

А.Е. Самотуга, П.С. Ложников

**ФУНКЦИОНАЛЬНАЯ БЕЗОПАСНОСТЬ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА**

## Понятие функциональной безопасности ИИ, управление безопасностью ИИ

Под функциональной безопасностью стоит понимать часть общей безопасности, относящейся к контролируемому оборудованию. Данное направление тесно связано с понятием функция безопасности, которая призвана контролировать риски безопасности, связанные с опасностями причинения вреда людям или окружающей среде.

Функциональная безопасность направлена на достижение безопасного состояния посредством обеспечения функций безопасности.

Функциональная безопасность связана с надлежащим проектированием технических и нетехнических функций безопасности для снижения риска безопасности или сдерживания уровня риска безопасности конкретного управляемого оборудования, от уровня компонентов до системного уровня. В том числе при наличии человеческой ошибки и эксплуатационных или экологических стрессов.

Функциональная безопасность фокусируется на технических функциях для снижения рисков безопасности и необходимых атрибутах этих функций, т.е. должны отсутствовать критические с точки зрения безопасности отказы выше некоторого определенного порога.

До появления программируемых систем функциональная безопасность ограничивалась исключительно аппаратной реализацией, которое включало два основных направления - это снижение вероятности отказов оборудования и снижение их последствий.

С программным обеспечением, используемым для реализации функций безопасности, акцент смещается на систематические сбои, допущенные во время разработки.

## Свойства и связанные с ними факторы риска безопасности технологий ИИ

Продукты, содержащие технологии ИИ, особенно технологии машинного обучения, следуют другой парадигме разработки. Они в меньшей степени зависят от спецификаций и больше зависят от наблюдения за данными, определяющими поведение системы.

Функциональная безопасность должна учитывать дополнительные аспекты, помимо случайных аппаратных и систематических сбоев при разработке.

Подходы, ориентированные на оценку рисков в других областях деятельности чаще всего являются статическими концепциями. Очевидно, что для систем ИИ с непрерывным обучением или переконфигурированием во время использования (к примеру, производственные линии, которые динамически перестраиваются для производства товаров разной номенклатуры). Для систем искусственного интеллекта нужны подходы, адаптированные под их особенности.

Далее приведем описание некоторых технологий искусственного интеллекта, их свойств и связанных с ними факторов риска безопасности.

На технологическом уровне возможности ИИ часто достигаются за счет комбинации алгоритмов и моделей. Модель часто представляет знания, которые достигают цели приложения. К примеру, знания о том, как различать и распознавать различные входные данные. Алгоритмы используются для получения дополнительной информации из знаний и входных данных, например, прогнозирование. Примеры моделей включают линейные функции, логические исчисления, динамические байесовские сети и искусственные нейронные сети. Модели могут быть созданы инженером вручную или синтезированы из данных с помощью алгоритмов машинного обучения, которые сами используют процесс систематического анализа. Алгоритмы и модели должны быть преобразованы в исполняемое представление, такое как машинный код (в случае программного обеспечения) или специальное оборудование, такое как программируемая пользователем вентильная матрица (FPGAs). Из чего следует факт того, что функциональная безопасность приложений, использующих технологию ИИ, зависит как от моделей, так и от алгоритмов.

Во многих случаях алгоритмы, взаимодействующие с моделями, содержат лишь ограниченный объем знаний или следствий о целях приложения. То есть сам алгоритм не нацелен на достижение безопасности, но корректность его работы критически важна для достижения функциональной безопасности. Таким образом, целостность алгоритмов в технологии искусственного интеллекта часто может быть обеспечена с помощью существующих принципов функциональной безопасности, аналогично общим программным компонентам. То же самое относится и к логике, связанной с переводом алгоритма и модели.

Модели часто содержат знания, связанные с целью систем, связанных с функциональной безопасностью. Существует несколько различных способов построения моделей, и могут использоваться различные подходы для оценки мероприятий по снижению риска.

Если модели разрабатываются инженерами вручную, они, скорее всего, будут отражать знания инженеров о приложении. Эти знания можно оценить в ходе процессов управления, используемых в рамках жизненных циклов функциональной безопасности. В этих случаях можно соблюдать жизненный цикл существующих стандартов функциональной безопасности.

Модели, полученные из данных с помощью алгоритмов машинного обучения, могут быть проанализированы и проверены после их создания. В качестве альтернативы можно анализировать модели, полученные с помощью алгоритмов машинного обучения, извлекать основные параметры и использовать их для расширения общих инженерных знаний, которые, в свою очередь, можно использовать для разработки новых моделей. В этом случае, при использовании проверенных инженерных знаний к новым моделям также может быть применен жизненный цикл на основе существующих стандартов функциональной безопасности.

В других случаях модели слишком сложны для понимания и следует выбрать другой подход к оценке риска для безопасности.

1. **Трёхкомпонентная структура системы ИИ. Технологические элементы для создания и исполнения модели ИИ. Уровень автоматизации и контроля ИИ**

Систему ИИ можно представить (см. рис. 1) структурой, состоящей из трех основных компонентов:

* получение данных (задача ввода);
* аккумуляция знаний из данных и знаний людей (задача обучения);
* обработка и генерация выходных данных (задача обработки).

Здесь под знаниями людей будем понимать знания из самых разных областей, в частности в области применения ИИ и области самого ИИ.

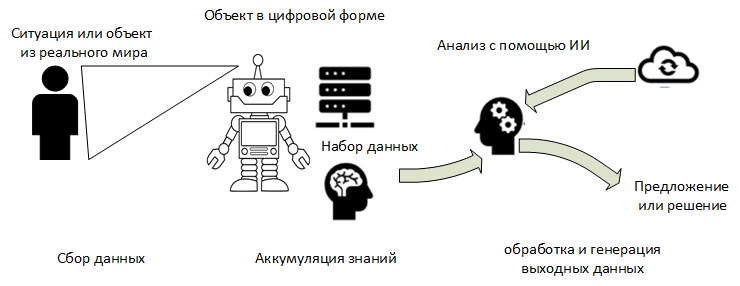


Рис. 1. Три компонента ИИ

Для данного трехкомпонентного рабочего процесса для получения критериев соответствия можно представить следующий набор процессов (см. рис. 2):

Желательные свойства определяются для каждого из трех компонентов;

Свойства связаны с тематиками и с подробными методами и приемами, касающимися этих тем;

Критерии соответствия выявляются из набора детализированных методов и приемов.

Свойства могут быть определены в каждом конкретном случае или получены из свойств, перечисленных в существующих стандартах.

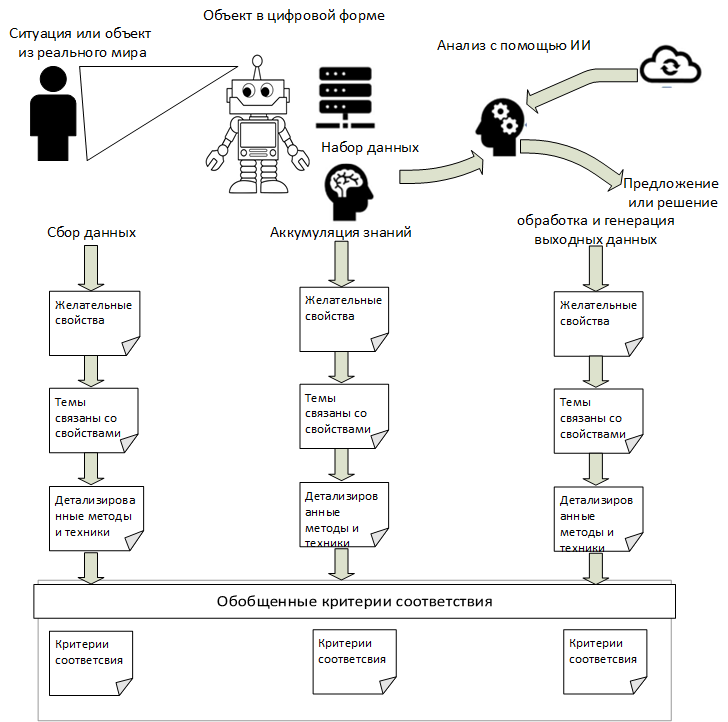


Рис. 2. Процессы для каждой из стадий

Создание и выполнение модели включает в себя различные технологические элементы. В таблице 1 перечислены эти элементы на основе функциональных уровней экосистемы ИИ. В таблице 2 перечислены технологические элементы, которые, например, можно использовать при создании и запуска нейронной сети.

Таблица 1. Примеры технологических элементов

|  |
| --- |
| Технологический элемент |
| Сервисы ИИ |
| Машинное обучение (разработка и использование модели; утилиты; данные) |
| Проектирование (разработка и использование модели; утилиты) |
| Облачные и периферийные вычисления, а также большие данные и источники данных |
| Пул ресурсов — вычисления, хранилище, сеть |
| Управление ресурсами – предоставление ресурсов |

Таблица 2. Примеры технологических элементов, используемых при создании модели и запуска в машинном обучении

|  |  |
| --- | --- |
| Технологический элемент(пример машинного обучения) | Пример языка/инструмента |
| Граф приложения | Граф GXF в YAML, граф rqt в ROS |
| Фреймворк машинного обучения | TensorFlow, PyTorch, Keras |
| Модель машинного обучения | ONNX, NNEF |
| Компилятор графов машинного обучения | TensorRT, GLOW |
| Комплекс расчетов | C++ |
| Библиотеки вычислительных операций | CUDA / C++ |
| Компилятор | gcc / nvcc |
| Исполняемый машинный код | aarch64, PTX |

Некоторые технологические элементы можно рассматривать с помощью существующих концепций функциональной безопасности. Например, обычно можно обрабатывать программное обеспечение, переводящее модель в исполняемое представление с существующими концепции функциональной безопасности

Как показано на рис. 3, элементы, содержащие технологию ИИ, используются на разных уровнях системы/приложения: элементы более высокого уровня (например, граф приложения) чаще всего зависят от варианта использования, поэтому соответствующие функции и данные зависят от приложения.

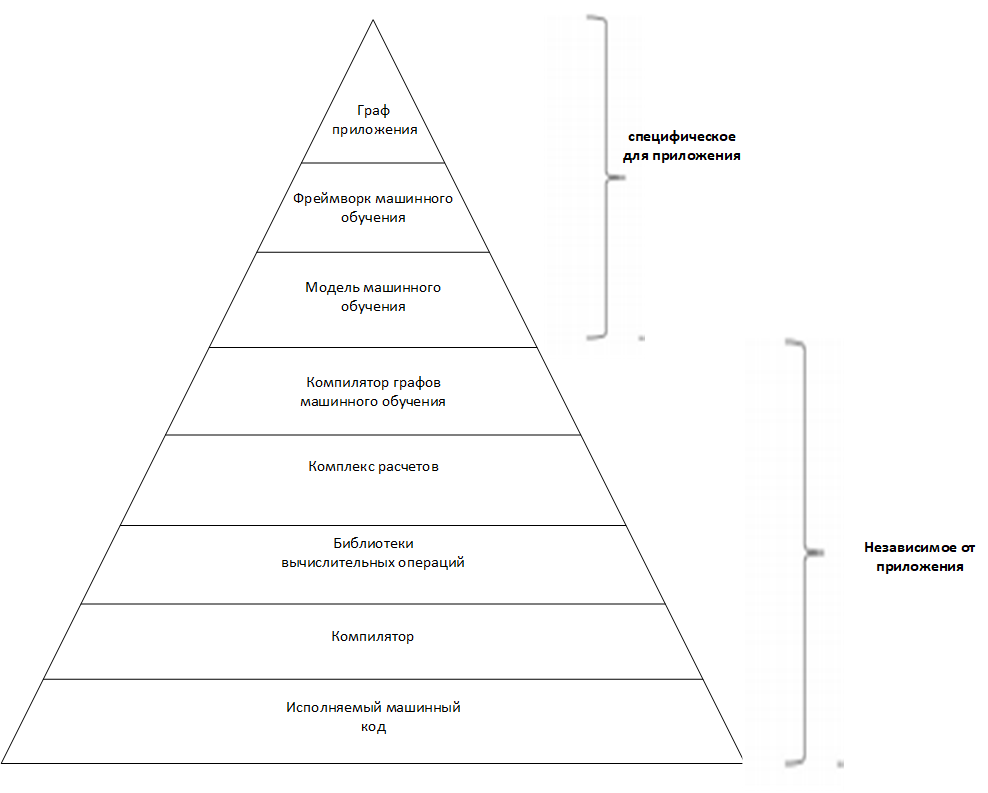


Рис. 3. Иерархия технологических элементов

Уровень автоматизации (иногда называемый в литературе «уровнями автономии») описывает степень, в которой система ИИ функционирует независимо от наблюдения и контроля со стороны человека. Таким образом, он определяет не только объем информации о поведении системы, доступный оператору, но также определяет возможности управления и вмешательства человека.

Измерения этой темы включают в себя то, насколько высоким должен быть уровень автоматизации для соответствующего приложения, а также насколько ограничены возможности управления пользователя. Системы с высоким уровнем автоматизации могут демонстрировать неожиданное поведение, которое трудно обнаружить и контролировать. Таким образом, высокоавтоматизированные системы могут представлять угрозу безопасности с точки зрения их надежности и безопасности.

При рассмотрении вопроса о том, достигается ли функциональная безопасность, важны несколько аспектов, таких как быстродействие системы ИИ и наличие или отсутствие «контроллера». В этом контексте «руководитель» служит для проверки или утверждения автоматизированных решений системы.

Такой «контроллер» может быть реализован функциями технического контроля. Например, можно добавить вторую инструментальную систему безопасности для критически важных элементов управления, которая будет назначена системе безопасности. Другой способ добавить «контроллера»— использовать человека, задача которого — вмешиваться в критические ситуации или подтверждать решения системы. Однако, даже если люди находятся в контуре и контролируют действия системы, это не приведет к автоматическому снижению таких рисков безопасности до приемлемого уровня и может привести к дополнительным рискам безопасности.

**Список источников**

1. ISO/IEC AWI TR 5469 [Электронный ресурс]. URL: https://www.iso.org/cms/render/live/en/sites/isoorg/contents/data/standard/08/12/81283.html (дата обращения: 14.04.2022).