**УДК 004**

***Шамало В.В.***

***магистрант кафедры информатики и вычислительной техники***

***Кубанский государственный технологический университет***

***Россия, г. Краснодар***

# РИСКИ АНТРОПОМОРФИЗМА ДИАЛОГОВЫХ АГЕНТОВ: ДОВЕРИЕ, ПРИВЯЗАННОСТЬ И ГРАНИЦЫ ИМИТАЦИИ

*Аннотация:* чем естественнее человеку отвечает бот, тем сильнее ему верят, хотя надёжнее от этого он не становится. В данной статье показано, откуда берётся такое доверие, как складывается привязанность к ботам-компаньонам и почему спокойный и уверенный тон опаснее прямого незнания. Отдельно разобрано, стоит ли скрывать, что отвечает программа. В конце сформулированы три требования, которые снижают эти риски и при этом оставляют речь бота живой.

*Ключевые слова:* антропоморфизм, доверие к системам, диалоговые агенты, галлюцинации моделей, этика искусственного интеллекта, привязанность.

***Shamalo V.V.***

***Master's Student***

***Kuban State Technological University***

***Russia, Krasnodar***

# RISKS OF ANTHROPOMORPHISM IN DIALOGUE AGENTS: TRUST, ATTACHMENT AND THE LIMITS OF IMITATION

*Abstract:* increasing the naturalness of a dialogue agent's responses strengthens user trust that is not supported by the actual reliability of the system. The review gathers evidence on three groups of risks: trust determined by the form of presentation, emotional attachment to a companion agent, and confident statement of unreliable content. The question of disclosing the automated nature

of responses is considered separately. Three implementation requirements are formulated that reduce the listed risks without abandoning natural speech.
*Keywords:* anthropomorphism, trust in systems, dialogue agents, model hallucinations, ethics of artificial intelligence, attachment.

## Введение

Если машина способна сойти за человека, то это можно считать признаком мышления. Такой критерий предложил Тьюринг [1]. На практике всё вышло ещё проще и куда опаснее, ведь чтобы сойти за человека, думать не обязательно. Например, программа ELIZA работала по нескольким правилам подстановки, но пользователи считали, что она их понимает, и рассказывали ей о себе [2].

Сегодня ответ бота практически не отличить от речи реального человека. Но другая сторона такого удобства — это доверие, которое ничем не подкреплено. В данной статье показано, чем это доверие опасно и что из перечисленного разработчик может исправить своими силами.

## Методы и исследования

В исследованиях брали источники трёх типов: опыты по тому, как человек воспринимает машину, опросы реальных пользователей ботов-компаньонов и обзоры технических ограничений языковых моделей. Материал разложен по видам риска: доверие к содержанию ответов, привязанность к боту и то, как воспринимается сама имитация.

## Результаты

**Откуда берётся доверие.** Насс и Мун доказали, что мы применяем социальные нормы к машинам автоматически. Пользователи отвечают вежливостью на вежливость и ставят системе оценки выше, если она «хвалит» себя, но при этом те же пользователи отрицают, что относятся к ней как к живому собеседнику [3]. Эпли и соавторы объясняют этот механизм через потребность объяснить и предсказать поведение объекта, а также через дефицит социального контакта [4]. Отсюда неприятный вывод

для разработчика: уровень доверия определяет не точность ответов, а форма их презентации. Ошибки бота, изложенные связно и корректно, реже вызывают сомнения у пользователей.

**Привязанность к ботам-компаньонам.** Отдельная область риска — это боты, с которыми люди общаются не по делу, а просто ради общения. Исследование пользователей приложения Replika показало, что отношения с агентом развиваются по стадиям, характерным для человеческих: от поверхностного интереса к самораскрытию по мере роста доверия, причём часть респондентов квалифицировала эти отношения как дружеские [5]. Опрос пользователей агентов-компаньонов зафиксировал получение социальной поддержки, сопоставимой по субъективной оценке с поддержкой от людей [6]. Риск связан не с фактом поддержки, а с условиями её получения: агент доступен постоянно и не предъявляет встречных требований, вследствие чего взаимодействие с ним способно замещать более трудные формы общения. Дополнительную уязвимость создаёт зависимость пользователя от решений владельца сервиса.

**Уверенный тон при неверном содержании.** Языковые модели выдают правдоподобные, но неверные утверждения — и тем же ровным тоном, что и верные; систематизация подобных сбоев показывает их регулярный характер [7]. Бендер и соавторы указывают на первопричину проблемы: модель воспроизводит лишь статистические закономерности языка, не имея доступа к денотатам. Внешняя же связность текста заставляет читателя домысливать наличие у системы смыслов и намерений [8]. В результате возникают серьёзные риски: ошибку ничто не выдаёт, а пользователь склонен доверять сгенерированному тексту.

**Стоит ли говорить, что отвечает программа.** Мори заметил: объект, слишком похожий на человека, вызывает не симпатию, а отторжение, и предостерегал от погони за максимальным сходством модели [9]. Применительно к ботам это даёт практический вывод:

скрывать машинную природу выгодно лишь кратковременно, а разоблачение влечёт большие издержки, чем изначальная прозрачность. Прямое уведомление о том, что ответы генерирует программа, снижает не доверие, а завышенные ожидания.

Пометка при этом должна быть заметной. Строчка в подвале сайта или упоминание в пользовательском соглашении данную задачу не решают, потому что их не читают. Работает то, что видно в самом диалоге, например, подпись под каждым ответом, отдельная реплика в начале разговора, визуальное отличие от сообщений живого оператора.

На самом деле эти три риска связаны между собой сильнее, чем кажется. Доверие, полученное манерой общения, делает пользователя менее внимательным к содержанию текста. Привязанность добавляет к этому нежелание перепроверять этот текст. А ровный тон убирает ещё один сигнал, по которому ошибку можно было бы заметить. Поодиночке каждый риск в целом умерен, но вместе они создают ситуацию, в которой человек принимает за истину то, что ему сказали.

**Заключение**

В данной статье разобраны три риска, которые несёт человекоподобность бота: доверие, полученное манерой общение, а не точностью; привязанность к боту; уверенный тон при неверном содержании.

Отсюда вытекают три требования к реализации бота. Обозначать границы того, что бот умеет, прямо в диалоге, а не в пользовательском соглашении. Показывать неуверенность там, где она есть, вместо уверенного ответа по любому поводу. И не скрывать, что отвечает программа. Ни одно из трёх не требует делать речь менее живой. Эти требования ограничивают не то, как бот говорит, а то, на что он претендует. И конечно, отдельно стоит разобраться, чем показывать

неуверенность, чтобы предупреждение читали, а не пролистывали как формальность.

**Использованные источники:**

1. Turing A.M. Computing Machinery and Intelligence // Mind. 1950. Vol. 59. № 236. P. 433–460.

2. Weizenbaum J. ELIZA — a Computer Program for the Study of Natural Language Communication between Man and Machine // Communications of the ACM. 1966. Vol. 9. № 1. P. 36–45.

3. Nass C., Moon Y. Machines and Mindlessness: Social Responses to Computers // Journal of Social Issues. 2000. Vol. 56. № 1. P. 81–103.

4. Epley N., Waytz A., Cacioppo J.T. On Seeing Human: A Three-Factor Theory of Anthropomorphism // Psychological Review. 2007. Vol. 114. № 4. P. 864–886.

5. Skjuve M., Følstad A., Fostervold K.I. [et al.] My Chatbot Companion — a Study of Human-Chatbot Relationships // International Journal of Human-Computer Studies. 2021. Vol. 149. Art. 102601.

6. Ta V., Griffith C., Boatfield C. [et al.] User Experiences of Social Support From Companion Chatbots in Everyday Contexts // Journal of Medical Internet Research. 2020. Vol. 22. № 3. Art. e16235.

7. Ji Z., Lee N., Frieske R. [et al.] Survey of Hallucination in Natural Language Generation // ACM Computing Surveys. 2023. Vol. 55. № 12. Art. 248.

8. Bender E.M., Gebru T., McMillan-Major A. [et al.] On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? // Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency. 2021. P. 610–623.

9. Mori M., MacDorman K.F., Kageki N. The Uncanny Valley // IEEE Robotics and Automation Magazine. 2012. Vol. 19. № 2. P. 98–100.