Секция «Этика»

The Alignment Problem: ценности и искусственный интеллект

Научный руководитель – Танюшина Александра Александровна

Муминова Парвина Диловаровна
Студент (магистр)
Московский государственный университет имени М.В.Ломоносова, Философский факультет, Москва, Россия
E-mail: muminovaprvn@gmail.com

В связи с широким распространением систем искусственного интеллекта возникает как множество возможностей, так и неопределенное количество качественно новых проблем, в том числе этических. Учитывая нашу ограниченную способность направлять и предвидеть результаты, достигаемые системами искусственного интеллекта, изучение данных проблем является первоочередной задачей.

Одной из наиболее важных проблем, с которой постоянно сталкиваются исследователи и разработчики искусственного интеллекта, является проблема согласования, соответствия, или так называемая «alignment problem» [2]. Впервые данная проблема была сформулирована Норбертом Винером, математиком и основателем кибернетики. Винер писал: *"Если мы используем для достижения наших целей какое-либо механическое устройство, в работу которого мы не можем вмешиваться ... нам лучше быть совершенно уверенными в том, что цель, заложенная в машину, - это та цель, к которой мы действительно стремимся"* [4].

В рамках данной проблемы ставится вопрос о том, какие или чьи ценности должны быть положены в основу при разработке систем искусственного интеллекта. Существует несколько противоположных подходов, которые отмечает исследователь Я. Габриэль [3]. Первый подход можно охарактеризовать как утилитаристский, поскольку, согласно этому подходу, следует разрабатывать искусственный интеллект таким образом, чтобы он приносил как можно больше блага наибольшему количеству людей и других разумных существ. Другой подход, который можно охарактеризовать как кантианский, требует установить руководящими принципами искусственного интеллекта те, которые могли бы быть универсальными законами. Например, принцип справедливости.

Важно отметить, что выбор ценностей для систем искусственного интеллекта сопряжен с еще большим количеством проблем. Какие ценности и принципы следует заложить в алгоритмы искусственного интеллекта и кто имеет право принимать подобные решения? На данные и подобные вопросы необходимо ответить как можно быстрее, поскольку вопрос о согласовании ценностей становится все более актуальным по мере того, как компьютерные системы работают с большей автономией и скоростью, которая может помешать людям оценивать, выполняется ли какое-либо действие этичным образом [1].

Задача согласования состоит из двух частей. Первая часть, техническая, концентрируется на более прикладной задаче: как запрограммировать ценности и этические принципы так, чтобы искусственный интеллект выполнял именно то, что от него требуется, обращая внимание на то, что у разных людей представление о ценностях может быть диаметрально противоположным. Важно привить системам искусственного интеллекта понимание человеческих ценностей, чтобы их действия и решения действительно соответствовали намерениям человека, который их использует. Для этого требуется более глубокий уровень понимания и интеграции человеческих ценностей в процессы принятия решений системами искусственного интеллекта. Вторая часть, нормативная, ставит вопрос о том, какие

ценности или этические принципы необходимо положить в основу алгоритмов систем искусственного интеллекта.

Подводя итог следует отметить, что только должным образом приведя системы искусственного интеллекта в соответствие с человеческими ценностями, мы сможем использовать весь потенциал искусственного интеллекта для улучшения жизни общества и предотвращения неблагоприятных последствий.

Источники и литература

1) Allen, C., Smit, I. and Wallach, W. (2005) “Artificial morality: Top-down, bottom-up, and hybrid approaches,” Ethics and information technology, 7(3), pp. 149–155. doi: 10.1007/s10676-006-0004-4.
2) Christian, B. (2023) The alignment problem: Machine learning and human values. New York, NY: WW Norton.
3) Gabriel, I. (2020) “Artificial intelligence, values, and alignment,” Minds and machines, 30(3), pp. 411–437. doi: 10.1007/s11023-020-09539-2.
4) Wiener, N. (1960) “Some Moral and Technical Consequences of Automation: As machines learn they may develop unforeseen strategies at rates that baffle their programmers,” Science (New York, N.Y.), 131(3410), pp. 1355–1358. doi: 10.1126/science.131.3410.1355.