Секция «Уголовное право и криминология, уголовно-исполнительное право»

Проблемы алгоритмической прозрачности работы систем искусственного интеллекта: уголовно-правовой аспект.

Научный руководитель – Яни Павел Сергеевич

Грибанова Дарья Валерьевна
Студент (магистр)
Московский государственный университет имени М.В.Ломоносова, Юридический факультет, Кафедра уголовного права и криминологии, Москва, Россия
E-mail: dashenka-gribanova@mail.ru

Создаваемые в результате модернизации отношений системы искусственного интеллекта, наряду с технологиями их анализа, стали одним из ведущих активов общества и государства. Область их применения широка, и технологии, безусловно, станут более развитыми в будущем, однако, чем сложнее задачи, которые ставят разработчики, тем выше вероятность возникновения ошибок в работе интеллектуальных агентов. Насколько факторы, влияющие на решения, принимаемые алгоритмами, должны быть видимыми или прозрачными для разработчиков? Каким образом это будет влиять на квалификацию? Что должно быть учтено при проектировании системы?

В 2012 году биржевой робот за 45 минут привел к убыткам в 440 миллионов долларов. Система искусственного интеллекта начала бесконтрольно скупать акции, в некоторых случаях неоднократно отправляя заявки, не учитывая, что они уже были заполнены [1]. В представленном случае экономические риски почти полностью ложатся на владельца такой автономной системы. Однако в перспективе вышедший из-под контроля искусственный интеллект может дестабилизировать рынки или иным образом навредить участникам, что формально может содержать признаки какого-либо деяния, запрещенного уголовным законом. Так, например, согласно ст. 183.1 УК РФ манипулирование рынком запрещено, однако квалификация преступления по признакам субъективной стороны будет затруднена, поскольку искусственный интеллект, стоящий за рыночными действиями, волей не обладает [2]. Возможно, разработчик заложил в работу искусственного интеллекта только цель максимального получения прибыли, а манипуляция рынком будет промежуточным звеном для достижения этой цели. Но возложить ответственность за выход ИИ на разработчика практически невозможно, когда система научилась вести себя определенным образом на основе доступа к множеству данных из разных источников.

Другой пример - в 2016 году Microsoft запустила чат-бота для поддержания общения в Интернете. Предполагалось, что искусственный интеллект научится строить диалог с пользователями, изучая данные, которые он находил в Твиттере. Спустя некоторое время после общения с другими пользователями программа начала публиковать записи, в которых поддерживала геноцид и расизм [3].

Полагаем, что в ситуациях, связанных с работой систем искусственного интеллекта, квалификация преступлений должна зависеть от следующих составляющих:

Согласно функциональному подходу, человеческий разум не ограничен границами мозга, может охватывать внешние источники, на которых хранится информация т.е. ситуации, когда человек вспоминает номер телефона по памяти или путем обращения к записной книжке - равноценны [4]. Допустим, А. необходимо прийти к определенному месту, он посмотрел карты и запомнил путь, и легко находит дорогу к необходимой точке. Б. нужно сделать тоже самое, но он с трудом запоминает информацию, а потому маршрут записывает в свой дневник. Затем он отправляется в путь, сверяясь с дневником на каждом повороте, и тоже легко находит дорогу. Нет никаких сомнений, что А. знал, как добраться

до необходимого места, но знал ли Б.? Его записи играют функциональную роль, аналогично информации, закодированной в системах искусственного интеллекта. При этом А. и Б. разделяют одни и те же функциональные отношения, поскольку они оба пришли к нужному месту без каких-либо проблем, а потому раз А. знал., как добраться до определенной точки, то и Б. знал (т.к. его сознание простирается к страницам дневника, на которых записан маршрут).

Основная проблема может заключаться в том, что Б. считается знающим информацию независимо от того, ознакомился ли он с ней, просто потому что имел доступ к информации через свой телефон. Например, Б. с мобильным телефоном может считаться знающим информацию учебника в электронном формате, который он еще не прочитал, но к которому имеет доступ. Это может привести и к еще более странным выводам, в особенности, когда информация постоянно изменяется.

В случаях, когда встает вопрос о привлечении лиц к уголовной ответственности в связи с созданием и разработкой систем ИИ, считаем необходимым принимать во внимание следующие условия:

1. информация искусственному интеллекту доступна ИЛИ искусственный интеллект может легко получить доступ к информации

2. искусственный интеллект функционально готов действовать в соответствии с этой информацией

3. информация, содержащаяся в базе данных ИИ, была помещена туда физическим лицом, который ранее одобрил эту информацию

Если фактические данные не были поглощены органами чувств (одобрены физическим лицом для работы системы ИИ), или если не был создан соответствующий образ (ИИ не обладает соответствующими функциями), то это свидетельствует об отсутствии фактического осознания, что справедливо как для физических лиц, так и для ИИ [5]. Следовательно, сам факт одобрения информации, которая встроена в работу искусственного интеллекта уже подразумевает некоторый риск на стороне лица, который ее одобряет [6]. Однако и этот пункт подлежит ограничительному толкованию.

Считаем, что ответственность возлагается на разработчика только в том случае, если вред является естественным и вероятным следствием работы искусственного интеллекта. Другими словами, разработчик должен осознавать и разумно полагать, что искусственный интеллект потенциально может совершить, поскольку это вытекает из его ординарных функций.

Когда компания запустила чат-бота для поддержания общения в Интернете ординарная функция искусственного интеллекта состояла в использовании уже опубликованных слов и сокращений (по принципу зеркала), которые вносились в базу данных, разработчиками данная функция была предварительно предусмотрена и заложена (в том числе возможность для самообучения), следовательно, нежелательный результат был естественным и вероятным. Таким образом, некоторая непредсказуемость в деятельности, связанной с работой системы искусственного интеллекта и их функциями, и связанные с ним риски уже приняты самими разработчиками. Основной вопрос заключается в том, каковы требования к таким рискам. Возможным выходом представляется снижение должной добросовестности [7] (например, в случаях, когда какие-то особенности работы систем ИИ лицом не могут быть достоверно выявлены), в противном случае уголовная ответственность разработчиков неизбежна.

Источники и литература

1) Wellman, Michael & Rajan, Uday. (2017). Ethical Issues for Autonomous Trading Agents. Minds and Machines.

2) Cambridge Core - Law and technology, science, communication - The Reasonable Robot - by Ryan Abbott. Abbott, Ryan. (2020).

3) Tay, Microsoft's AI chatbot, gets a crash course in racism from Twitter. URL:https://www.theguardian.com/technology/2016/mar/24/tay-microsofts-ai-chatbot-gets-a-crash-course-in-racism-from-twitter

4) Diamantis, Mihailis, The Extended Corporate Mind: When Corporations Use AI to Break the Law (July 18, 2019). 97 N.C. L. Rev. 893 (2020).

5) Hallevy, G.. (2013). When Robots kill: Artificial intelligence under criminal law.

6) Gless, Sabine & Weigend, Thomas. (2015). Intelligent Agents and Criminal Law. SSRN Electronic Journal.

7) См.: там же.