana, and Riaan Rudman. "COBIT 5 compliance: best practices cognitive computing risk assessment and control checklist." Meditari Accountancy Research (2019).
[30] The IIA's Artificial Intelligence Auditing Framework https://www.theiia.org/en/content/articles/global-perspectives-and-insights/2017/the-iias-artificial-intelligence-auditing-framework-practical-applications-part-ii/ Retrieved: Dec, 2023
[31] REALIZE THE FULL POTENTIAL OF ARTIFICIAL INTELLIGENCE https://www.coso.org/Shared%20Documents/Realize-the-Full-Potential-of-Artificial-Intelligence.pdf Retrieved: Dec, 2023
[32] Do Foundation Model Providers Comply with the Draft EU AI Act? https://crfm.stanford.edu/2023/06/15/eu-ai-act.html Retrieved: Dec, 2023
[33] AI Risk Management Framework https://www.nist.gov/itl/ai-risk-management-framework Retrieved: Dec, 2023
[34] ISO/IEC 23894 – A new standard for risk management of AI https://aistandardshub.org/a-new-standard-for-ai-risk-management Retrieved: Dec, 2023
[35] Raji, Inioluwa Deborah, et al. "Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing." Proceedings of the 2020 conference on fairness, accountability, and transparency. 2020.
[36] New research proposes a framework for evaluating general-purpose models against novel threats https://www.deepmind.com/blog/an-early-warning-system-for-novel-ai-risks Retrieved: Jun, 2023
[37] Shevlane, Toby, et al. "Model evaluation for extreme risks." arXiv preprint arXiv:2305.15324 (2023).

[38] Markert, Thora, Fabian Langer, and Vasilios Danos. "GAFAI: Proposal of a Generalized Audit Framework for AI." INFORMATIK 2022 (2022).
[39] Auditing and Certification of AI Systems https://www.hhi.fraunhofer.de/en/departments/ai/technologies-and-solutions/auditing-and-certification-of-ai-systems.html Retrieved: Dec, 2023
[40] Towards Auditable AI Systems Retrieved: Jun, 2023 https://www.hhi.fraunhofer.de/fileadmin/Departments/AI/TechnologiesAndSolutions/AuditingAndCertificationOfAiSystems/2022-05-23-whitepaper-tuev-bsi-hhi-towards-auditable-ai-systems.pdf Retrieved: Dec 2023
[41] AI TRiSM https://www.gartner.com/en/information-technology/glossary/ai-trism Retrieved: Dec, 2023
[42] Datarobot https://www.datarobot.com/platform/trusted-ai/ Retrieved: Dec, 2023
[43] IBM Trustworthy https://research.ibm.com/topics/trustworthy-ai Retrieved: Dec, 2023
[44] Ruparelia, Nayan B. "Software development lifecycle models." ACM SIGSOFT Software Engineering Notes 35.3 (2010): 8-13.
[45] Explaining W-shaped Learning Assurance https://daedalean.ai/tpost/pxl6ih0yc1-explaining-w-shaped-learning-assurance Retrieved: Dec, 2023
[46] Force, DA EASA AI Task, and A. G. Daedalean. "Concepts of Design Assurance for Neural Networks (CoDANN)." Concepts of Design Assurance for Neural Networks (CoDANN). EASA, Daedalean (2020).
[47] EASA roadmap https://www.easa.europa.eu/en/domains/research-innovation/ai Retrieved: Dec, 2023
[48] G-34 Artificial Intelligence in Aviation https://standardsworks.sae.org/standards-committees/g-34-artificial-intelligence-aviation Retrieved: Dec, 2023
[49] DO-178 continues to adapt to emerging digital technologies https://militaryembedded.com/avionics/safety-certification/do-178-continues-to-adapt-to-emerging-digital-technologies Retrieved: Dec, 2023
[50] Vidot, Guillaume, et al. "Certification of embedded systems based on Machine Learning: A survey." arXiv preprint arXiv:2106.07221 (2021).
[51] EASA Artificial Intelligence Roadmap 1.0. https://www.easa.europa.eu/sites/default/files/dfu/EASA-AIRoadmap-v1.0.pdf Retrieved: Dec, 2023
[52] Dmitriev, Konstantin, Johann Schumann, and Florian Holzapfel. "Towards Design Assurance Level C for Machine-Learning Airborne Applications." 2022 IEEE/AIAA 41st Digital Avionics Systems Conference (DASC). IEEE, 2022.
[53] "Concepts of design assurance for neural networks (CoDANN)," European Aviation Safety Agency, Tech. Rep., 2020.
[54] "Report. concepts of design assurance for neural networks (CoDANN) II," European Aviation Safety Agency, Tech. Rep., 2021.
[55] "EASA concept paper: First usable guidance for level 1 machine learning applications," European Aviation Safety Agency, Tech. Rep., 2021.
[56] Li, Linyi, Tao Xie, and Bo Li. "Sok: Certified robustness for deep neural networks." 2023 IEEE symposium on security and privacy (SP). IEEE, 2023.
[57] Namiot, Dmitry, Eugene Ilyushin, and Ivan Chizhov. "On a formal verification of machine learning systems." International Journal of Open Information Technologies 10.5 (2022): 30-34.
[58] Stroeva, Ekaterina, and Aleksey Tonkikh. "Methods for Formal Verification of Artificial Neural Networks: A Review of Existing Approaches." International Journal of Open Information Technologies 10.10 (2022): 21-29.
[59] Brix, Christopher, et al. "The Fourth International Verification of Neural Networks Competition (VNN-COMP 2023): Summary and Results." arXiv preprint arXiv:2312.16760 (2023).
[60] Zhang, Bohang, et al. "Towards certifying robustness using neural networks with l-dist neurons." arXiv preprint arXiv:2102.05363 (2021).
[61] Kuprijanovskij, V. P., D. E. Namiot, and S. A. Sinjagov. "Demistifikacija cifrovoj jekonomiki." International Journal of Open Information Technologies 4.11 (2016): 59-63.
[62] Kuprijanovskij, V. P., et al. "Roznichnaja torgovlja v cifrovoj jekonomike." International Journal of Open Information Technologies 4.7 (2016): 1-12.