УДК 004.85

РАЗРАБОТКА МОДУЛЯ БЕЗОПАСНОСТИ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА НА ОСНОВЕ ОГРАНИЧЕНИЙ ПОВЕДЕНИЯ СИСТЕМЫ ПОДДЕРЖКИ И ПРИНЯТИЯ РЕШЕНИЙ

Краснов Т.А., Компаниец Р.И. (ВКА им. А.Ф. Можайского)
Научный руководитель – кандидат технических наук, Дудкин А.С.
(Военно-космическая академия им. А.Ф.Можайского)

Введение. В последние десятилетия неуклонно наблюдается рост разработок в области искусственного интеллекта (далее - ИИ). Возьмем, к примеру, “OpenAI five” (команда компьютерной игры dota, основанная на ИИ, которая превосходит даже самых опытных игроков), или же модель ИИ от разработчиков Facebook, который разговаривает с другим ИИ на языке, который человек не в состоянии понять. В наше время системы ИИ превосходят людей в научных и инженерных способностях. Автономные обучающиеся системы учатся по ходу работы, а это значит, что гораздо сложнее точно знать, как они будут вести себя при развертывании. Основные проблемы, такие как прерывание таких систем, влияющие на то, как они учатся вести себя, или как отсутствие руководителя влияет на действия, которые они решают предпринять, в большинстве случаев остаются нерешенными. Кроме того, системы ИИ могут небезопасно исследовать свою среду и вызывать негативные побочные эффекты. Возникает необходимость создания некоторого модуля ограничений, который будет интегрирован в систему ИИ и регулировать поведение его системы поддержки и принятия решений, тем самым предотвращая его от принятия небезопасных решений.

Основная часть. На основе анализа принимаемых небезопасных решений систем ИИ решаются три типа задач:

1) Задача максимального снижения или полного аннулирования вероятности принятия системами ИИ небезопасных решений, заключающаяся в поиске оптимального подхода к снижению ошибок системы поддержки и принятия решений, на основе чего будет разработан модуль безопасности ИИ, ограничивающий ее поведение.
2) Задача создания “предохранителя” от принятия небезопасных решений для систем ИИ, целью которого будет предотвращение небезопасных решений систем ИИ, вплоть до ее полного отключения, и который будет составлять основу модуля безопасности ИИ.
3) Задача интеграции разработанного модуля безопасности ИИ в существующие системы ИИ с возможностью его реконфигурации для достижения максимальной безопасности.

Выводы. Проведен анализ принятия небезопасных решений системами ИИ, разработан модуль ограничений поведения системы поддержки и принятия решений ИИ.

Список использованных источников:

1. Баррат, Д. Последнее изобретение человечества: искусственный интеллект и конец эры Homo sapiens: [пер. с англ.] // Д. Баррат. - 2-е изд. - М. : Альпина нон-фикшн. – 2018. – С. 53–82.
2. Демкин В. И. История и перспективы развития нейронных сетей / В.И. Демкин, Д. К.Луков // Вестник современных исследований. – 2018. – № 6.1 (21). – С. 366–368.

Краснов Т.А. (автор) подпись

Компаниец Р.И. (автор) подпись

Дудкин А.С. (научный руководитель) подпись