

Искусственный интеллект и защита информации

En Artificial Intelligence and Information Security

V. A. Konyavskiy,
PhD (Eng., Grand Doctor), Head
of the Department¹, Scientific Director²
konyavskiy@gospochta.ru

S. V. Konyavskaya-Schastnaya,
PhD (Philology), Associate Professor¹,
Deputy Director²
cd@okbsapr.ru

GigaChat,
Russian-language AI by Sber
<https://giga.chat>

¹ MIPT

² OKB SAPR

The article is devoted to the widely discussed today problem of regulating artificial intelligence (AI). To solve this problem, it is proposed to first highlight, identify and classify the subject of regulation: AI systems (as opposed to algorithms, models and other entities related to AI in one way or another). The reason for regulating exactly AI systems is that a person interacts exactly with AI systems. As nothing more than AI systems directly influence a person's life. The purpose of regulation is to control the influence of AI on human life. Productive classification can be built on the basis of the strength of influence on human life, that depends on the degree of autonomy of AI systems. Something that can have a negative impact on human life and health obviously should be regulated more strictly than something that cannot. To elicit factors that are significant for identifying AI, the opinion of the GigaChat AI by Sber was taken into account.

Keywords: AI, AI system, trusted AI, regulation of AI, knowledge, abilities, skills, consciousness, models

УДК 004.89

Статья посвящена широко обсуждаемой сегодня проблеме регулирования искусственного интеллекта (ИИ). Для решения этой проблемы предлагается в первую очередь выделить, идентифицировать и классифицировать предмет регулирования: системы искусственного интеллекта (в отличие от алгоритмов, моделей и иных так или иначе связанных с ИИ сущностей). Основанием для регулирования систем ИИ (СИИ) является то, что человек взаимодействует именно с СИИ, и, стало быть, именно они могут оказывать непосредственное влияние на жизнь человека. Классификация, продуктивная для регулирования, целью которого является контроль над влиянием СИИ на жизнь человека, может быть построена на основании степени автономности СИИ разных классов и связанной с этим степени их влияния на жизнь человека: те, что могут оказать негативное влияние на жизнь и здоровье человека, очевидно, должны регулироваться строже, чем те что не могут. При выделении значимых для идентификации ИИ факторов учтено мнение СИИ GigaChat от Сбера.

Ключевые слова: ИИ, СИИ, доверенный ИИ, регулирование ИИ, знания, умения, навыки, сознание, модели

Валерий Аркадьевич Конявский,
доктор технических наук, заведующий
кафедрой¹, научный консультант²
konyavskiy@gospochta.ru

**Светлана Валерьевна
Конявская-Счастливая,**
кандидат филологических наук, доцент¹,
заместитель генерального директора²
cd@okbsapr.ru

GigaChat,
русскоязычная нейросеть от Сбера
<https://giga.chat/>

¹ Московский физико-технический институт

² АО «ОКБ САПР»

Введение

Как только человечество сталкивается с чем-то новым, оно начинает испытывать в его отношении серьезные опасения и глубокую обеспокоенность, даже если это новое абсолютно рукотворно. Оставляя в сто-

роне такую давнюю историю, как луддиты, даже в новейшем времени, времени постоянного и очень быстрого технического прогресса, можно почерпнуть массу таких примеров. Мы (люди) были обеспокоены роботами еще до их появления, компьютерами в целом, компьютерными вирусами, 5G-вышками, ИНН. Нет ничего удивительного в том, что искусственный интеллект тоже вызывает глубокую обеспокоенность и желание что-то с этим сделать.

Опыт предыдущих обеспокоенностей подсказывает, что на самом деле важно – это определить, во-первых, в какой мере то, с чем мы имеем дело, действительно так уж качественно ново, а во-вторых, насколько оно сможет на нас повлиять.

Опасения, которые всколыхнули достижения в области искусственного интеллекта (ИИ), вообще говоря, вечные, присущие человеку как фундаментальные:

- что власть над людьми захватят какие-либо нечеловеческие сущности;
- что конкретный человек потеряет свою ценность для общества (или стоимость на рынке труда);
- что сознанием, мнением, чувствами человека будут манипулировать.

Отсюда, например, рассуждения о необходимости информирования людей о начале взаимодействия не с человеком (в различных формулировках) с целью получения их информированного согласия на это взаимодействие. При этом, скажем, объявление следующей станции в вагоне метро тоже делается не человеком, и отсутствие предупреждения об этом пока не привело ни к какой катастрофе. К практике, кстати, такое требование зачастую мало применимо: например, в какой момент нужно предупредить пассажиров о том, что трамвай, в который они заходят, – беспилотный? Как будет выглядеть поездка при условии получения информированного согласия каждого пассажира? И почему этого не нужно делать, когда самолет сажает автопилот?

Манипулирование сознанием и чувствами людей – это, действительно, проблема, и то, что человечество задумывается о ней при каждом удобном случае – хорошо, или по крайней мере, нормально. Однако вряд ли именно ИИ представляет в этом отношении самую главную угрозу.

Как правило, феномен делается менее пугающим, становясь более изученным. Так наверняка будет и с ИИ.

(Искусственно) интеллектуальная беседа

С кем еще поговорить об искусственном интеллекте, как не с ним самим? Поболтать с ChatGPT мы пробовали раньше, впечатлились, но пользы не получили. Не смогли добиться нового знания, как ни старались. На этот раз попробовали поговорить с GigaChat. Результат примерно такой же. Отличие – существенно выше амбициозность, если такую характеристику можно применить к программе. Ну, чтобы было

понятно, о чем это мы – ниже выдержки из нашего разговора.

В стенограмме ВК – это Валерий Коняевский, который позволил себе существенно сократить многие ответы GigaChat в той их части, в которой отсутствие релевантной информации маскировалось избытком лишних слов, лишь слегка касающихся обсуждаемых вопросов. В этом взаимодействии мы старались понять, при каких условиях компьютерная программа может считаться искусственным интеллектом, как их можно классифицировать, и что можно и нужно защищать: программы, технологии или системы. В результате мы не всё смогли обсудить, но установили, что наш собеседник не владеет методами научного анализа и достаточной совокупностью знаний, использует комбинации бытовых и/или сомнительных положений, не очень-то улавливает контекст, но может реагировать на вопросы примерно, как студент начальных курсов смежных специальностей, и при должном усердии его можно «переубедить».

Итак, перейдем к стенограмме, а потом обсудим результаты.

ВК: Что такое искусственный интеллект?

GigaChat: Искусственный интеллект (AI) – это область компьютерных наук, которая занимается разработкой и исследованием методов и программного обеспечения, позволяющих машинам демонстрировать интеллект, аналогичный человеческому. <...>

ВК: Что такое технология искусственного интеллекта?

GigaChat: Технология искусственного интеллекта (ИИ) представляет собой научную технологию моделирования процессов познания и мышления человека с помощью вычислительных устройств. <...>

ВК: Что такое технология?

GigaChat: Технология – это совокупность методов, процессов и инструментов, применяемых для достижения конкретных целей, особенно в области науки и промышленности.

ВК: Совокупность – это множество, элементы которого применяются в произвольном порядке. разве можно произвольным набором операций изготовить изделие?

GigaChat: Нет, технология подразумевает упорядоченную и систематизированную последовательность действий, а не произвольный набор операций. Это ключевое различие между технологией и кустарным производством. В технологии каждый шаг строго определен и направлен на достижение предсказуемого результата.

ВК: Упорядоченную последовательность действий разве можно назвать совокупностью?

GigaChat: Да, вы правы, слово «совокупность» может ввести в заблуждение, так как оно обычно ассоциируется с множеством элементов без строгого порядка. Более точным будет сказать, что технология – это система методов, процессов и инструментов, применяемых в определенной последовательности для достижения конкретных целей.

ВК: Совокупность понимается как множество не «обычно» – это подразумевает бытовой, ненаучный подход. Мне кажется, что использовать для определения технологии понятие совокупность – это серьезная научная ошибка, вводящая в заблуждение других. Так?

GigaChat: Вы абсолютно правы. Использование понятия «совокупность» действительно может вводить в заблуждение, поскольку оно подразумевает отсутствие четкой структуры и последовательности. Лучше всего определить технологию следующим образом:

Технология – это система методов, приемов и средств, применяемых в определенной последовательности для достижения заданных целей, особенно в сфере производства, обработки информации или предоставления услуг.

ВК: Это определение ничуть не лучше. Чтобы это понять, нужно вспомнить, что такое «система», и мы увидим, что в определении системы упорядоченности нет.

GigaChat: Давайте, попробую дать еще более точное и простое определение.

Технология – это упорядоченная последовательность действий, методов и инструментов, применяемая для достижения конкретного результата.

ВК: *А вот теперь можно подумать, а что же такое «последовательность инструментов»? Похоже, это довольно бессмысленный набор слов.*

GigaChat: Понял вашу точку зрения. Давайте, я предложу окончательное определение, которое будет максимально ясным и точным.

Технология – это упорядоченный процесс применения знаний и методов для преобразования ресурсов в конечный продукт или услугу.

ВК: *Близко. Но и это не точно. Подумай еще.*

GigaChat: Хорошо, давайте, попробуем еще раз.

Технология – это упорядоченное применение научных знаний и технических методов для выполнения определенных задач и достижения конкретных результатов.

ВК: *Такое определение спекулятивно. В нем нет конкретики. Ведь определение должно быть конструктивным.*

GigaChat: Попробую сформулировать еще одно определение, учитывая ваши замечания.

Технология – это организованная система знаний, методов и оборудования, предназначенная для эффективного выполнения производственных, информационных или иных задач, обеспечивающая стандартизованный и воспроизводимый результат.

ВК: *Целью любой технологии является изготовление изделия на выходе из материалов на входе. Так?*

GigaChat: В целом да, но стоит уточнить, что цель технологий может варьироваться в зависимости от контекста. Основная идея заключается в том, чтобы преобразовать исходный материал или ресурс в ко-

нечный продукт, который обладает определенными свойствами или функциями. Это может включать:

- производство: изготовление товаров для потребления или использования людьми (например, автомобили, бытовая техника);
- обработка данных: преобразование информации в полезную форму (например, анализ больших данных для принятия решений);
- энергетика: производство энергии из различных источников (нефть, газ, солнечная энергия и т. п.);
- медицина: разработка лекарств, методов лечения и диагностики заболеваний.

Таким образом, технология направлена на создание чего-то нового или улучшение уже существующего продукта/процесса, используя определенные ресурсы и методы.

ВК: *Даже из этих примеров видно, что целью является изготовление известного, а не нового.*

Так, невозможно предложить технологию разработки новых лекарств, методов лечения и т. д.

Наоборот, – в медицине есть регламенты (то есть технологии) лечения тех или иных болезней.

А вот разработка новых регламентов и новых лекарств – это научное творчество.

И попытка в единую кучу свалить творчество и технологии – неконструктивна.

GigaChat: Вы поднимаете интересный момент о различии между технологией и научным творчеством. Действительно, технологии часто направлены на стандартизацию процессов и достижение предсказуемых результатов, тогда как научные исследования и инновации связаны с созданием нового знания и открытием неизвестных ранее возможностей. <...>

Однако стоит отметить, что границы между этими понятиями иногда размыты. Часто результаты научных исследований становятся основой для создания новых технологий, а развитие технологий открывает возможности для дальнейших научных изысканий.

ВК: *Садись, два! <...>*

ВК: *Знания – умения – навыки. Известная триада. А как они связаны?*

GigaChat: Триада «Знания – Умения – Навыки» отражает взаимосвязь между теоретической подготовкой и практическим применением знаний. <...>

1. Знания – это информация, которую человек получает через обучение, чтение, опыт других людей и так далее. <...>

2. Умения – это способность применять знания на практике для решения конкретных задач.

3. Навыки – это автоматические действия, которые выполняются без сознательного контроля. <...>

Таким образом, эта триада помогает понять процесс перехода от простого получения информации к эффективному применению её в реальной жизни.

ВК: *В книге есть информация? <...>*

GigaChat: Да, в книгах содержится информация.

ВК: *А знания есть в книгах?*

GigaChat: Да, книги содержат знания.

ВК: *Значит, если я куплю книгу, то я уже получил знания?*

GigaChat: Нет, покупка книги сама по себе не означает получение знаний. Чтобы получить знания, необходимо прочитать и усвоить информацию из книги.

ВК: *Так есть ли в книге знания?*

GigaChat: Книга содержит знания, но, чтобы ими овладеть, читатель должен прочитать и понять материал. <...>

● Информация – это конкретные данные, факты, цифры, события, описания объектов и процессов. Она представлена в виде текста, таблиц, схем и диаграмм.

● Знания – это понимание смысла этих данных. <...>

ВК: *Тогда следующий вопрос – а содержит ли книга понимание смысла? <...>*

GigaChat: Нет, в книге нет готового понимания смысла. Понимание возникает у читателя в процессе

чтения и осмысления содержания книги.

ВК: *Если знание – это понимание смысла, а его в книге нет, то это означает, что знаний в книге нет. Так? <...>*

GigaChat: Если определять знания как понимание смысла, то в книге самих по себе знаний нет, поскольку понимание возникает у читателя в процессе взаимодействия с текстом.

ВК: *Отлично, уже лучше! Тогда, где же знания? <...>*

GigaChat: <...> знания находятся в голове человека.

ВК: *Без знаний интеллект возможен?*

GigaChat: Интеллект без знаний возможен, но его эффективность будет ограничена. Знания служат основой для применения интеллекта в решении задач и принятии решений.

ВК: *Интеллект без знаний возможен? Приведи пример.*

GigaChat: Представьте младенца, который ещё ничего не знает о мире вокруг, но уже обладает способностью учиться, исследовать окружающее пространство и реагировать на новые стимулы. Этот младенец демонстрирует базовые формы интеллекта, такие как любопытство и способность к обучению, несмотря на отсутствие накопленных знаний.

ВК: *Любопытство у кота – он интеллектуален?*

GigaChat: Любопытство у кота говорит о его природной склонности к исследованию окружающей среды, но это не делает его интеллектуальным в человеческом понимании. Интеллект подразумевает более сложные когнитивные процессы, такие как логическое мышление, решение сложных задач и способность к абстрактным понятиям, которыми коты не обладают.

ВК: *А младенцы этим обладают – логическое мышление, решение сложных задач и способность к абстрактным понятиям, которыми коты не обладают.*

GigaChat: Нет, младенцы тоже не обладают этими сложными когнитивными функциями сразу при рождении. Эти способности развиваются постепенно по мере роста и обучения. Младенцы начинают с базовых форм восприятия и реакции на окружающую среду, а затем, с возрастом, приобретают более сложные навыки мышления и решения задач. <...>

ВК: *Супер! Вернемся к моему вопросу – возможен ли интеллект без знаний?*

GigaChat: Абсолютный интеллект без всяких знаний невозможен. Даже для самых простых действий и решений требуются некоторые базовые знания и навыки.

ВК: *Хорошо, а вот теперь перейдем к искусственному интеллекту. Возможен ли ИИ без знаний? <...>*

GigaChat: Да, искусственный интеллект без знаний невозможен. <...> Знания искусственного интеллекта хранятся в его программном обеспечении, базе данных и моделях, которые используются для обработки информации и принятия решений.

ВК: *То есть знания не обязательно в голове у человека? Могут быть и в компьютере? <...>*

GigaChat: Да, знания могут быть в компьютере.

ВК: *Ок. Если знания в компьютере – то они в каком виде?*

GigaChat: Знания в компьютере могут быть представлены в разных формах, таких как данные, алгоритмы, модели, программный код. Эти формы позволяют компьютеру хранить, обрабатывать и использовать информацию для выполнения задач.

ВК: *А если знания в голове – то они в каком виде?*

GigaChat: Знания в голове человека существуют в виде нейронных связей и паттернов активности мозга.

ВК: *То есть в голове – это связи, а в компьютере – это данные, модели и связи? Так?*

GigaChat: Верно. В голове человека знания представлены в виде ней-

ронных связей, а в компьютере – в виде данных, моделей и программного кода.

ВК: *То есть знания у человека ограничены множеством связей, а у компьютера – расширяются еще и еще за счет данных и моделей?*

GigaChat: Да, знания у человека ограничены возможностями мозга, а компьютер может расширять свои знания за счет добавления новых данных и моделей.

ВК: *А человек не может этого сделать, так как не может манипулировать данными и моделями. Так? <...>*

GigaChat: Да, человек не может манипулировать данными и моделями так, как это делает компьютер.

ВК: *То есть возможности человека ограничены, а возможности компьютера – нет. Так?*

GigaChat: Возможности человека действительно ограничены биологическими факторами, такими как емкость памяти и скорость обработки информации. Компьютер, напротив, может постоянно увеличивать свои ресурсы и производительность, поэтому его возможности теоретически неограниченны. <...>

ВК: *Ок, мы поговорили о знаниях, и пришли к забавному выводу – возможности человека ограничены, а компьютера нет. Конечно, это глупость, и позднее я это продемонстрирую на противоречиях из твоих ответов. Давай сейчас поговорим об умениях. Что это?*

GigaChat: Умения – это способность человека или системы выполнять определенные действия на основе имеющихся знаний и опыта. Это практические навыки, которые позволяют решать задачи и достигать целей. <...>

Для формирования умений обычно требуется сочетание знаний и опыта. Знания предоставляют необходимую информацию, а опыт позволяет применить эти знания на практике и развить навыки.

ВК: *То есть без знаний умений быть не может? <...>*

GigaChat: Некоторые базовые умения могут существовать без глубоких знаний, но большинство сложных умений требуют наличия знаний.

ВК: Пример. <...>

GigaChat: <...> есть умения, которые не требуют знаний, например, рефлекс сосания у новорожденных.

ВК: Рефлекс сосания реализуется в действии. В процессе. А пока он не реализуется, он же существует, ну, хотя бы в виде некоторого описания. Так? <...>

GigaChat: Да, именно так.

ВК: Наверняка описывается связями нейронов. Другого же механизма нет. Так?

GigaChat: Верно, другого механизма нет.

ВК: Чем тогда это отличается от знаний?

GigaChat: Отличие заключается в том, что знания – это осознанная информация, которую человек может использовать для принятия решений, анализа ситуаций и планирования действий. Они формируются на основе опыта, обучения и мышления. Рефлексы же являются автоматическими реакциями организма на стимулы, происходящими без участия сознания.

ВК: Описание и там, и там – связи.

Отличия – не в связях, а в условиях активации. Так?

GigaChat: В целом да, отличия заключаются в условиях активации и участия сознания. <...>

Описание в обоих случаях действительно сводится к связям между нейронами. Разница лишь в том, как эти связи используются: рефлексы работают автоматически, а знания требуют осознания и могут использоваться произвольно.

ВК: Хорошо.

Значит, данные, модели и программный код в системах ИИ должны использоваться осознанно, иначе они не поведут, а рефлекс? <...>

Интеллект у человека – осознанный, а у машины – неосознанный. Так?

GigaChat: Именно так. <...>

ВК: <...> И так – можно ли рефлексы считать умениями, если они <...> неосознанные?

GigaChat: Правильно, рефлексы лучше не называть умениями.

ВК: Хорошо! Тогда мы установили цепочку «знания – умения». Без знаний нет умений. Это важный вывод.

GigaChat: Да, это важная связь. Знания служат основой для формирования умений, так как умения предполагают применение знаний на практике. <...>

ВК: Хорошо. Перейдем теперь к «навыкам». Мы выделили аналогии между знаниями и умениями, и видим различия, которые принципиальны и отличаются только способом фиксации. Но человек кроме знаний и умений еще владеет и некоторыми навыками. Есть ли аналогия в СИИ?

GigaChat: Отличный вопрос! Да, существует определенная аналогия между человеческими навыками и тем, что мы можем наблюдать в системах искусственного интеллекта (ИИ). <...>

Различие заключается в том, что навыки человека обычно связаны с физическим выполнением действий (например, игра на пианино), а навыки ИИ ограничены виртуальной сферой (например, генерация текста) <...> и человек, и ИИ развивают свои способности через практику и обучение.

ВК: Умения – через обучение. Навыки – через практику. Не так ли?

GigaChat: Верно подмечено!

● Умения обычно приобретаются через процесс обучения, где человек получает новые знания и учится применять их на практике. Например, чтобы научиться играть в шахматы, нужно сначала изучить правила игры, понять стратегии и тактику.

● Навыки, напротив, развиваются через многократную практику и за-

крепление полученных знаний до уровня автоматизма. Продолжим пример с шахматами: игрок совершенствует свои навыки, регулярно играя партии, анализируя ошибки и улучшая свою игру. <...>

ВК: Очень плохо! Играя в шахматы, человек совершенствует умение. Навык передвигать фигуры – не очень важен в этом процессе. Пока не вижу аналогии в СИИ. <...>

Навык используется в автоматическом режиме, как правило, неосознанно, без прямого использования знаний. Для СИИ аналогии пока не видно.

GigaChat: Спасибо за уточнение!

<...> Основное различие между навыками у человека и ИИ заключается в том, что человеческий навык основан на физическом опыте и мышечной памяти, тогда как навык ИИ проявляется в цифровой среде и зависит от вычислительных мощностей и архитектуры модели. Тем не менее, в обоих случаях навык подразумевает выполнение определенных операций без активного участия сознания (у человека) или явного пересчета всех возможных вариантов (в случае ИИ).

ВК: Ну нет! И человек не обучается завязыванию шнурков при каждом акте обувания, и СИИ не переобучается. Но СИИ обрабатывает данные, как всегда, а человек нет. Потому что у человека навыки – не в сознании, а в подсознании. А у СИИ и сознания – то нет.

GigaChat: Понимаю твою мысль. Действительно, есть важное различие между тем, как навыки реализуются у человека и у ИИ. <...>

Основное различие заключается в том, что у человека навыки закреплены на уровне физиологии и психологии, тогда как у ИИ они представляют собой результат сложных математических вычислений. И хотя обе системы могут демонстрировать высокую эффективность в выполнении определенных задач, механизмы, лежащие в основе этой эффективности, существенно различаются. Поэтому можно сказать, что понятие «навыка» в контексте ИИ является скорее метафорой, чем

точным эквивалентом человеческих навыков.

ВК: Ты повторил мои аргументы. Это хорошо, но неинформативно.

Суммируем итоги беседы

В диалоге с СИИ мы с огромным трудом выяснили, что интеллект и его проявления плотно связаны с объемом знаний, умений и навыков, которые, в свою очередь, как-то связаны с такой категорией, как информация. В процессе изучения научных работ [1 и др.] усвоение этих истин проходит гораздо быстрее. Мы понимаем, что информация (отражение движения объектов материального мира в системах живой природы) интерпретируется в организме живого существа (например, человека) в сведения (*запечатленные в организме результаты* отражения движения объектов материального мира), которые преобразуются в знания (*производные сведения о закономерностях изменения* состояния отражаемых объектов материального мира). Разные организмы интерпретируют одну и ту же входную информацию в разные сведения и получают из них разные знания (в зависимости и от разницы стартовых условий, и от того, какие именно результаты движения объектов материального мира релевантны для данного организма).

В диалоге с СИИ мы выяснили также, что знания – в голове человека, и фиксируются они (предположительно) связями между нейронами. Но в этой беседе мы не нашли аналога в компьютерной системе, так что об этом следует поговорить. Не обнаружилось и ничего, что было бы похоже на сознание, позволяющее изменять представления об объектах вещного мира. И, конечно, мы обратили внимание, что наш собеседник считает себя и себе подобных существом *программным продуктом*, который может влиять на виртуальную, цифровую сферу, но не может влиять на аналоговую, физическую жизнь.

Если принять эту точку зрения, то ничего делать не нужно. Конечно, это не так, к сожалению, уже (воз-

можно, не «уже», а «пока») есть печально известные примеры совершенно буквального, физического влияния на жизнь человека, например, беспилотного трамвая. Очевидно, что положение о том, что ИИ способен влиять только на виртуальную сферу – это неумное положение, связанное, возможно, с ограниченностью процесса обучения нашего собеседника. Но, может быть, не все так просто, и дело в том, что научного анализа развития СИИ пока нет? В частности, отсутствуют даже основы классификации таких систем. Без классификации не только нереально разрабатывать нормативную правовую базу, но и весьма затруднительно формулировать требования по технической защите.

Нельзя определять ИИ через сравнение некоторых заранее зафиксированных функций (алгоритмов, моделей) с возможностями человека. Калькулятор выполняет деление многозначных чисел существенно быстрее человека, антивирусные программы идентифицируют последовательности бит как «вирус» точно лучше большинства людей, человек и отдельные биты различает с трудом, но ни калькулятор, ни антивирус не относятся к ИИ.

Как же идентифицировать ИИ? Похоже, что с появлением больших языковых моделей (БЯЗ, LLM) известный принцип Тьюринга устарел. При этом попытки регулирования в области ИИ постоянно возникают, и без уверенной идентификации эти попытки могут остановить перспективное направление, а не стимулировать его развитие. Ясно, что алгоритмизации формирования связанных фраз мало для признания процесса интеллектуальным, но столь же ясно, что это один из признаков такового. Пусть необязательный, но важный. На что нужно опереться? На имитацию общения? На возможность построения суррогатных моделей в тех случаях, когда нет возможности построить прямую модель? На правдоподобность прогнозов? На псевдоинтеллектуальное поведение?

Все это можно использовать, но базового признака интеллектуальности ни в чем из этого нет. А в беседе с GigaChat мы его установили –

это наличие знаний. Нет знаний – нет интеллекта.

Далее – то, что пока установить не удалось, а именно, как нужно хранить знания в компьютерной системе. Добиться от СИИ этого знания не удалось. Хотя специалисты давно хорошо понимают, что целое есть единство формы и содержания. Числа – это форма. Содержание – это семантика. Алан Тьюринг этого еще не заметил: у него семантикой оперировал человек, изучающий последовательности символов – если семантика появилась, то ключ шифрования найден. А вот Марвин Минский уже хорошо это знал, написав поистине замечательный труд «Фреймы для представления знаний». Кейслер, Тарский, Поспелов – известные ученые, изучавшие теорию моделей, семиотику, теорию языков и смежные аспекты. Сейчас исследования в части семантической интероперабельности продолжают ученые мира, в том числе и российские, например, Ю. М. Акаткин [2]. Общее положение заключается в том, что знания фиксируются в виде семантических активов (СА) – словарей, справочников, классификаторов, онтологий и других, при этом не в виде цифр и чисел, а в виде связанных данных. СА содержат описания нашего представления об объектах и сущностях предметной области. Заметим, не описания объектов, а описания представления об объекте. От мирового уровня здесь мы уже отстаем, но новые результаты появляются. Нам же пока достаточно понимать, что основа – это СА и инструмент работы с ними – система управления знаниями (СУЗ). А связанность – это связь формы и содержания стандартизированной конструкцией, например, фреймами М. Минского или образами У. Гренандера. Однако с высоким уровнем уверенности можно сказать, что необходимой характеристикой наличия ИИ является наличие формализованных СА и СУЗ.

Знания, умения, навыки

Знания – фиксируются в виде СА. Управление знаниями – похоже на сознание, особенно с учетом контек-

та, особенно в русском языке. Понятие «навыки» похоже на «алгоритм». Зачастую навыков достаточно для успешного функционирования в быту, особенно при выполнении относительно однообразных и достаточно несложных операций. А при наличии некоторых «умений» изменять алгоритм и модифицировать поведение, соответствующее «навыку», исполнителя зачастую можно идентифицировать уже как специалиста.

Каждую из составляющих триады можно считать некоторым проявлением интеллекта, а каждое сочетание свойств – вполне хорошее основание для классификации алгоритмической составляющей систем ИИ по глубине присущей им интеллектуальности.

Принципиально важным для данной классификации является понимание, что только владение всеми составляющими триады позволяет говорить о наличии у человека той или иной степени интеллекта. И, значит, говорить об ИИ можно лишь тогда, когда умения и навыки основаны на знаниях. Знания должны быть формализованы для возможности их обработки в СИИ и представляться в виде семантических активов (СА) – словарей, справочников, классификаторов, онтологий, и систем управления знаниями (СУЗ). Иначе алгоритмическая составляющая характеризуется некоторой интеллектуальностью, но и не более того.

Есть и другое основание для классификации: функция (в «человеческой» терминологии – профессия). Это будет классификация моделей (ведь речь тут идет об «умениях»). Здесь можно выделить:

- простые (например, большие языковые модели (БЯЗ или LLM), кредитный скоринг и др.);
- расширенные (например, БЯЗ с контекстом – RAG + LLM);
- гибридные (например, с использованием систем управления знаниями (СУЗ) – СУЗ + RAG + LLM).

С точки зрения безопасности классификация будет уже другой – классификацией систем ИИ. Здесь уже мы должны говорить не об ИИ, моделях и технологиях, а именно, о СИИ.

Действительно, вряд ли технология может сама по себе чем-то навредить человеку. Используются системы, человек взаимодействует с системами, и поэтому требования по безопасности должны выдвигаться в отношении систем.

В свою очередь, системы ИИ вполне можно разделить по степени их влияния на аналоговый мир на физическом уровне, что детерминировано степенью их автономности от оператора. Так, системы искусственного интеллекта можно разделить на:

- программные (функционируют под управлением человека, не оказывают влияния на аналоговый мир на физическом уровне, то есть результаты их работы – виртуальные);
- кибер-физические (функционируют под контролем человека (например, автопилот), оказывают влияние на аналоговый мир на физическом уровне, но человек может в случае необходимости перехватить управление и предотвратить негативное влияние);
- автоматические (функционируют полностью автономно (например, робот-доставщик), оказывает влияние на аналоговый мир на физическом уровне, и оно не может быть скорректировано управляющим персоналом, так как этого персонала нет).

Для каждого из этих видов СИИ нужно формулировать свои требования по защите информации. При этом основы доверенности должны закладываться от этапа сбора данных для обучения до принятия и исполнения решений.

Доверие к ИИ

Доверенность СИИ, в частности, разговорных систем, робототехнических систем, беспилотного транспорта необходимо рассматривать в нескольких взаимосвязанных аспектах:

- доверенность методов и алгоритмов (при каких условиях их можно считать доверенными, как установить доверенное состояние системы);
- функциональная безопасность (безопасность функционирования

при обеспечении условий доверенности);

- устойчивость системы в условиях агрессивной среды (при фиксированных уровнях нарушения условий доверенности);
- защищенность от стороннего вмешательства (например, в структуру и содержание базы знаний, в том числе в процессах дообучения, а также в последовательность сигналов управления).

На основе анализа публикаций в открытой печати можно считать, что эта тематика пока недостаточно раскрыта в научных исследованиях и практике. В частности, широко известны случаи перехвата хакерами управления беспилотным автотранспортом и БПЛА, что в достаточной степени иллюстрирует опасность применения систем ИИ без использования механизмов обеспечения доверия.

Развитие направления обеспечения доверенного характера разрабатываемых математических моделей, алгоритмов искусственного интеллекта, программного или аппаратно-программного обеспечения связано с анализом и фиксацией обучающих выборок, анализа устойчивости и обеспечения целостности гибридных схем ИИ, выработкой требований, рекомендаций и набора контрольных процедур для установления соответствия требованиям доверенности при дообучении системы, обеспечения доверенного взаимодействия по каналам связи с помощью доверенных аппаратно-программных средств защиты информации.

Основная задача – обеспечить доверенное взаимодействие как в массовых применениях, так и при попытке реализации целевой атаки. При этом рассмотрение должно проводиться в разрезе масштабов (человек, организация, общество), и по стадиям жизненного цикла (разработка, развертывание, использование).

Должны быть прописаны строгие требования по организации инфраструктуры сбора, хранения и оборота моделей данных и собственно данных.

Необходимо также разработать способы оценки сложности атак, сложности механизмов защиты от

этих атак, рассмотреть возможные конфликты между робастностью моделей и конфиденциальностью обрабатываемых в них данных.

Кроме того, важнейшим моментом является оценка интерпретируемости моделей.

Значительная часть требований к защите информации вполне может быть позаимствована из числа уже давно сформулированных, описанных и освоенных. Так, если говорить о программных системах, то на этапе эксплуатации практически не видится принципиальных отличий СИИ от любой информационной системы (ИС) аналогичной области применения.

Область применения тут имеет принципиальное значение. Пример – наша беседа, приведенная выше. Ничем эта беседа не отличается от беседы за чаем с соседом по даче. Ну поговорили, ну улыбнулись потоку штампов и забавным заблуждениям, и разошлись – кто за компьютер, кто за лопату. Однако если это будет информационная система государственных органов, государственных унитарных предприятий, государственных учреждений, или, тем более, государственная информационная система (№ 216-ФЗ в редакции от 08.08.2024), то необходимо как минимум выполнить требования по защите обрабатываемой в этих ИС информации по требованиям, устанавливаемым регуляторами.

Однако в СИИ, обученных на больших наборах данных, возникают и новые уязвимости. Так, по существу, значительная часть таких систем может быть описана как статистическая база данных с неконтролируемым потоком запросов. В этом случае, анализируя большое число результатов запросов, зачастую можно восстановить исходные наборы данных. Это является косвенной утечкой, требований к поиску которых пока нет. Возникает задача обеспечения дифференциальной конфиденциальности, включая блокирование целевых атак. Это может быть существенным для, например, задач кредитного и страхового скоринга, антифрода и других задач этого класса. На этом пути уже получен ряд фундаментальных результатов

в анализе и блокировании вредоносных запросов при известных алгоритмах работы СИИ.

Существенно сложнее обстоят дела в кибер-физических системах. Примером может быть беспилотный транспорт в различных своих вариантах (взаимодействие «Сапсана» с БПЛА, летящим перед ним для контроля состояния пути, или же БПЛА, используемые в качестве летающих светофоров для управления потоками перемещения людей и автотранспортных средств). И здесь сохраняется тот же принцип: должны быть применены все известные меры, которые при этом следует расширить за счет механизмов блокирования специфических угроз.

А вот в автоматических системах вряд ли есть существенные различия, так как безопасность человека, животных и растений регулируется техническими регламентами, что целесообразно зафиксировать как правильную меру.

При этом во всех перечисленных видах СИИ остается чрезвычайно важной проблема обновления. Эта проблема усугубляется возможностью атак, связанных с дообучением и переобучением, и пока не нашла существенных фундаментальных решений. На данный момент здесь видится только одна радикальная мера: *разделение функций обучения моделей и их применения*. Обновление, естественно, должно проводиться с соблюдением всех требований, обеспечивающих контроль авторства, целостность и конфиденциальность.

Отдельным направлением исследования может быть применение систем ИИ для поиска уязвимостей в системах ИИ же, а также в повышении уровня доверия, например, на основе мажоритарных механизмов с близкими обучающими последовательностями и близкими семантическими средами. Необходимо изучить возможность применения подходов класса «слой доверия» из группы семантических стандартов W3C, анализируя возможность фиксации источника происхождения данных, присвоения ему некоторого индекса (уровня) доверия, и в соответствии с этим оценивать доверие

к выводам, сделанным на основе этих данных.

Краткие итоги

Итак, интеллект человека – это знания, умения и навыки. Именно знания, а не сведения или сообщения. Соответственно, аналогия для ИИ – не данные, а семантические активы. Инструмент работы с ними – система управления знаниями, аналог сознания человека (в отличие от рефлексов). Умения – это модели, навыки – алгоритмы. Наличие всей триады говорит о наличии ИИ.

Проявления интеллектуальности могут характеризовать алгоритмическую составляющую системы (и это может быть основой для их классификации как более или менее интеллектуальных), структурная сложность характеризует модели (и по этой характеристике их можно разделить на простые, сложные и гибридные), а СИИ характеризуются степенью автономности и степенью влияния на аналоговый мир на физическом уровне, и именно это важно для предъявления к ним требований по защите информации.

В свое время воображение слушателей доклада Б. Н. Мирошникова поразил тезис, что если человека убили, ударив по голове компьютером, то к компьютерным преступлениям это деяние все же не относится. Сейчас пришло время провести следующие аналогии этого типа и снова сделать правильные выводы и провести конструктивные границы.

Ну и, наконец, проблемы людям, как правило, доставляет не интеллект, а его отсутствие. Поэтому всем исследователям и разработчикам СИИ желаем успеха, и сами попробуем внести свой скромный вклад в обеспечение доверенности таких систем. ■

ЛИТЕРАТУРА

1. Стрельцов А. А. Правовое обеспечение информационной безопасности России: теоретические и методологические основы / А. А. Стрельцов. – Минск, 2005. – 304 с.
2. Акаткин Ю. М. Цифровая трансформация государственного управления: датацентричность и семантическая интероперабельность / Ю. М. Акаткин, Е. Д. Ясиновская. – М: URSS: ЛЕНАНД, 2019. – 724 с.