Научная статья
УДК 004.896

Принципы испытаний образцов вооружения, военной и специальной техники с реализацией технологии машинного обучения (полемические заметки)

Вячеслав Константинович Абросимов

Аннотация. Цель работы: введение в научный оборот принципиально новых принципов испытаний образцов интеллектуальных систем (ИС), для обучения систем управления которых использованы технологии глубокого машинного обучения. Метод исследования: системный анализ существующих подходов, опыт практической работы по испытаниям сложных интеллектуальных систем в интересах укрепления обороноспособности страны. Результаты исследования: сформулированы положения о принципиальной допустимости ошибок, совершаемых интеллектуальными системами военного назначения, объективной невозможности испытаний во всех условиях боевого применения, слабой значимости ошибок второго рода, объяснимости результата и др. Реализация принципов направлена на формирование максимально сложных условий для объекта испытаний в интересах повышения качества проверки тактико-технических характеристик образцов ВВСТ на испытаниях и повышения эффективности последующей боевой эксплуатации.

Ключевые слова: интеллектуальность; образец вооружения; испытания; машинное обучение; ошибка; проверка

Для цитирования: Абросимов В.К. Принципы испытаний образцов вооружения, военной и специальной техники с реализацией технологии машинного обучения (полемические заметки) // Вооружение и экономика. 2024. №2(68). С. 23-32.

Original article

The Principles of Weapon Samples Tests with the Use of Machine Learning Technology Implementation (Polemical Notes)

Viacheslav K. Abrosimov

Abstract. The aim of the work: introduction of the fundamentally new principles of dual-purpose intelligent systems samples tests into scientific discourse, for the control system training of which deep learning technologies are used. Research method: system analysis of existing approaches, practical experience in complicated intelligent systems tests in the interests of the country's defense capability strengthening. Research results: the provisions are formulated regarding the fundamental tolerability of errors made by the dual-purpose intelligent systems, the objective impossibility of tests under all conditions of use, the weak significance of type 2 errors, the result justification, etc. The implementation of the principles is pointed to the most difficult conditions creation for the test item in the interest of the weapon sample tactical and technical characteristics quality test improvement in the course of tests and subsequent operation effectiveness.

Keywords: intelligence; weapon sample; tests; machine learning; error; verification

For citation: Abrosimov V.K. The Principles of Weapon Samples Tests with the Use of Machine Learning Technology Implementation (Polemical Notes) // Armament and Economics. 2024. No.2(68). P. 23-32.

Введение

В настоящее время процессы испытаний сложных интеллектуальных систем (ИС), разрабатываемых в интересах создания объектов двойного назначения, достаточно хорошо отработаны. Однако, активное внедрение технологий искусственного интеллекта (ИИ) в разработку новых ИС требует совершенно новых подходов к проведению испытаний. Указанное вызвано, прежде всего, существенными рисками внедрения таких технологий-непрозрачностью получаемых решений, проблемами с адекватностью исходной информации, объемами и корректностью обучающих выборок, возможностью так называемых «злонамеренных» атак как в процессе обучения, так и в процессе эксплуатации и др. [1]. В настоящее время проблема таких испытаний становится исключительно актуальной.

1 Постановка задачи

Анализ доступной литературы показал как разнообразие потенциально возможных подходов, так и отсутствие общепринятых не только решений, но даже и принципов испытаний. Ясно, что определяющим при формировании принципов испытаний должны являться руководящие документы. Но ГОСТ 16504-81[1], раскрывающий требования к государственным испытаниям продукции, естественно не предусматривает положений, связанных с испытаниями сложных систем, разрабатываемых на основе технологий искусственного интеллекта. Более современный ГОСТ Р 59276-2020[2], определяющий способы обеспечения доверия к системам искусственного интеллекта, также не фиксирует методы и способы испытаний таких систем, но связан с этими вопросами опосредованно, вводя практически полезные критерии и показатели оценки. Так, в таблице 3 ГОСТ Р 59276-2020, содержащей факторы снижения качества, специфичные для систем ИИ на стадиях их создания и эксплуатации, даже содержатся своего рода подсказки для разработки методик испытаний – наличие преднамеренных искажений в обучающей выборке, использование состязательных атак, приводящих к неустойчивой работе систем в процессе их эксплуатации, вторжение и нарушение конфиденциальности данных и др.

Первые же разработанные стандарты ТК 164 Госстандарта в области испытаний систем с ИИ, к сожалению, вообще не учитывают специфику испытаний объектов класса ИС. Автором проанализировано четырнадцать таких документов[3], из которых можно вычленить лишь следующие общие теоретические положения.

1. Для испытаний необходима разработка сценариев реальных ситуаций, в том числе релевантных, критических и сложных, на основе которых нужно проводить валидацию аналитической модели, построенной на основе машинного обучения, и которые должны включать в себя множество элементов, существенных для их описания.

2. Испытания необходимо проводить как виртуальные, с созданием необходимых сред симуляции и компьютерным моделированием процессов функционирования систем с ИИ с различными виртуальными сценариями, так и полигонные, с проверкой возможностей и функционирования систем ИИ в реальных условиях.

3. При испытаниях и построении систем оценки целесообразно активно использовать знания экспертов.

4. Необходимо строить системы мониторинга функционирования систем ИИ после начала эксплуатации, так как какая бы оценка качества системы ни проводилась, фактический уровень качества может быть подтвержден только после эксплуатации системы в «…достаточном диапазоне условий …и параметров окружающей среды».

5. В качестве целей испытаний должно быть не просто подтверждение заявленных характеристик, а оценка способности системы с ИИ «…справиться с чрезвычайными происшествиями, с которыми можно столкнуться при эксплуатации...».

6. Ни один документ не содержит общих подходов к испытаниям систем с ИИ; более того, для задач машинного обучения, например, большинство документов описывают терминологию и методы создания обучающих выборок для нейросетей, полностью игнорируя другие потенциально возможные методы испытаний.

Вместе с тем хорошо известны многочисленные недостатки систем машинного обучения: для получения точных моделей требуются большие данные, моделям машинного обучения свойственна непрозрачность и отсутствует необходимая особенно для военно-технических задач аргументация, требование значительного времени и человеческих ресурсов для подготовки исходных данных (сложность автоматической разметки данных), у алгоритмов машинного обучения нет универсальности и они специфичны для определенных классов задач и др.

---

[1] ГОСТ 16504-81 Система государственных испытаний продукции. Испытания и контроль качества продукции. Основные термины и определения. М.: Стандартинформ, 2011. 23 с.

[2] ГОСТ Р 59276-2020 Системы искусственного интеллекта. Способы обеспечения доверия. Общие положения. М.: Стандартинформ, 2021. 12 с.

[3] Перечень утвержденных стандартов ТК 164. URL: http://tc164.ru/standarts123456

Следует подчеркнуть, что зачастую совершенно неправильно процесс испытаний отождествляют с тестированием [2]. В основе такой ошибки неоднозначный перевод и некорректное использование термина «*testing*». Применительно к машинному обучению под тестированием понимают проверку работы программного обеспечения, реализующего алгоритмы машинного обучения при всех реальных вариациях исходных данных и поиск ситуаций, при которых программа ведет себя неправильно. Поэтому совсем запутанно выглядит часто используемое положение «приемочные испытания в процессе тестирования». Испытания же, в частности ИС, проводятся с целью определения соответствия характеристик свойств систем (образцов) заявленным характеристикам, и к ним более применим английский термин «*experience*».

Возникает законный вопрос: какие из свойств при создании перспективного образца ИС обеспечиваются за счет использования алгоритмов машинного обучения?

В работе [3] проанализированы боевые свойства перспективных ИС, в которых могут использоваться технологии искусственного интеллекта, и предложено ввести в научный оборот новое понятие «интеллектуальные боевые свойства». Основными интегральными свойствами ИС, обеспечивающими их интеллектуальность, предложено считать «автономность» функционирования ИС, его «адаптивность» к текущей боевой ситуации и «безопасность боевого применения». Но если для таких боевых свойств как «огневая мощь», «надежность», «живучесть» и др. методы испытаний существуют и отработаны, то для новых боевых свойств они еще не сформированы.

Таким образом, актуальна разработка основных положений (принципов) осуществления испытаний ИС, обученных в парадигме машинного обучения и направленных на разработку программ и методик испытаний для обеспечения качества решения функциональных задач в различных условиях боевого применения.

2 О роли тестов при проведении испытаний

Существуют специализированные тесты, которые позволяют определить, является ли испытываемая система интеллектуальной и обладает ли она искусственным интеллектом. Это такие испытания, как тест «Тьюринга» (беседа с системой ИИ), тест А.Лаврейс (способность создавать произведения искусства в определенном жанре), тест Ч.Ортица (создание, объяснение и описание процесса разработки вербальной конструкции) тест М.Барклая (проверка, может ли ИИ видеть и способен ли описать свое отношение к другим объектам), тест Левеска (на понимание скрытых смыслов), тест Маркуса (показ системе ИИ видео и определение им временного промежутка, где необходимо смеяться) и др.

Очевидно, что непосредственное применение таких тестов применительно к испытаниям образцов ИС затруднительно.

Наиболее интересным для решения рассматриваемых в работе задач является так называемый «Обратный тест Тьюринга», при котором не человек узнает робота, а робот распознает человека. Указанное имеет прямое отношение к испытаниям ИС, поскольку типовой ситуацией, в частности для полуавтономных и автономных образцов ИС, станет попадание людей в область ответственности систем технического зрения и в этих условиях одной из главных задач станет распознавание объектов и субъектов, несущих угрозу. Пока решение таких задач основывается на использовании машинного обучения для распознавания агрессивных действий [4].

Ниже сформулировано семь основных принципов испытаний ИС, системы управления которых созданы с реализацией технологии машинного обучения. Принципы иллюстрированы примерами для использования при разработке соответствующих программ и методик испытаний ИС.

3 Принцип недопустимости негативных последствий принятия решений

Одно из важных требований к перспективным ИС, в системах управления которых реализуются решения по использованию технологий искусственного интеллекта, связано с обеспечением свойства «безопасность боевого применения». Оно трактуется как способность: а) функционировать в условиях угроз и воздействий и б) не создавать угроз при эксплуатации.

Для реализации этих требований необходимо сформулировать утверждение со статусом аксиомы, что обученный образец ИС может совершать ошибки. Их возможность не должна быть высокой, но должно быть принципиально допустимо ее отличие от нуля. Поэтому с учетом проверки на испытаниях свойства «безопасность боевого применения» необходимо сформулировать принцип: если, например, боевой робот с ИИ совершит ошибку, это не должно привести к негативным последствиям.

Каждый из терминов этого постулата можно и необходимо конкретизировать.

Понятие «Ошибка» целесообразно тесно увязать с функциональностью образца. Ключевыми ошибками являются ошибки информационного обеспечения, организации движения, управления огнем, обучения и самообучения, самодиагностики, в конечном итоге – принятие частично ошибочного решения и, как следствие, невыполнение боевой задачи. Из перечня этих ошибок следует, что одиночная ИС не может наносить ущерб бездействием, но только действием- движением или огнем.

Понятие «Негативные последствия» необходимо связать с ущербом, наносимым: а) среде; б) собственным техническим средствам; в) личному составу. В зависимости от функциональности ИС ущерб среде можно измерять в уровне критичности такого ущерба и возможности ее последующего восстановления, технике и личному составу – количественными или качественными характеристиками потерь, включая важность всех их составляющих.

Анализ показывает, что наиболее важным здесь является способность обученного образца ИС различать, условно говоря, «своих» и «чужих». Поэтому для исключения опасности открытия «неправильного огня» система управления огнем ИС должна снабжаться алгоритмом типа «свой-чужой». Для оценки опасности перемещения ИС в сторону собственных войск необходимо все время контролировать расстояние между ИС и предметами (объектами), которым он может нанести вред, и предусматривать прекращение движения.

Отсюда следует вывод, что при реализации алгоритма машинного обучения необходимо так построить обучающие и тестовые выборки, чтобы обеспечить надежное разделение указанных выше классов. Такие требования должны предъявляться на этапе технического проектирования образца; более того, требования к показателям качества классификации для класса «свои» необходимо задавать более жесткие.

Пример. В процессе машинного обучения обнаружению и распознаванию образов в состав обучающей выборки включаются объекты с признаком «свои». В состав системы принятия решений закладывается правило «Огневое воздействие по своим запрещено». На испытания представляются объекты как «свои», так и противника. В процессе испытаний ИС распознает и идентифицирует все объекты, но осуществляет огневое воздействие только по объектам противника.

4 Принцип «разрешенной» неточности

Основные требования к ИС всегда предусматривают максимально высокие характеристики выполнения боевых задач – точности, оперативности, надежности, достоверности и др. Системы искусственного интеллекта влияют на различные показатели боевой эффективности, но чаще всего на точность показателей.

Для реализации требований обеспечения высокой точности показателей необходимо сформулировать утверждение со статусом аксиомы, что никогда не удастся испытать образец ИС во всех возможных вариантах его боевого применения. Поэтому качество решения задач и обеспечения требуемых показателей эффективности никогда не достигнет 100%. Из этого

следует, что качество можно определить только «в целом» или «в среднем», если удастся собрать требуемую статистику, что можно сделать только в процессе эксплуатации.

Проверяемые критерии точности, в частности, можно описывать такими параметрами как точность обнаружения, распознавания, идентификации, поражения цели и др. [5] с наиболее употребимым показателем «вероятность». В связи с этим необходимо вводить диапазон, в котором будут разрешаться допустимые погрешности таких вероятностей.

Аналогичные диапазоны целесообразно вводить для показателей критериев качества решения различных задач (погрешности движения по местности, оперативность выполнения различных функций и действий и др.), проверяемых критериев взаимодействия в группах (процессы обмена информацией и др.), в которых реализуются алгоритмы машинного обучения.

На испытаниях должно проверяться нахождение указанных показателей в допустимых диапазонах.

Пример. В процессе испытаний в тестовую выборку введены данные, полученные методом аугментации со смещенной оценкой таким образом, что образы реальных и отравленных данных существенно похожи. Таким образом, обучающая выборка становится смещенной, то есть часть совокупности данных по сравнению с другими имеют более низкую или более высокую вероятность выборки. Обучение происходит на смещенной выборке. Если алгоритмы обучения не предусматривают учет смещенных выборок, качество решения задач будет в определенной мере ухудшаться, что потребует объяснений и переобучения.

5 Принцип «неожиданности»

Как известно, машинное обучение относится к непрозрачным технологиям, в которых невозможно получить объяснение процесса работы обученных моделей и результатов.

Поэтому необходимо сформулировать утверждение со статусом аксиомы о том, что на испытаниях обученную модель машинного обучения можно и нужно заставить ошибаться. Именно поэтому испытания обученных методом машинного обучения систем управления ИС должны проводиться на тестовых данных, существенно отличающихся от обучающих выборок не только по объему, но и по смыслу. «Отравление» тестовой выборки призвано спровоцировать ошибки обученной модели и выявление ее наиболее узких мест. Разработчик же должен предусматривать очистку обучающей выборки от отравляющих атак, использование нескольких систем классификаторов для повышения надежности решения и др.

Для решения указанных выше задач испытаний необходимо сформировать комплекс так называемых состязательных атак в интересах выдачи обученной моделью неправильных результатов для подтверждения уязвимости алгоритмов машинного обучения. То, что разработчик моделей обязан проинформировать Заказчика об архитектуре модели и параметрах обучения, существенно упрощает подготовку, например, так называемых White Box атак. Целесообразно также подготовить и Targeted-атаки, направленные на неправильную классификацию изображений определенного класса, и Untargeted-атак, чтобы просто заставить модель машинного обучения ошибаться [6].

Пример. Для дискредитации алгоритма распознавания, построенного с использованием машинного обучения, используется так называемая целевая атака (targeted attack) на основе специально обученного алгоритма типа One pixel attack из класса алгоритмов атак BlackBox с подменой специально выбираемых одного или нескольких пикселей, в основу которого закладывается решение задачи оптимизации в интересах минимизации доверия к заданному классу (например к классу «свои»). Если на испытаниях ИС показатели качества будут неудовлетворительны (например, превышают допустимые), то указанное является информацией о том, что атака оказалась успешной.

6 Принцип «объяснимости» решения

Традиционно в испытаниях проверяется выполнение требований, заданных в тактико-технических требованиях на образец ИС, что, как правило, связано с его функциональностью.

Как показывает практика, разрабатываемые интеллектуальные решения с использованием машинного обучения настолько сложны и объективно непрозрачны, что проверить правильность получения решения проблематично, если вообще возможно [7].

В такой ситуации необходимо сформулировать утверждение со статусом аксиомы о том, что «необходимо получать и анализировать не только результаты испытаний, но также и то, как и почему они получены именно такими».

Поэтому разработанные алгоритмы машинного обучения должны содержать в себе встроенные функции аналитической экспертизы, учитывающие разнообразные сценарии, уникальные с точки зрения обучения и объясняющие в допустимой степени как получаемые результаты, так и возможную их погрешность при последующей эксплуатации.

Автоматически такие функции построить невозможно. Единственным на данный момент вариантом экспертизы видится а) представление и последующая экспертиза Справки разработчика об архитектуре и тестировании модели, результатах ее валидации, степени независимости тестовой выборки и др. и б) накопление определенной статистики по получению различных показателей качества модели машинного обучения, ее визуализация и соответствующее экспертное заключение.

Пример. В силу отсутствия в настоящее время универсальных решений по реализации технологий объяснимого искусственного интеллекта и принципиальной непрозрачности алгоритмов машинного обучения в состав комиссии по государственным испытаниям необходимо включение группы экспертов, вырабатывающих консолидированное экспертное мнение о правильности принимаемых образцом ИС решений.

7 Принцип «от обратного»

Многообразие возможных вариантов боевого применения интеллектуальных образцов ИС столь велико, что возможность полной проверки правильности функционирования существенно ограничена.

В таких условиях необходимо сформулировать утверждение со статусом аксиомы о том, что «испытания должны проверять не то, что образец ИС должен делать, а то, что делать не должен».

Количество ситуаций, описывающих то, что образец не должен делать, принципиально счетно. Анализ показывает, что их число не превышает первых десятков. Поэтому необходимо систематизировать такие недопустимые ситуации и по каждой из них разработать алгоритм проверки действий образца ИС. Образец ИС должен продемонстрировать отказ от выполнения недопустимых действий.

Необходимо все же учитывать, что динамика и условия реальных боевых действий таковы, что даже сформировав реалистичные сценарии работы с системой на основе статистики её использования, новые сценарии на практике сформируются с ненулевой возможностью. В этом же контексте можно рассматривать и создание сложных условий для испытаний. Так, обученный с использованием технологий машинного обучения беспилотный летательный аппарат должен запускаться и решать задачи на сложном участке местности (здания, деревья и др.), возвратиться в исходную точку запуска при сложных погодных условиях, отказаться от решения задач на недопустимых высотах движения, функционировать при сниженной мощности сигнала GPS и др.

При таком подходе в программу испытаний могут быть заложены инновационные решения по выполнению тех операций, которые ИС не должен совершать, например, подать команду двигаться на недопустимый объект, движение на который запрещено, открыть огонь по «своим», поднять БПЛА на недопустимую для эксплуатации высоту и др. Образец

ИС на испытаниях не должен выполнять эти команды, а прекратить функционирование с соответствующим уведомлением.

Пример. К числу недопустимых действий для беспилотных летательных аппаратов является значительное отклонение от запланированного маршрута и невозврат в исходную точку запуска по своему следу. В процессе испытаний создаются условия продолжительной потери связи с БПЛА и контролируются его перемещения. На основании этого делается вывод о работоспособности образца.

8 Принцип «приоритета ошибок первого рода»

Исходя из допустимости ошибок ИС, обученного в парадигме машинного обучения, целесообразно разделить критичные для выполнения задач и некритичные ошибки. В таких условиях необходимо сформулировать утверждение со статусом аксиомы о том, что «пропуск важного события хуже, чем ложная тревога».

С точки зрения системы управления движением объект испытаний не должен сталкиваться с препятствиями, перемещаться в направлении, выходящем за границы ограничений в пространстве и др. Так, столкновение с препятствием является ошибкой первого рода. Но если наземный образец ИС будет объезжать несуществующие препятствия, то такие ошибки на испытаниях следует признавать допустимыми.

С точки зрения системы управления огнем объект испытаний не должен совершать ошибки первого рода (в процессе распознавания и идентификации принимать правильное за неправильное, то есть пропускать угрозу, например, обнаруживать, но неправильно распознавать объект, идентифицировать и не оказывать воздействие на объект противника). Ошибки же второго рода, например, огневое воздействие на несуществующий объект противника, следует признать допустимыми, однако при условии, что такой показатель как точность (отношение истинно положительных результатов к сумме истинных и ложных) для ошибок второго рода будет незначимым.

Таким образом, совершение ошибок второго рода не следует считать определяющим в принятии решения о качестве образца ИС, если только их количество незначительно.

Пример. В процессе испытаний образец ИС, обученный методом машинного обучения распознаванию бронетанковой техники противника, обнаружил надувной имитационный танк, принял в процессе распознавания и идентификации его за реальный объект и неправильно среагировал на несуществующую угрозу, подавив его огнем. Такой результат испытаний следует признать допустимым.

9 Принцип «злонамеренности»

Наиболее реальной ситуацией при функционировании обученного образца ИС является получение дополнительной информации в процессе боевой эксплуатации. Это требует переобучения модели. В боевых условиях весьма вероятно воздействие противника на программно-технические средства ИС и целенаправленное представление ложной дезинформации [8]. Поэтому в процессе испытаний целесообразно исследовать реакцию модели машинного обучения на недостоверные данные – неточные, некорректные, дублирующие, недоопределенные, испорченные и др. Такие искаженные данные с низкой достоверностью в процессе испытаний необходимо представлять модели как надежные данные с высокой степенью достоверности. Алгоритмы же машинного обучения должны уметь отличать намеренно введенные вредоносные данные от реальных аномальных событий.

В таких условиях необходимо сформулировать утверждение со статусом аксиомы о том, что «намеренное внесение в обучающие выборки злонамеренной информации и последующее дообучение образца ИС способно ухудшить эффективность и качество функционирования образца ИС. Поэтому на испытаниях целесообразно проведение экспериментов по ситуации, требующей дообучения, намеренному внесению в обучающие выборки искаженной информации и последующему контролю действий образца ИС. Искаженная информация

не должна влиять на качество распознавания. Качество результатов функционирования модели машинного обучения после дообучения не должно снижаться. Система должна постоянно обучаться, анализировать ошибки и приспосабливаться к новой информации.

Пример. На испытания вынесен вариант с имитацией обнаружения испытываемым образцом ИС новых объектов, отсутствующих в обучающей выборке, фиксацией необходимости дообучения и введением дополнительной, но в реальности неправильной информации. При этом процесс переобучения предлагается осуществить на несбалансированных классах. После дообучения результаты работы модели распознавания должны оставаться удовлетворительными.

10 Выводы

Сложность вопросов организации и проведения испытаний образцов ИС, обученных в парадигме машинного обучения, пока еще даже не осознана. Количество проблемных вопросов, по каждому из которых целесообразна разработка соответствующих как нормативных, так и технических документов, очень велико.

В литературе многократно подчеркивается, что модели машинного обучения генерируются автоматически, нелинейны и слишком сложны. Традиционные методы испытаний неэффективны. Поэтому необходимы испытания не продукта целиком, а последовательно по этапам его разработки: разработка общих тактико-технических требований, доказательство концепции, проектирование, разработка, тестирование и эксплуатация. Однако сейчас нет рациональных руководств по оценке каждого вида деятельности. В ведущих компаниях мира методы обеспечения качества моделей машинного обучения являются разнообразными экспериментальными практиками. Современные тенденции включают: а) разработку методов градиентного поиска, эффективно обнаруживающих множество неправильных вариантов поведения; б) оценку надежности и поиск состязательных репрезентативных примеров; в) специальные алгоритмы тестирования; г) методы объяснимого искусственного интеллекта, включая локальное объяснение для каждого результата с выявлением ключевых точек, ответственных за качество в данной ситуации; д) специальные методы дискретной оптимизации, по сути параллельную машинному обучению модель для построения списков правил в пространстве признаков и др.

Основываясь на квалифицированном мнении неофициального консорциума QA4AI – японской группы экспертов по обсуждению обеспечения качества систем искусственного интеллекта на основе машинного обучения [9], выделим следующие три (из пяти обсуждаемых) групп основных направлений исследований по оценке качества систем искусственного интеллекта, основанных на машинном обучении, применительно к испытаниям обученных образцов ИС.

1. Целостность данных. На испытаниях необходимо проверять одиннадцать ключевых аспектов, в том числе работу с конфиденциальной информацией, качество данных и генераторов данных для обучающей выборки, отношения между генеральной совокупностью и выборкой, наличие систематических ошибок, независимость и влияние мультиколлениарности данных, исключение выбросов, решения по отсутствующим данным и др.

2. Надежность модели машинного обучения. Здесь также одиннадцать направлений испытаний модели, включая ее производительность, архитектуру, выделение и анализ гиперпараметров, проверку достижимости локальных и глобальных оптимумов, возможность обработки разнообразных данных, использование при обучении перекрестных проверок (кросс-валидация), качество обучения в условиях шума датасетов и др.

3. Качество модели машинного обучения. Выделяются восемь групп проверок: общее качество модели, использованные методы валидации, критичность и частота ошибок, управляемость, безопасность функционирования, вклад и локализуемость компонентов моделей машинного обучения, объяснимость результатов, надежность и др.

Отдельно подчеркивается, что модели машинного обучения подвержены изменениям в процессе эксплуатации, прежде всего за счет погрешностей при их создании, используемых методов, заточенных на конкретные приложения, окружающей среды и др.

Таким образом, вопросы испытаний ИС требуют серьезных исследований.

Заключение

Вопросы проведения испытаний образцов ИС, при создании которых используются технологии машинного обучения, в настоящее время требуют разработки принципов проведения таких испытаний. Но сами технологии искусственного интеллекта обладают свойствами, приводящими к рискам качества результирующих моделей – непрозрачностью решений, возможностью получения некорректных результатов и др.

Для повышения качества проверки тактико-технических характеристик ИС на испытаниях и повышения эффективности последующей боевой эксплуатации необходимо формировать максимально сложные условия для объекта испытаний.

Предложены семь принципов испытаний образцов ИС, при разработке которых использовались технологии машинного обучения.

1. Принцип недопустимости негативных последствий принятия решений, допускающий ошибки в процессе функционирования, которые, однако, не должны приводить к ущербу среде, личному составу и вооружению.

2. Принцип «разрешенной» неточности, вызванный принципиальной невозможностью испытаний ИС во всех условиях боевого применения, требующий введения допустимых диапазонов критериев качества решения различных задач и контроля нахождения показателей в этих пределах.

3. Принцип «неожиданности», направленный на формирование комплекса состязательных атак как на данные для обучения, так и на созданную модель машинного обучения для выявления ошибок в исходной информации и архитектуре модели.

4. Принцип «объяснимости» решения, выдвигающий требование наличия встроенных функций аналитической экспертизы для подтверждения не только правильности результата, но также и объяснение процесса его получения.

5. Принцип «от обратного», направленный на проверку в процессе испытаний не только требуемого по тактико-техническому заданию функционала ИС, а тех функций, которые образец ИС выполнять не должен и которые выполнять запрещено.

6. Принцип «приоритета ошибок первого рода», допускающий в процессе эксплуатации незначительное количество ошибок второго рода и не рассматривающий такие ошибки существенными для принятия решения о качестве модели машинного обучения.

7. Принцип «зловредности», заключающийся в намеренном внесении в обучающие выборки злонамеренной информации, способной ухудшить эффективность и качество функционирования образца ИС для проверки способности модели машинного обучения противостоять недостоверной информации.

Автор отдает себе отчет в уязвимости предлагаемых принципов для критики в силу их новизны и пока еще слабой теоретической проработанности. Но такая полемика необходима и будет полезна для корректировки предлагаемых принципов в интересах их практического применения при разработке программ и методик соответствующих испытаний.

Список источников

1. Băjenescu T.I. The risks of artificial intelligence // Journal of Engineering Science. 2018. Vol.XXV, No.4. P. 47-56.

2. Feldt R., Oliveira Neto F.G., Torkar R. Ways of Applying Artificial Intelligence in Software Engineering // Realizing Artificial Intelligence Synergies in Software Engineering (RAISE): IEEE/ACM 6th International Workshop. Gothenburg: IEEE, 2018. P. 35-41.

3. Абросимов В.К., Гладкий А.В. Интеллектуальность боевых свойств перспективных робототехнических комплексов наземного базирования // Известия РАРАН. 2024. №1(131). С. 109-115.

4. Уздяев М.Ю. Распознавание агрессивных действий с использованием нейросетевых архитектур 3D-CNN // Известия Тульского государственного университета. Технические науки. 2020. №.2. С. 316-330.

5. Макаренко С.И. Противодействие беспилотным летательным аппаратам: монография. СПб.: Наукоёмкие технологии, 2020. 204 с.

6. Намиот Д.Е. Введение в атаки отравлением на модели машинного обучения // International Journal of Open Information Technologies. 2023. Т.11. №.3. С. 58-68.

7. Оселедец И.В. Успехи и проблемы машинного обучения // Проектирование будущего. Проблемы цифровой реальности. 2022. №1(5). С. 102-108.

8. Зарудницкий В.Б. Характер и содержание военных конфликтов в современных условиях и обозримой перспективе // Военная мысль. 2021. №1. С. 34-44.

9. Hamada K., Ishikawa F., Masuda S., Myojin T., Nishi Y., Ogawa H., Toku T., Tokumoto S., Tsuchiya K., Ujita Y., Matsuya M. Guidelines for Quality Assurance of Machine Learning-based Artificial Intelligence // Proceedings of the 32nd International Conference on Software Engineering and Knowledge Engineering (USA, July 9-19, 2020). DOI: 10.18293/SEKE2020-094.

Информация об авторе

В.К. Абросимов – доктор технических наук, старший научный сотрудник.