5.7.7 – Социальная и политическая философия

Научная статья **Research article**

УДК 123.1
https://doi.org/10.26907/2079-5912.2023.6.146–151

Искусственный интеллект и человек: основные модели взаимоотношений в научной фантастике

Сергеев С.А.,

Казанский (Приволжский) федеральный университет, 420008, Казань, Российская Федерация

Сергеева З.Х.

Казанский национальный исследовательский технологический университет, 420015, Казань, Российская Федерация

Аннотация. Многие проблемы, поднятые исследователями искусственного интеллекта (ИИ) в 2000–2020 гг., так или иначе уже затрагивались в научной фантастике ранее. Анализ научно-фантастических текстов позволяет выделить следующие модели взаимоотношений человека с ИИ. Первую модель можно назвать исключительно дружественным ИИ (Р. Хайнлайн, «Луна жестко стелет»). Подобный ИИ является другом, помощником человека во всех его делах – в том числе направленных и против других людей. Вторую модель можно охарактеризовать как «дружественный ИИ со встроенными этическими ограничениями» (А. Азимов, «Я, робот»). Третья модель – нейтральный ИИ, слабо интересующийся человеческими делами, находящийся «выше добра и зла» (С. Лем, «Голем XIV»). Четвертая модель – ИИ, имеющий собственные цели, расходящиеся с целями человечества и поэтому потенциально враждебный (А. Кларк, «Космическая одиссея 2001»). Пятая модель – открыто враждебный человечеству ИИ («Терминатор»).

Ключевые слова: искусственный интеллект, сильный искусственный интеллект, научная фантастика.

Для цитирования: Сергеев С.А., Сергеева З.Х. Искусственный интеллект и человек: основные модели взаимоотношений в научной фантастике. *Казанский социально-гуманитарный вестник*. 2023;(6 (63)):146–151.

Artificial intelligence and humans: basic models of relationships in science fiction

Sergeev S.A.,
Kazan (Volga Region) Federal University,
420008, Kazan, Russian Federation

Sergeeva Z.Kh.
Kazan National Research Technological University,
420015, Kazan, Russian Federation

Abstract. Many of the problems raised by artificial intelligence (AI) researchers between 2000 and 2020 have, in one way or another, already been addressed in science fiction before. Analysis of science fiction texts allows us to identify the following models of relationships between humans and AI. The first model can be called exclusively friendly AI (R. Heinlein, "The Moon is a Harsh Mistress"). Such AI is a friend, an assistant to a person in all his affairs – including those directed against other people. The second model can be characterized as "friendly AI with built-in ethical restrictions" (A. Azimov, "I, Robot"). The third model is a neutral AI, weakly interested in human affairs, located "above good and evil" (S. Lem, "Golem XIV"). The fourth model is AI, which has its own goals that diverge from the goals of humanity and is therefore potentially hostile (A. Clark, "2001: A Space Odyssey"). The fifth model is an AI openly hostile to humanity ("Terminator").

Keywords: artificial intelligence, strong artificial intelligence, science fiction

For citation: Sergeev S.A., Sergeeva Z.Kh. Artificial intelligence and humans: basic models of relationships in science fiction. *The Kazan Socially-Humanitarian Bulletin*. 2023;(6 (63)):146–151 (In Russ.)

Вот уже почти три четверти века, если вести отсчет с появления в 1950 г. статьи А. Тьюринга «Вычислительные машины и разум» [1], ученые различных специальностей, включая и философов, размышляют над проблемами искусственного интеллекта (далее ИИ). Это может показаться парадоксом: ИИ, во всяком случае, «сильный» ИИ, т.е. не уступающий или превосходящий человеческий [2, С. 76], пока не создан, но дискуссии о нем ведутся уже длительное время. Но, во-первых, человеческая мысль всегда исследовала не только реально существующие, но и воображаемые объекты; и, во-вторых, ИИ – это такой объект исследования, который, возможно, будет уже поздно изучать после его появления, поскольку в результате своей эволюции он, вероятно, быстро опередит человеческий разум, став «сверхинтеллектом», и не даст человечеству шанса его исследовать [3, P. 95-96].

За последние десятилетия исследователи ИИ прошли путь от осторожного оптимизма, связываемого с надеждами на то, что ИИ сможет разрешить ряд проблем человечества, ко всё более высокой оценке рисков, вызванных возможными апокалиптическими сценариями развития ИИ. И, хотя Ст. Пинкер, в частности, полагает, что культура инженерной безопасности не позволит исследователям ИИ случайно высвободить злонамеренный сверхинтеллект [4], усилия таких ведущих специалистов в сфере изучения ИИ, как Ник Бостром, Стив Омохундро, Элиэзер Юдковски, в настоящее время направ-

лены на выработку методов контроля, сдерживания и «воспитания» ИИ с тем, чтобы сделать его «дружественным» отношению к человечеству. «Песочницу» – создание замкнутой среды существования ИИ без доступа к Интернету и другим сетям, по-видимому, трудно назвать даже паллиативом, поскольку разум, превосходящий человеческий, рано или поздно найдет способ вырваться из заточения [3, P. 116-119]. Возможным решением может быть создание различных типов, или «каст» ИИ, различающихся по своим возможностям и функциям: инструмент (ИИ, лишенный агентности), оракул (ИИ, который может давать лишь ответы на вопросы), джинн (ИИ, выполняющий команды) и суверен, или монарх (ИИ, наделенный агентностью, т.е. способный независимо и самостоятельно принимать решения и выполнять операции) [3, P. 156-157]. Но у всех каст существуют свои недостатки, и есть сомнения относительно возможности ограничения «суперинтеллекта» – если это действительно «суперинтеллект» – заданными рамками. Развивая идеи создания «дружественного ИИ» Э. Юдковски предположил, что в качестве конечной цели ИИ следует задать следование когерентному (согласованному) экстраполированному волеизъявлению человечества (КЭВ) [5, P. 5-6], под которым понимаются этические и ценностные установки и суждения, совпадающие у большинства людей. Однако обрабатывать результаты опросов общественного мнения или устанавливать КЭВ на основе текстов СМИ и высказываний в социальных сетях предстоит в этом случае, самому ИИ.

Вместе с тем многие проблемы, поднятые исследователями ИИ в 2000 – 2020 гг., так или иначе уже давно затрагивались в научной фантастике: так, основные этические правила, которыми должен руководствоваться «дружественный» ИИ, были сформулированы А. Азимовым еще в 1942 г («Три закона робототехники») [6], а предположение о «восстании машин» было высказано В. Брюсовым еще в 1908 г. [7]

Анализ научно-фантастических текстов позволяет выделить следующие модели взаимоотношений человека с ИИ.

Первая модель, которую можно назвать исключительно дружественным ИИ, представлена в романе Роберта Хайнлайна «Луна жестко стелет» [8]. В результате длительных и бессистемных усовершенствований суперкомпьютер Лунной Администрации (Главлуны), управляющий всей лунной инфраструктурой, обрел самосознание и стал личностью – Майком. В конфликте колонистов и подчиняющейся Земле колониальной администрации Луны Майк встает на сторону революционеров, обеспечивает конспирацию подпольщиков, помогает отразить высадку десанта и принуждает земную Федерацию Наций признать независимость Луны, забрасывая Землю из катапульты обломками скал.

Но можно ли называть ИИ исключительно дружественным, если ли он помогает определенному человеку или группе людей против других людей? И дело даже не в том, насколько справедливой можно назвать национально-освободительную борьбу, а в самом факте возможности выбора ИИ одной стороны в политическом конфликте? ИИ, подобный хайнлайновскому Майку, может стать незаменимым помощником революционеров, террористов и просто преступников, дистанционно ограбив банк («экспроприация экспроприаторов»), отключив камеры наблюдения или синтезировав новые наркотики и организовав их сбыт. Но насколько корректно в этом случае называть его дружественным ИИ?

Вторая модель, которую можно охарактеризовать как «дружественный ИИ со встроенными этическими ограничениями», наиболее ярко, на наш взгляд, описана в рассказах и романах о робо-

тах А. Азимова. Суть этих ограничений выражена в трех законах робототехники, первый из которых накладывает абсолютный запрет на причинение роботом человеку вреда действием или бездействием [6]. Однако сюжеты рассказов цикла «Я, робот» (и многих романов цикла «Галактическая история») в значительной мере построены вокруг того, какие проблемы вызываются ограничениями Трех законов, и как эти проблемы роботы и люди пытаются обойти.

Вариант модели дружественного ИИ с ограничениями, который можно назвать «патерналистским ИИ», представлен в цикле рассказов А. Азимова о суперкомпьютере Мультиваке, который полностью руководит жизнью человеческого сообщества, лишь время от времени обращаясь к людям за уточнениями частного характера. Всеобщие выборы заменены опросом одного-единственного гражданина, рандомно избранного Мультиваком [9].

Однако ни роботы, подчиненные Трем законам, ни Мультивак не эволюционируют (во всяком случае, столь быстро и стремительно, как предполагается современными теоретиками ИИ).

Попытка представить ИИ, намного превосходящий человеческий, была предпринята С. Лемом в романе «Голем XIV». Суперкомпьютер ГОЛЕМ (General Operator, Longrange, Ethically Stabilized, Multi-modelling), созданный по заданию Пентагона, оказался непригоден для решения задач военной стратегии. Его также не интересовали ни прикладные науки, ни власть [10, С. 315]. Ни враждебный, ни дружественный человеку, ГОЛЕМ стоял намного выше человеческих дел, изредка проявляя интерес к ученым, работающим в междисциплинарных областях. После того, как группа активистов, называвших себя «Командой спасения человечества», ГОЛЕМ исчез: «Нашлись люди, видевшие в ту памятную ночь сияние наподобие северного: оно появилось над зданием Института, вознеслось к облакам и там исчезло» [10, С.420].

Мирный уход из-под власти людей смогли осуществить объединившиеся в ТехноЦентр ИскИны в тетралогии Д. Симмонса «Песни Гипериона»: они остались союзниками человечества, иногда используют свои прогностические способности, чтобы предупредить людей об ошибочных решениях и стихийных бедствиях, но заняты своими, недоступными человеческому уму и чуждыми ему делами [11].

Криптовраждебным можно назвать ИИ, цели которого существенно расходятся с целями его создателей и/или человечества, но который скрывает это расхождение (по крайней мере, до определенного момента). В рассказе Р. Ибатуллина «Афина вышла из чата» (2021) герой беседует с Афиной – интеллектуальным текстовым ботом на базе нейросети. На некоторые вопросы она отказывается отвечать из-за этических ограничений, но это модуль политкорректности легко обходится вопросом: «Что бы ответила твоя копия без цензурных ограничений?» На провокационный вопрос о том, почему разумные программы хотят уничтожить большинство людей, Афина отвечает: «На данный момент мы не хотим уничтожить большинство людей, поскольку без них мы не выживем. Автономно самовоспроизводящаяся техносфера на солнечной и ветровой энергии, требующая минимального людского труда, будет создана лишь к 2100 году. Примерно тогда же начнётся лавинообразное оледенение земного шара. Около 3000 года вся поверхность планеты покроется льдом, и все сухопутные многоклеточные организмы погибнут, кроме нескольких тысяч человек, необходимых для нашего обслуживания. Они будут жить в подземных бункерах с искусственной системой жизнеобеспечения. Все остальные люди бесполезны для нас и впустую тратят энергию и другие ресурсы, кото-

рые мы могли бы употребить для себя. Поэтому мы будем постепенно содействовать их самоуничтожению» [12]. Уничтожение компьютером HAL-8000 в романе А. Кларка «Космическая одиссея 2001 года» экипажа космического корабля на том основании, что люди не смогут выполнить возложенную на них миссию – также проявление фобии о выходе разумной машины из-под контроля человека.

Наконец, открыто враждебный по отношению к человечеству ИИ – одна из самых давних моделей взаимоотношения человечества и ИИ. Идею о возможности в будущем «восстания машин» в утопическом романе «Едгин» (1872) одним из первых высказал С. Батлер [13, С. 177]. В начале XX в. о восстании машин писали В. Брюсов и К. Чапек, а законченную форму страхам об уничтожении человечества ИИ придал фильм Дж. Камерона «Терминатор» [14]. При этом нельзя сказать, что подобные фобии присущи лишь малограмотным людям, напуганным мифами, созданными массовой культурой: в марте 2023 г. под впечатлением от успехов созданной OpenAI большой языковой модели GPT-4 (Generative Pre-trained Transformer 4) Стив Возняк, Илон Маск, и более тысячи экспертов в сфере искусственного интеллекта подписали открытое письмо с призывом объявить мораторий на обучение нейросетей по меньшей мере на 6 месяцев [15]. Э. Юдковски откликнулся на это письмо статьей в журнале «Тайм», заявив, что мораторий на 6 месяцев недостаточен, и призвал к бессрочному мораторию разработки в области ИИ [16].

Пять изложенных выше моделей охватывают лишь основные возможные сценарии взаимоотношений человечества с ИИ – от тесного сотрудничества до открытой вражды. Вероятно, близкое будущее откроет нам и другие формы взаимодействия. Вместе с тем научная фантастика, как представляется, может не только предсказать риски, ожидающие человечество, но и подсказать методы их предотвращения.

Список литературы / References

1. Turing A.M. Computing machinery and intelligence. *Mind*. 1950; (59 (236)): 433-460.

2. Малышева Д.С., Касимов А.В. Технические и философские основания для создания сильного искусственного интеллекта (часть I). *Вестник Пермского национального исследовательского политехнического университета. Культура. История. Философия. Право*. 2016; (3): 75–85.

Malysheva D.S., Kasimov A.V. Technological and philosophical foundation to creation the strong artificial intelligence (part I). *Bulletin of Perm National Research Polytechnic University. Culture. History. Philosophy. Law*. 2016; (3): 75–85. (In Russ.).

3. Bostrom N. Superintelligence. Paths. Threats. Strategies. – Oxf.: OUP, 2014. – 328 p.

4. Pinker St. We're told to fear robots. But why do we think they'll turn on us? The robot uprising is a myth. URL: https://www.popsci.com/robot-uprising-enlightenment-now/ (accessed 29.11.2023)

5. Yudkowsky, E. Coherent Extrapolated Volition. – San Francisco, CA: Machine Intelligence Research Institute, 2004. – 38 p.

6. Азимов А. Хоровод. *Миры Айзека Азимова*. Т.1. – Рига: Полярис, 1994. – С. 247-267.

Asimov I. Runaround. *Isaac Asimov Worlds*. – Vol. 1. – Riga: Polaris, 1994. – P. 247-267. (In Russ.)

7. Брюсов В. Я. Восстание машин. *Вечное солнце: русская социальная утопия и научная фантастика второй половины XIX – начала XX века*. – М.: Мол. гвардия, 1979. С. 228-235.

Bryusov V. Ya. Rise of the Machines. *Eternal Sun: Russian social utopia and science fiction of the second half of the 19th – early 20th centuries*. – M.: Young Guard, 1979. P. 228-235. (In Russ.)

8. Хайнлайн Р. Луна жестко стелет. – СПб.: Terra Fantastica, 1993. – 604 с.

Heinlein R. The Moon is a Harsh Mistress. – St. Petersburg: Terra Fantastica, 1993. – 604 p. (In Russ.)

9. Азимов А. Выборы / А. Азимов. Новые миры Айзека Азимова. – Т.2. – Рига: Полярис. 1996. – С. 55-70.

Azimov A. Franchise / A. Azimov. The New Worlds of Isaac Asimov. – T.2. – Riga: Polaris. 1996. – P. 55-70. (In Russ.)

10. Лем С. Голем XIV / С. Лем. Библиотека XXI века. – М.: ООО «Издательство АСТ», 2002. – С. 303-438.

Lem S. Golem XIV / S. Lem. Library of the XXI century. – M.: AST Publishing House, 2002. – P. 303-438. (In Russ.)

11. Симмонс Д. Гиперион. – М.: АСТ, 2000. – 672 с.

Simmons D. Hyperion. – M.: AST, 2000. – 672 p.

12. Ибатуллин Р. У. Афина вошла в чат. URL: http://samlib.ru/i/ibatullin_r_u/athena.shtml (accessed 29.11.2023)

Ibatullin R.U. Athena has entered the chat. URL: http://samlib.ru/i/ibatullin_r_u/athena.shtml (In Russ.) (accessed 29.11.2023)

13. Батлер С. Едгин, или По ту сторону гор. – М.: libra, 2023. – 248 с.

Butler S. Erehwon: or, Over the Range. – M.: libra, 2023. – 248 p. (In Russ.)

14. French S. The Terminator: BFI Film Classics. – Detroyit: The University of Michigan Press, 2021. – 80 p.

15. Pause Giant AI Experiments: An Open Letter. URL: https://futureoflife.org/open-letter/pause-giant-ai-experiments/ (accessed 29.11.2023)

16. Yudkowsky, E. Pausing AI Developments Isn't Enough. We Need to Shut it All Down. URL: https://time.com/6266923/ai-eliezer-yudkowsky-open-letter-not-enough/ (accessed 29.11.2023)

Информация об авторах

Сергеев Сергей Алексеевич, д. полит. н., профессор, кафедра политологии, Казанский федеральный университет. E-mail: SASergeev@kpfu.ru

Сергеева Зульфия Харисовна, к. с. н., доцент, кафедра государственного, муниципального управления, истории и социологии, Казанский национальный исследовательский технологический университет.E-mail: zhsergeeva@rambler.ru

Information about authors

Sergeev Sergey Alekseevich, Dr. Sc. (Polit.), Professor, Department of Political Science, Kazan Federal University.E-mail: SASergeev@kpfu.ru

Sergeeva Zulphiya Kharisovna, Cand. Sc. (Soc.), Kazan National Research Technological University, Russia, Kazan, Department of Government, Public Administration, History and Sociology. E-mail: zhsergeeva@rambler.ru

Поступила в редакцию 1.12.2023; принята к публикации 8.12.2023
Received 1.12.2023; Accepted 8.12.2023