DOI: 10.51790/2712-9942-2021-2-3-1

О ПРОБЛЕМЕ ДОВЕРИЯ К ТЕХНОЛОГИЯМ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА

В. Б. Бетелин
Федеральное государственное учреждение «Федеральный научный центр Научно-исследовательский институт системных исследований Российской академии наук», г. Москва, Российская Федерация, ORCID: http://orcid.org/0000-0001-6646-2660, betelin@inbox.ru

Для цитирования: Бетелин В. Б. О проблеме доверия к технологиям искусственного интеллекта. *Успехи кибернетики*. 2021;2(3):6–7. DOI: 10.51790/2712-9942-2021-2-3-1.

CAN WE TRUST THE ARTIFICIAL INTELLIGENCE TECHNOLOGIES?

V. B. Betelin
Federal State Institution "Scientific Research Institute for System Analysis of the Russian Academy of Sciences", Moscow, Russian Federation, ORCID: http://orcid.org0000-0001-6646-2660, betelin@inbox.ru

Cite this article: Betelin V. B. Can We Trust the Artificial Intelligence Technologies? *Russian Journal of Cybernetics*. 2021;2(3):6–7. DOI: 10.51790/2712-9942-2021-2-3-1.

Национальной стратегией развития искусственного интеллекта до 2030 года (далее — Стратегия), утвержденной Указом Президента Российской Федерации от 10.10.2019 № 490, определены цели и основные задачи развития искусственного интеллекта в Российской Федерации. Вместе с тем в разделе II Стратегии отмечается, что «***отсутствие понимания*** *того, как искусственный интеллект* ***достигает результатов****, является одной из причин* ***низкого уровня доверия*** *к современным технологиям искусственного интеллекта и может стать* ***препятствием для их развития».***

В разделе IV Концепции развития регулирования отношений в сфере технологий искусственного интеллекта и робототехники до 2024 года (далее — Концепция) также констатируется, что «***одним из основных препятствий*** *для расширения применения систем с использованием искусственного интеллекта и робототехники является* ***отсутствие достаточной степени доверия*** *к ним со стороны общества»* [1]. Преодоление этого препятствия авторы Концепции видят во введении **регуляторных ограничений на применение** систем ИИ и робототехники, для которых отсутствует **понимание** того, как эти системы достигают результата. То есть, вообще говоря, тем самым постулируют **непознаваемость систем ИИ**, как теми, кто создает эти системы, так и теми, кто их использует. Однако это противоречит одному из основных принципов Стратегии (раздел III, п. 19в) — *«прозрачности: объяснимости работы искусственного интеллекта и процесса достижения им результата».*

Отсутствие понимания того, как функционируют системы ИИ, связано в нашей стране, прежде всего, с использованием для их создания разработанных за рубежом библиотек полуэмпирических нейросетевых алгоритмов, доступных в сети Интернет. Для этих алгоритмов **отсутствуют** какие-либо **теоретические обоснования устойчивости и сходимости**, что **не гарантирует** системам ИИ на их основе **получение надежного результата** и, кроме того, делает их уязвимыми для кибератак. Об этом свидетельствуют, например, многочисленные публикации об авариях и катастрофах с автомобилями TESLA и Toyota, управляемыми системами ИИ на основе таких алгоритмов. По сути дела, именно **отсутствие математических обоснований** функционирования систем ИИ и **является причиной низкого уровня доверия** к ним.

Согласно Стратегии (раздел I, п. 5а) ***«искусственный интеллект*** *— комплекс технологических решений, позволяющий . . . получать при выполнении* ***конкретных задач*** *результаты,* ***сопоставимые****, как минимум, с* ***результатами интеллектуальной*** *деятельности* ***человека****».*

По сути дела, это определение искусственного интеллекта в Стратегии **основывается на предположении о возможности** создания **цифровых двойников человека**, которые интеллектуально и профессионально **сопоставимы со своими реально существующими прототипами или даже превосходят их**. Таких, например, как человека-водителя и человека-пилота, управляющих **реальными** автомобилями и самолетами, двигающимися в **реальных** дорожных и воздушных пространствах. Однако проблемы создания цифровых двойников и водителя, и пилота относятся к классу **некорректных**

задач, решение которых либо отсутствует, либо множественно, либо неустойчиво и сводится к проблеме их **регуляризации**. То есть к «подгонке» специалистами автомобильных и авиационных КБ и НИИ, водителями и пилотами, в рамках стендовых и натурных испытаний, универсальных математических методов аппроксимации и оптимизации под специфику этих задач. Конечно, с доказательством сходимости и устойчивости этих «подогнанных» методов и, в конечном счете, к оценке их адекватности первоначальной цели — созданию цифрового водителя и пилота и тем самым — к **обоснованию ограничений** возможности **их существования**. Такой «подгонкой» универсальных, **но полуэмпирических** нейросетевых алгоритмов является, по сути дела, процесс обучения нейросети для решения какой-либо конкретной задачи. Однако без доказательства сходимости и устойчивости.

Решение этих математически сложных проблем управления беспилотным транспортом на основе ИИ обеспечит их ***прозрачность****: объяснимость работы искусственного интеллекта и процесса достижения им результата* и тем самым — необходимый уровень доверия к этим системам ИИ со стороны общества. В том числе доверие к планируемой Минтрансом к 2024 г. коммерческой эксплуатации беспилотных грузовиков по трассе М-11, соединяющей Москву и Санкт-Петербург [2].

Наш журнал начинает цикл публикаций, отражающих различные подходы, в том числе и дискуссионные, к решению проблемы доверия к технологиям искусственного интеллекта.

ЛИТЕРАТУРА

1. *Концепция регулирования отношений в сфере технологий искусственного интеллекта и робототехники до 2024 года*. Утверждена распоряжением Правительства Российской Федерации от 19.08.2020 № 2129-р.
2. Мальгавко С. *Цифровая трасса: как грузовики-беспилотники помогут снизить число аварий*. Режим доступа: https://национальныепроекты.рф/news/tsifrovaya-trassa-kak-gruzoviki-bespilotniki-pomogut-snizit-chislo-avariy?utm%5C_source=Yandex%5C_Net%5C&utm%5C_medium=CPC%5C&utm%5C_content=All18-55drivers%5C&utm%5C_campaign=bezopasnye-i-kachestvennye-avtomobilnye-dorogi%5C&yclid=4862431727482615636