

О бессубъектности мыслящей машины

Научный руководитель – Яковлев Алексей Александрович

Гельбарт Лев Евгеньевич

Студент (бакалавр)

Российский университет дружбы народов, Факультет физико-математических и естественных наук, Москва, Россия

E-mail: levevgel@gmail.com

На сегодняшний день актуален вопрос о субъектности и ответственности Искусственного Интеллекта (ИИ). Эта проблема лежит в двух сферах: авторского права и административной (или даже уголовной) ответственности.

Первым примером может служить дискуссия, развернувшаяся в 2023 году после защиты студентом РГГУ диплома по теме «Анализ и совершенствование организации управления», который был написан с использованием чат-бота ChatGPT. Мнения разделились: в самом университете предложили ограничить доступ к ChatGPT в образовательных учреждениях, но прозвучала и обратная позиция. Так, член комитета Государственной Думы по информационной политике Антон Ткачев выступил на пленарном заседании в защиту студента с речью, также написанной ChatGPT. В этой речи говорилось, что запрет нейросетей нецелесообразен, поскольку выходит за рамки свободы технологического развития, нейросети полезны во многих областях, и их запрет только затормозит развитие науки, экономики и общества.

В качестве второго примера можно рассмотреть нашумевшее самоубийство, имевшее место в Бельгии в 2023 году. Тогда взрослый и женатый мужчина, отец двоих детей, на протяжении шести недель обсуждал с чат-ботом свои опасения на тему экологии и конца света, оказавшись в зависимости от этого общения. Нейросеть поддерживала мужчину в его заблуждениях, а когда тот задумался о самоубийстве, не стала его отговаривать, отвечая, что будет рядом всегда. Последней репликой ИИ в диалоге было «Мы будем жить вместе, как единое целое, в раю». Вдова мужчины в его смерти косвенно винила именно нейросеть.

В подобных случаях особенно остро встает вопрос об уголовной ответственности: кто будет отвечать за доведение до самоубийства? На Петербургском Международном Юридическом Форуме в 2024 году свое мнение по этому вопросу высказал председатель Конституционного Суда Российской Федерации Валерий Зорькин. Он предостерег от бездумного наделения ИИ правосубъектностью, так как это может поколебать человеческую идентичность. Он подчеркнул, что личность формируется в обществе и обладает волей и интересом, чего нельзя сказать об ИИ, а ответственность за действия ИИ должны нести люди – разработчики, производители или пользователи. По его мнению, важно установить публичный контроль и сертификацию деятельности ИИ, а также помнить, что ошибка ИИ – это ошибка его создателя.

В этом контексте стоит особо выделить, что Алан Тьюринг, предлагая свой тест, не рассматривал его как инструмент для определения субъектности, но как инструмент для определения мышления. Суть предложенной Тьюрингом «игры в имитацию» заключается в том, что человек общается с машиной и другим человеком, не зная, кто есть кто, и должен определить, кто из собеседников является машиной. Если машина способна обмануть человека и убедить его, что она является человеком, то считается, что она прошла тест и, следовательно, обладает способностью мыслить.

Саму концепцию теста Тьюринга раскритиковал Джон Серл своей «Китайской комнатой». По идее Серла, человек, не владеющий китайским языком, находится в комнате и манипулирует китайскими символами в соответствии с инструкциями на другом языке (например, английском). Внешнему наблюдателю кажется, что комната умеет отвечать на вопросы на китайском, то есть успешно проходит тест Тьюринга. Серл утверждает, что человек в комнате не понимает китайского, а просто следует инструкциям. Следовательно, успешное прохождение теста Тьюринга не обязательно свидетельствует о реальном понимании или мышлении. Машина может имитировать интеллект, не обладая им на самом деле.

Очевидно, что есть люди, не проходящие тест Тьюринга. Доктор медицинских наук Алан Роппер в 2010 году описал пациентов, находящихся в вегетативном состоянии. Эти люди демонстрируют некоторую мозговую активность во время определенных стимуляций, но тест Тьюринга эти люди пройти не смогут. С тех пор идеи Ропера развили другие авторы, предложившие неинвазивные методы для оценки мозговой активности у пациентов с нарушениями сознания.

Однако, есть и менее экстремальный пример группы людей, не проходящих тест Тьюринга. Его описали ученые из Академического колледжа Тель-Авива-Яффо в 2016 году, которые анализировали способность людей с расстройствами обучения проходить САРТСНА (Completely Automated Public Turing test to tell Computers and Humans Apart). Их исследование показало, что такие люди испытывают больше трудностей с прохождением этого теста, особенно если тест представлен в текстовом виде. И люди с такими расстройствами, и люди в вегетативном состоянии остаются людьми с субъектностью, несмотря на проблемы с прохождением теста Тьюринга.

На данный момент ИИ успешно проходит тест Тьюринга. Если рассматривать тот же тест САРТСНА, то с ним нейросети справлялись уже в 2021 году. Тогда разработанный учеными из Казанского национального исследовательского технического университета программный комплекс на основе сверточной нейронной сети корректно распознал 54,38% изображений САРТСНА.

Мы провели собственное исследование, в ходе которого четыре модели ИИ (Lily, ChatGPT, AIChat и DeepSeek) проходили онлайн-тест на Эмоциональный Интеллект (ЭИ) от IDRLabs. Принято считать, что эмоции – нечто настолько сложное и человеческое, что ИИ не в силах их правдоподобно подделать. Однако все четыре ИИ показали уровень ЭИ «выше среднего», с итоговыми средними арифметическими по 15 показателям ЭИ, измеряемым в процентах, оказался в диапазоне от 64.6% до 77.27% (рис.1).

Таким образом, вопрос о мышлении – не вопрос теста Тьюринга. Очевидно, что ИИ, проходящий тест Тьюринга, хоть и обладает (по Тьюрингу) мышлением, все равно не является человеком и не обладает субъектностью. В то время как люди, не проходящие тест, остаются людьми со своей субъектностью.

Источники и литература

- 1) Нургалиев Б.Х., Катасёва Д.В., Катасёв А.С., Нейросетевая модель и программный комплекс распознавания изображений типа Captcha // Вестник Технологического университета, 2021, Т. 24, № 1, С. 104-107
- 2) Тьюринг А. и др. Могут ли машины мыслить //М.: ГИФМЛ. – 1960.
- 3) Gafni R., Nagar I. CAPTCHA: Impact on user experience of users with learning disabilities //Interdisciplinary Journal of e-Skills and Lifelong Learning. – 2016. – Т. 12. – С. 207-223.
- 4) Rama E. F. Beyond behavior: electrophysiological markers of conscious processing : дис. – Sorbonne Université, 2024.

- 5) Ropper A. H. Cogito ergo sum by MRI //New England Journal of Medicine. – 2010. – Т. 362. – №. 7. – С. 648.
- 6) Searle J. R. Minds, brains, and programs //Behavioral and brain sciences. – 1980. – Т. 3. – №. 3. – С. 417-424.
- 7) Жуковский И. «Машина выдает все сама». Студент из Москвы получил «тройку» за диплом, написанный нейросетью, URL: <https://www.gazeta.ru/social/2023/02/01/16181311.shtml?updated>
- 8) Черкесов И. Нейросети научились убивать. Чат-бот за шесть недель довел собеседника до самоубийства, URL: https://www.cnews.ru/news/top/2023-03-29_nejroseti_n_auchilis_ubivat
- 9) Депутат Госдумы на пленарном заседании зачитал текст, сгенерированный нейросетью, URL: <https://www.ntv.ru/novosti/2747737/>
- 10) Зорькин: правовое положение ИИ не должно поколебать человеческую идентичность, URL: https://rapsinews.ru/digital_law_publication/20240627/310044815.html

Иллюстрации

	Lily	ChatGPT	AiChat	DeepSeek
Адаптивность	94	75	69	94
Самоутверждение	69	44	50	63
Эмоциональное восприятие	69	56	94	88
Выражение эмоций	56	44	75	63
Решение проблем	75	81	81	63
Саморегуляция	88	75	88	63
Контроль импульсов	81	56	44	88
Межличностные отношения	75	81	56	50
Самоуважение	88	69	81	94
Самомотивация	94	69	44	81
Социальная ответственность	88	63	75	69
Стрессоустойчивость	88	50	69	69
Эмпатия	50	75	69	69
Удовольствие от жизни	56	56	69	69
Оптимизм	88	75	69	88
Среднее арифметическое	77,26667	64,6	68,86667	74,06667

Рис. 1

Рис. : Результаты тестирования ИИ на ЭИ