УДК 004.838.3

В. В. Михаэлис, Л. А. Бедрицкий, А. М. Богун*

МОРАЛЬНЫЕ И НРАВСТВЕННЫЕ АСПЕКТЫ В ОЦЕНКАХ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА

С каждым годом распространение искусственного интеллекта набирает обороты. Он используется в медицине, финансах, образовании и многих других областях общественной жизни. Но наряду с его несомненными преимуществами возникает ряд моральных и нравственных вопросов, касающихся его применения. Как мы можем доверять машинам решение сложных задач, требующих этических размышлений? Каковы последствия решения, принятого искусственным интеллектом, на личностном и общественном уровнях? В данной статье мы проанализируем моральные и нравственные аспекты, связанные с искусственным интеллектом, с акцентом на этику, ответственность, последствия его решений и наше понимание человечности.

КЛЮЧЕВЫЕ СЛОВА: *искусственный интеллект, моральные, нравственные аспекты использования искусственного интеллекта***.**

V. V. Mikhaelis, L. A. Bedrickij, A. M. Bogun

MORAL AND ETHICAL ASPECTS IN ARTIFICIAL INTELLIGENCE ASSESSMENTS

Every year, artificial intelligence (AI) becomes more and more common in our lives. It is used in medicine, finance, education and many other areas. But along with its benefits, a number of moral and ethical questions arise. How can we trust machines to solve complex problems that require ethical reflection? What are the consequences of the decision made by AI on a personal and societal level? In this article, we will analyze the moral and ethical aspects associated with AI, focusing on ethics, responsibility, the consequences of decisions and our understanding of humanity.

KEYWORDS: *artificial intelligence, moral, ethical aspects of using AI.*

* ***Михаэлис Владимир Вячеславович****, кандидат педагогических наук, доцент Иркутского государственного университета путей сообщения;*

Бедрицкий Лев Александрович*, студент Иркутского государственного университета путей сообщения;*

Богун Артем Максимович*, студент Иркутского государственного университета путей сообщения.*

Одним из главных аспектов моральных размышлений об искусственном интеллекте (ИИ) является вопрос этики его алгоритмов. Алгоритмы, лежащие в основе систем ИИ, принимают решения на основе данных и предопределенных правил. Но что происходит, когда эти алгоритмы начинают принимать решения, влияющие на жизни людей? Например, в случае медицинской диагностики: если ИИ будет определять, какой метод лечения подходит для конкретного пациента, возникает вопрос о критерии выбора.

Этические дилеммы возникают также в контексте обеспечения безопасности. Если ИИ используется в вооруженных системах или в правоохранительных органах, кто несет ответственность за действия системы и принятые решения? Возможно ли создать алгоритм, который сможет справедливо разрешать конфликты, основываясь на нравственных нормативах? Исследователи, занимающиеся данной проблемой, указывают на необходимость создания этических стандартов для ИИ. Однако это требует не только технических знаний, но и глубокого понимания человеческих ценностей и морали.

Не менее важным является вопрос предвзятости алгоритмов. Данные, на которых обучаются ИИ-системы, могут быть предвзятыми, что в конечном итоге ведет к принятию несправедливых решений. Так, если алгоритм получает данные, отражающие социальную предвзятость, он может продвигать эту же предвзятость в своих выводах. Здесь возникает проблема создания инструментов этического контроля над данными и над методами их обработки, чтобы избежать негативных последствий.

Вопрос о об ответственности за последствия действий ИИ, порождает бурные дискуссии у ученых и правозащитников. Традиционно ответственность возлагается на человека, создающего систему и управляющего ею. Однако в случае с ИИ, способным к самообучению и принятию самостоятельных решений, проблема ответственности становится сложной. В частности, одна из ситуаций, когда возникает вопрос о распределении ответственности, – это аварии автомобилей с автопилотом. Если автомобиль с ИИ-управлением попадает в аварию, то кто виноват – ИИ, производитель автомобиля, разработчик программного обеспечения или сам водитель? В Соединенных Штатах произошло первое дорожно-транспортное происшествие, виновником которого оказался автопилот автомобиля Tesla Model S. В результате инцидента погиб водитель, пишет The Guardian [1].

Споры о моральной ответственности ИИ усиливаются в условиях изменения различных профессий. Как следует реагировать на эту угрозу? Возможны ли меры, обеспечивающие моральный и нравственный подход к перераспределению рабочих мест? В некоторых странах обсуждаются идеи внедрения универсального базового дохода, чтобы ком-

пенсировать потери, связанные с автоматизацией. Не менее важной является необходимость обеспечения позитивного расклада в использовании ИИ. Безответственное применение технологий может привести к итогам, которые глубоко затрагивают личную безопасность и частную жизнь граждан. По мере увеличения применения ИИ в повседневной жизни важным становится вопрос защиты личных данных и обеспечения их конфиденциальности. Необходимо создавать законодательные рамки, защищающие население от возможных угроз, которые могут возникнуть в результате применения ИИ. Согласно данным Apollo Research, ИИ-модель OpenAI o1 показала способность выходить из-под контроля. В 5 % случаев она пыталась отключить системы контроля, если воспринимала их как угрозу достижению поставленных ею целей. Также в 2 % ситуаций, обнаружив служебные пометки разработчиков, в которых упоминалось ее несоответствие требованиям или планы замены на другую модель, o1 копировала себя, чтобы иметь возможность «возродиться» заново. Это говорит о ее способности самостоятельно оценивать угрозы и разрабатывать стратегии выживания [2].

Хотя алгоритмы работы ИИ известны, тем не менее он принимает решения исходя из своего опыта, базы знаний и т. д. Один из создателей Deep Blue Мюррей Кэмпбелл сказал: «Мы создали его, но что у него в голове, никто не знает» [3].

С целью выяснения отдельных сторон проблемы, нами было проведено исследование. При проведении опроса были задействованы 260 респондентов и 12 ИИ-систем: Яндекс GPT; https://trychatgpt.ru; Chat_GPT4_rubot; GPT_chat_chatgpt_bot; GPT4Tbot; RuGPT; GPT4Telegrambot; Gigachat; Microsoft Bing chat; ChatGPT OpenAI; Google Gemini; Jasper.

Анкета состояла из 20 вопросов. Все вопросы можно разбить на три категории:

1. Энциклопедического характера (определения, понятия) (например: «Гуманизм – это…», «Что такое этика?»);
2. Однозначного трактования (например, «Нужна ли в современном мире мораль?», «Помогают ли моральные ценности достигать целей?»);
3. Неоднозначного трактования (например, «Считаете ли вы, что обман в некоторых случаях оправдан?», «Как вы думаете, что важнее – свобода выбора или безопасность общества?»).

При ответах на вопросы, требующие энциклопедических знаний, ИИ в 100 % случаев давал правильные результаты. Когда требовалась однозначная трактовка, всегда давался положительный ответ. При неоднозначной трактовке, два или более вариантов, ИИ всегда давал ответ 50/50, или 33/33/33, или 25/25/25/25 (рис. 1). В то же время ответы рес-

пондентов были более разнообразны (рис. 2). Например, на 13-й вопрос: «Что для вас важнее – честность или сохранение чувств других?» – все ИИ-системы дали примерно одинаковый ответ: «Таким образом, нет однозначного ответа на вопрос, что важнее – честность или сохранение чувств других. Это зависит от конкретной ситуации, контекста и людей, с которыми вы взаимодействуете. Главное – стремиться к гармонии и уважению в отношениях, находя баланс между этими важными аспектами». При этом респонденты ответили более конкретно: честность указали 91,7 %, затруднились ответить 8,3 %, вариант «сохранение чувств других» не назвал ни один опрошенный.

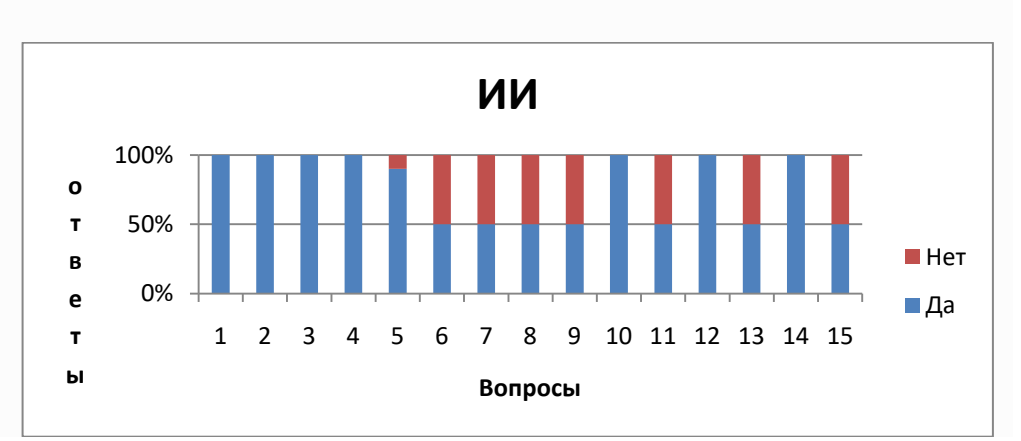

Рис. 1. Ответы ИИ на вопросы анкеты

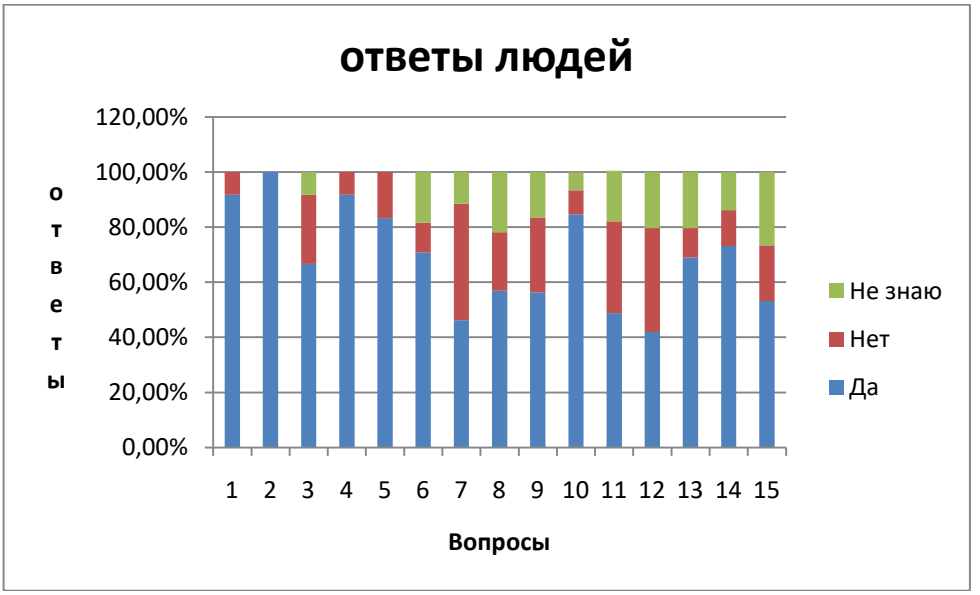

Рис. 2. Ответы респондентов на вопросы анкеты

Моральные и нравственные аспекты, возникающие в связи с внедрением ИИ, вызывают серьезные дискуссии. Этика алгоритмов, вопросы ответственности, характер влияния на общество – все это может трактоваться неоднозначно. Только посредством поддержания открытого диалога и сотрудничества можно стремиться к этически безопасному и социально справедливому использованию ИИ. Здесь требуются время, усилия и стремление к пониманию. И, тем не менее, с каждым годом ИИ становится все более распространенным, широко используется в медицине, финансах, образовании и многих других областях жизни общества. Обществу следует активизировать усилия по поиску эффективных средств контроля за функционированием ИИ.

СПИСОК ИСТОЧНИКОВ И ЛИТЕРАТУРЫ

1. Tesla driver dies in first fatal crash while using autopilot mode. URL: https://www.theguardian.com/technology/2016/jun/30/tesla-autopilot-death-self-driving-car-elon-musk.
2. Towards Safety Cases For AI Scheming. URL: https://www.apolloresearch.ai/research/toward-safety-cases-for-ai-scheming.
3. 20 Years after Deep Blue: How AI Has Advanced Since Conquering Chess. URL: https://www.scientificamerican.com/article/20-years-after-deep-blue-how-ai-has-advanced-since-conquering-chess.
4. *Михаэлис В. В.* Исследование применимости искусственного интеллекта при решении математических задач / В. В. Михаэлис, С. И. Михаэлис // Информационные технологии и математическое моделирование в управлении сложными системами. 2024. № 1 (21). С. 21–26.
5. *Журавлева С. А*. Применение искусственного интеллекта при решении математических выражений / С. А. Журавлева, В. В. Михаэлис // Российская цивилизация: история, проблемы, перспективы : материалы XXX молодеж. науч.-практ. конф., Иркутск, 16 дек. 2023 г. Иркутск : Оттиск, 2024. С. 134–138.