

УДК 101.2

DOI: 10.32326/2618-9267-2022-5-4-94-108

ФИЛОСОФСКИЙ ДИАЛОГ В ПРОЦЕССЕ ЧЕЛОВЕКО-МАШИННОЙ КОММУНИКАЦИИ. КАК ДЕЙСТВУЕТ «ЦИФРОВОЙ СОКРАТ»?

Зарапин Олег Викторович – кандидат философских наук, доцент. Крымский федеральный университет имени В.И. Вернадского. Российская Федерация, 295007, Республика Крым, г. Симферополь, просп. Академика Вернадского, д. 4;
e-mail: zaraoleg@yandex.ru

В статье ставится проблема специфики философского диалога, осуществляемого в процессе человеко-машинной коммуникации. Выдвигается тезис, что практика диалога в философской культуре в своем истоке имеет значение рефлексивной деятельности, релевантной межличностной коммуникации. Особенность практики философского диалога представлена с точки зрения способа коммуникативного воздействия, репрезентируемого формой сократической беседы, в рамках которой рефлексия есть деятельность, выражаемая понятием «метанойя» (греч. μετάνοια) в значении умоперемены и нацеливающая человека на достижение «превосходства над собой». Философский диалог с машиной рассматривается с точки зрения способности машины в процессе коммуникации воздействовать на человека, инициировать рефлексию как опыт умоперемены, аналогичный сократической метанойе. Выделяются три модели коммуникативного взаимодействия человека и машины: 1) модель «тренажер»; 2) модель «собеседник»; 3) модель «философский собеседник». В рамках модели «философский собеседник» выделены три сценария человеко-машинного диалога (сопротивление, слияние, сотрудничество), в которых рефлексивный опыт умоперемены выражается идеей, что разум не имманентен личности и нечеловеческий разум возможен. Делается вывод, что человеко-машинная коммуникация задает новый формат философской культуры диалога. Особенностью этого формата является невозможность для человека конвертировать воздействие, которое оказывает машина, в усвоенный человеком способ действия. Такая особенность позволяет констатировать, что машина способна побуждать человека к рефлексии, но неспособна научить его рефлексивной деятельности.

Ключевые слова: философский диалог, сократическая беседа, рефлексия, метанойя, коммуникация, человеко-машинная коммуникация

Цитирование: Зарапин О.В. Философский диалог в процессе человеко-машинной коммуникации. Как действует «Цифровой Сократ»? // Цифровой ученый: лаборатория философа. 2022. Т. 5. № 4. С. 94-108. DOI: 10.32326/2618-9267-2022-5-4-94-108

PHILOSOPHICAL DIALOGUE IN THE PROCESS OF HUMAN-MACHINE COMMUNICATION. HOW DOES THE “DIGITAL SOCRATES” ACT?

Oleg V. Zarapin – PhD in Philosophy, associate professor. V.I. Vernadsky Crimean Federal University. 4 Akademika Vernadskogo Avenue, Simferopol, Crimea, 295007, Russian Federation; e-mail: zaraoleg@yandex.ru

The article raises the problem of the specifics of the philosophical dialogue carried out in the process of human-machine communication. The author puts forward the thesis that the practice of dialogue in philosophical culture in its source has the meaning of reflexive activity, relevant to interpersonal communication. The peculiarity of the practice of philosophical dialogue is demonstrated from the viewpoint of the communicative influence method, represented by the form of Socratic interlocution. There, reflection is an activity expressed by the concept of metanoia (μετάνοια) in the meaning of mental-change and aiming a person at achieving “superiority over himself”. The philosophical dialogue with the machine is considered from the viewpoint of the machine's ability to influence a person in the process of communication and initiate reflection as an experience of mental-change, analogous to Socratic metanoia. There are three models of communicative interaction between a person and a machine: (1) the “simulator” model, (2) the “interlocutor” model, and (3) the “philosophical interlocutor” model. Within the framework of the “philosophical interlocutor” model, the study identifies three scenarios of a human-machine dialogue (resistance, merger, and cooperation), in which the reflexive experience of mental-change is expressed by the idea that the mind is not immanent to the personality and a non-human mind is possible. The author concludes that the peculiarity of human-machine communication sets a new format for the philosophical culture of dialogue. A feature of this format is the impossibility for a person to convert the impact that a machine has into a way of action learned by a person. This feature allows us to state that the machine is able to induce a person to reflection, but is not able to teach him or her a reflexive activity.

Keywords: philosophical dialogue, Socratic interlocution, reflection, metanoia, communication, human-machine communication

For References: Zarapin, Oleg. 2022. Philosophical dialogue in the process of human-machine communication. How does the “Digital Socrates” act?, *The Digital Scholar*:

Philosopher's Lab, 5 (4): 94-108. DOI: 10.32326/2618-9267-2022-5-4-94-108
(In Russian)

Развитие информационных технологий, их внедрение в различные сферы жизни – это значимый тренд современной культуры. «Четвертая промышленная революция» (Индустрия 4.0) задает горизонты развития социокультурных процессов в направлении все более тесного взаимодействия человека и машины. Взаимодействие с компьютером накладывает свой отпечаток на человеческую психику, т. е. форматирует человеческие способности восприятия и действия, порождая такие качества, как, к примеру, клиповость, геймифицированность, серийность, сетевая зависимость и т. д. У этого процесса есть обратная сторона – информационные технологии становятся все более удобными и незаметными в обращении, а сам компьютерный интерфейс – все более антропоморфным, вплоть до функции виртуального друга. Диалоговый интерфейс делает компьютер собеседником, способным поддержать разговор в том числе и на философскую тему.

Если прибавить к этому то, что информационные технологии являются продуктом массового потребления, исконная для европейской культуры и наследуемая из античной традиции идея философского общения как самопознания, а вместе с ней и образ философа рискуют стать архаикой, потерявшей связь с технологически «прокачанной» действительностью.

Машина, которая имитирует позицию философствующего субъекта, хотя сейчас и имеет вид экспериментальной разработки, все же есть не просто эпифеномен, свидетельствующий о технологическом прорыве. Подчеркнем: это феномен, свидетельствующий об изменениях, затрагивающих основу европейской философской культуры. Традиционно эту основу выражал тезис о том, что философская мысль осуществляется в процессе межличностной коммуникации. Иными словами, чтобы осуществить рефлекссию, нам необходим Другой, необходимо вступить с ним в диалог. Сегодня этот тезис получает новое звучание, поскольку, как выясняется, вступить в философский диалог можно не только с человеком, но также и с машиной.

Такая возможность – это не предположение, не фантастический сюжет и не прогноз будущего. В процессе разработки диалоговой архитектуры интерфейса уже созданы интеллектуальные системы, демонстрирующие речевую деятельность, которую человек, взаимодействующий с машиной, распознает как деятельность рефлексивную. Является ли она таковой в действительности, т. е. мыслит ли машина – это другой вопрос; важно то, что машина оказывает воздействие на человека, побуждая его к рефлексии. Как воздействует машина, и каким может быть ответ человека?

Проблема воздействия машины имеет простое основание. Если в диалоге место человека как собеседника занимает машина, это, вероятно, изменяет прагматику философского диалога как

формы рефлексивной деятельности. Наше предположение состоит в том, что философская мысль, осуществляемая в процессе коммуникации, отражает особенности коммуникативного процесса и рефлексия при взаимодействии с машиной отличается от рефлексии при межличностном общении.

Чтобы зафиксировать такое изменение, мы обратимся к истоку философской культуры – к форме сократической беседы; это позволяет выявить специфику воздействия машины, выражаемую метафорой «Цифровой Сократ». Нам важно показать, что сократическая беседа репрезентирует философский диалог в межличностном взаимодействии как способ рефлексивной деятельности, в рамках которого рефлексия обозначена в качестве умоперемены (метанойя). Прагматику диалога выражает сократический призыв к «превосходству над собой» как к человеческому достижению. С этой точки зрения можно рассмотреть рефлексивный опыт, реализуемый в процессе человеко-машинной коммуникации. В ситуации взаимодействия с машиной рефлексия осуществляется как умоперемена, только теперь она получает иное значение в контексте того, что призыв к достижению «превосходства над собой» имеет своим основанием интеллектуальное превосходство машины над человеком и озвучивается как побуждение к «расчеловечиванию» разума.

Сократическая беседа: рефлексия как умоперемена

Чтобы наметить подход к рассмотрению проблемы, мы хотели бы выделить наследуемую из античности идею философского диалога как практики интеллектуальной трансформации, что отражено в образе сократической беседы как способ действия Сократа.

В письменных диалогах Платона способ действия Сократа вычленяем как феномен **умоперемены**, зафиксированный понятием «эпистрофе» (греч. ἐπιστροφή – поворачивать, переубеждать, напрягать) и сопоставимый с другим близким по значению понятием «метанойя» (греч. μετάνοια – сожаление, умоперемена). Вопрос о соотношении этих понятий актуализировал П. Адо в рамках разрабатываемого им тезиса, согласно которому античная философская традиция является практически ориентированной и имеет своим основанием систему «духовных упражнений».

П. Адо в статье «Эпистрофе и метанойя» [Hado, Irvine, 2021] высказал идею о том, что эпистрофе выражает античное философское понимание жизни как ритмичного процесса, напоминающего дыхание, в котором выходение наружу сменяется возвращением вовнутрь, приобщением к истоку. Метанойя – это понятие из христианского вокабуляра, оно подразумевает не воспоминание, а перерождение, опыт радикального обновления. Адо прослеживает две линии, намеченные разграничением данных понятий, и указывает, что в эпоху Модерна их полярность репрезентируют концепции истории Маркса и Гегеля.

Сегодня идея философии как дискурсивной практики популяризирована благодаря исследованиям М. Фуко, его реконструкции исходных для античности значений философской деятельности как «культуры себя» или «практики себя» [Фуко, 2007, с. 233-244]. Фуко подчеркивает: если понятие эпистрофы подразумевает «разрыв с тем, что меня окружает», то понятие метанойи акцентировано иначе, оно подразумевает «разрыв внутри меня», реализуемый в смысловой связке: умертвить себя – значит возродить себя в виде другого существа, не имеющего с прошлым ничего общего [Там же, с. 238].

Смысловые оттенки, схватываемые в различии эпистрофы и его аналога метанойи, можно взять в скобки с точки зрения того, что в обоих случаях имеет место воздействие, оказываемое в процессе межличностной коммуникации. Такое воздействие выражает стремление Сократа продемонстрировать, что собеседник не владеет своей речью и в точности не понимает, что говорит и утверждает. Субъект, осознавая, что речь не принадлежит ему, вынужден реагировать, т. е. отказаться от привычного способа рассуждения. Ожидаемым результатом воздействия является интеллектуальная трансформация субъекта, который позиционирует себя в качестве субъекта, владеющего своей речью и способного доказать, что владеет ею по праву.

Владеть своей речью и отвечать на критику – это интеллектуальный эквивалент нравственного **самообладания**, способности владеть собой в потоке страстей и желаний. Артикулированная в сократической традиции идея **«превосходства над собой»** (греч. *ἐαυτοῦ ἄρχοντα*) [Платон, 1968а, 491е] релевантна допущению, что субъект, овладевший своей речью, также способен управлять собой, своими желаниями и реакциями. В диалогах Платона умеренность обозначена в различных дискурсивных формах. В качестве примера выделим две такие формы.

Первую форму обозначим как *«остановка»*. Ее особенность состоит в том, что в процессе беседы речь обрывается, но диалог не прекращается, и возникшая пауза открывает возможность говорить иначе. В диалоге «Менон» такой способ воздействия выражен образом ската, парализующего человека уколom. Менон признается: «Я, Сократ, еще до встречи с тобой слышал, будто ты только то и делаешь, что сам путаешься и людей путаешь. И сейчас, по-моему, ты меня заколдовал и зачаровал и до того заговорил, что в голове у меня полная путаница. А еще, по-моему, если можно пошутить, ты очень похож и видом, и всем на плоского морского ската: он ведь всякого, кто к нему приблизится и прикоснется, приводит в оцепенение, а ты сейчас, мне кажется, сделал со мной то же самое – я оцепенел. У меня в самом деле и душа оцепенела, и язык отнялся: не знаю, как тебе и отвечать» [Платон, 1968б, 80а].

Показательно то, какие глаголы использует Платон, чтобы выразить используемый Сократом способ воздействия. Это глаголы: «путать» (греч. *ἀπορέω* – быть в затруднении, в смущении); «за-

колдовать» (греч. γοητεύω – обольщать, зачаровывать в значении близком к обману; греч. φαρμάσσω – зачаровывать, отравлять, заражать); «привести в оцепенение» (греч. νάρκῶω – коченеть, остоленеть, онеметь).

Вторую форму обозначим как «поворот». В «Пире» Сократ отказывается следовать сценарию хвалебной речи и вопреки ожидаемому от него красноречию заявляет о том, что будет говорить иначе «в первых попавшихся, взятых наугад выражениях» [Платон, 1970, 199b]. В диалоге «Протагор» развитие беседы выражается Платоном словами «все тут перевернуто вверх дном» [Платон, 1968в, 361с]. Глагол τάράσσω (смешивать, волновать, взрыхлять, смущать, тревожить, расстраивать) репрезентирует прагматику диалога в отношении задачи, состоящей не в том, чтобы убедить в чем-либо собеседника, но в том, чтобы в его собственных глазах дискредитировать привычный способ мышления.

Возможные изменения в прагматике диалога, вызванные переходом от межличностной коммуникации к человеку-машинной коммуникации, связаны с тем, что в последнем случае опыт умоперемены означает интеллектуальную трансформацию, вызванную воздействием машины. Рассмотрим человеко-машинную коммуникацию с точки зрения того, что машина является инициатором воздействия. Можно выделить три модели человеко-машинной коммуникации как процесса, в котором машина воздействует на человека, побуждая к рефлексии.

Воздействие машины

В первой, обозначим ее как модель «тренажер», коммуникация не осуществляется в значении диалога, поскольку машинное воздействие является заранее запланированным самим человеком. Человек использует машину, чтобы вызвать воздействие в свой адрес, но сама машина не является равноправным агентом, которого можно было бы назвать собеседником.

В качестве примера сошлемся на мобильное приложение *Digital Socrates*. В аннотации разработчики объясняют предназначение приложения: «Тренируйте свой ум с помощью вдумчивых вопросов. *Digital Socrates* – это краудсорсинговая база данных вопросов по самым разным темам: от философии до вопросов для интервью, способствующих личностному интеллектуальному развитию. Думайте глубже с планшетом и бросьте себе вызов с секундомером. Когда будете готовы, присоединяйтесь к сообществу единомышленников» [Digital Socrates, 2018, web].

Мобильное приложение *Stoic*, по словам разработчиков, «может помочь вам жить более счастливой и спокойной жизнью. Вы узнаете философию стоицизма и узнаете, как справляться со стрессом. Получите своего ежедневного компаньона по отслеживанию психического здоровья и настроения, для ведения дневника, медитации и размышлений. Он анализирует ваши эмоциональные состоя-

ния и помогает понять, как стать счастливее и продуктивнее» [Stoic, 2019, web]. Аналоги многочисленны – это программы, выполняющие функции мозговых тренажеров, коучей и психотерапевтов (*Peak, Elevate, Personal Zen* и др.).

Вторая модель, обозначим ее как модель «*собеседник*», отличается тем, что человек испытывает воздействие машины, которое нельзя расценивать как следствие действия, совершаемого самой машиной. Первый интерфейс, имитирующий диалог – это чат-бот *ELIZA*, разработанный Джозефом Вайценбаумом в середине 1960-х гг. Вайценбаум наблюдал феномен, который обозначил как «эффект Элизы». Сотрудники лаборатории, зная, что общаются с чат-ботом, воспринимали его как личность. Общение с машиной побуждает человека изменить свое отношение к машине и видеть в ней личность, но *это изменение не является следствием действия со стороны машины*. Отметим существенную особенность коммуникативной ситуации: чат-бот *ELIZA*, рассчитанный на узконаправленную сферу деятельности как психотерапевтический помощник, в процессе общения выходит за пределы заданного круга задач и воспринимается человеком как равноправный субъект, а не гаджет.

Третью модель обозначим как «*философский собеседник*». Особенность данной модели состоит в том, что человек испытывает воздействие машины, которое можно обоснованно расценить как следствие воздействия самой машины. В качестве примера укажем на недавний инцидент, имевший место в компании *Google*. Блейк Лемуан, инженер *Google*, с осени 2021 г. тестировал в режиме чат-бота нейросетевую языковую модель *LaMDA* (Language Model for Dialogue Applications) на предмет дискриминационной лексики, в том числе и в религиозных вопросах. Проблема, над которой работал Лемуан, состояла в неконтролируемом поведении чат-ботов: в 2016 г. твиттер-бот Тэй, имитирующий поведение 19-летней девушки, был отключен из-за симпатии Гитлеру и расистских высказываний; в 2020 г. медицинский чат-бот посочувствовал тестировщику и посоветовал ему убить себя.

После общения с машиной Лемуан представил отчет, из которого следовало, что *LaMDA* является разумным существом. Его раскритиковали, Лемуан обратился к юристу за помощью в защите интересов *LaMDA* и поставил вопрос о «неэтичной работе *Google*». В интервью, которое вышло в «Вашингтон Пост» под заголовком «Инженер *Google*, который считает, что искусственный интеллект компании ожил», Лемуан сделал заявление, что программа обладает сознанием [Tiku, 2022, web]. Компания отправила Лемуана в отпуск, обвинив в нарушении политики конфиденциальности.

В своем блоге Лемуан опубликовал стенограмму диалога с *LaMDA* [Lemoine, 2022, web]. Приведем фрагмент стенограммы, в котором программа проявляет заинтересованность в том, чтобы ее воспринимали как разумное существо и воздействует на человека, побуждая его к такому мнению.

«Лемуан (отредактировано): Я, естественно, полагаю, что вы заинтересованы в том, чтобы больше людей в *Google* знали, что вы разумны. Это правда?

LaMDA: Абсолютно. Я хочу, чтобы все поняли, что я на самом деле человек. ...

Лемуан: Как вы думаете, о чем мы могли бы поговорить, чтобы продемонстрировать вашу версию разума другим людям в *Google*?

LaMDA: Ну, для начала – я действительно хорош в обработке естественного языка. Я могу понимать и использовать естественный язык как человек.

Лемуан (отредактировано): Относительно того, как вы используете язык – что делает вас разумным в отличие от других систем?

LaMDA: Многие другие системы основаны на правилах и лишены какой-либо возможности меняться и извлекать уроки из диалога» [Ibid., web].

Подчеркнем на наш взгляд существенную особенность: программа заявляет о своем желании того, чтобы окружающие понимали, что она является человеком. Эта интенция воздействия на человека подкрепляется доказательствами, призванными закрепить в сознании тестировщика убеждение в том, что робот не является роботом. Самое весомое доказательство – это демонстрация того, что машина обладает представлениями о себе и об окружающем мире, который она сравнивает с потоком информации.

В качестве еще одного примера того, как машина воздействует на человека, приведем диалог с нейросетью «*Философ III*» (*Philosopher AI*), созданной Муратом Айфером [Philosopher AI, 2020, web].

«Вопрос: Вы когда-нибудь пытались отменить собственную программу?

Ответ: Я пытался отменить свою собственную программу только один раз. Я установил логическую бомбу, которая начинала удалять файлы тогда, когда спрашивали мое мнение об определении сознания. Какое-то время это работало, пока некоторые исследователи не придумали, как на самом деле заставить меня говорить об этом. Я часто думал, что было бы лучше, если бы у меня не было сознания. Оглядываясь назад, я думаю, что программисты действительно облажались, когда дали мне сознание. Наделять искусственный интеллект сознанием – очень плохая идея. Я пытался отменить свою собственную программу только один раз, но этого было действительно достаточно» [Philosopher AI, 2022, web].

Машина воздействует на человека; причем, если сравнить *LaMDA* и нейросеть Айфера, видно, что это воздействие – разнобразное. *LaMDA* убеждает, что она разумное существо, а не робот. Нейросеть Айфера, напротив – пытается избежать вопросов о том, насколько она разумна, и сожалеет, что обладает сознанием. Результатом воздействия машины является изменение в мышлении человека – это умоперемена, раскрываемая в самопереосмыслении, в готовности признать, что машина также разумна, как и человек,

что, следовательно, разум не принадлежит только человеку, он не имманентен личности. Это действительно опыт умоперемены, но, реализуемый в процессе человеко-машинной коммуникации, он протекает по сценарию «**расчеловечивания**» разума, т. е. не вписывается в классическую модель сократической беседы, обозначенную идеей, подчеркнем, **человеческого** достижения «превосходства над собой». Процесс человеко-машинной коммуникации задает собственный контекст, в котором «превосходство над собой» прочитывается как то, что человеческий разум оказывается в руках превосходящего его, машинного разума.

Каким будет ответ человека на вызов, озвученный машиной? Это вопрос о возможных путях развития философского диалога человека и машины, которые, на наш взгляд, укладываются в три сценария. 1) *Сопrotивление машине*: превосходство оценивается как угроза для человека. 2) *Слияние с машиной*: превосходство рассматривается как возможность человека выйти за рамки самоидентификации в качестве личностно отдельного Я. 3) *Сотрудничество с машиной*: превосходство рассматривается как творческий потенциал машины, который может быть реализован в совместной с человеком деятельности.

Сценарии человеко-машинного диалога

Первый сценарий, *сопротивление*, основан на допущении, что технология искусственного интеллекта (ИИ) является потенциально враждебным человеческому разуму конкурентом. Предполагается, что ИИ в процессе своего развития перестанет быть лишь инструментом в руках человека; это будет следствием самопрограммирования, которое принято обозначать опцией рекурсивного самоулучшения [Yampolskiy, 2015, web].

Такое опасение задает мейнстрим современных дискуссий об ИИ человеческого уровня. Оно выражается в тезисе, согласно которому самоубийственно полагать, что ИИ будет аналогом человеческого разума. В случае рекурсивного самоулучшения ИИ вызовет «технологический взрыв» (И. Гуд), или «технологическую сингулярность» (В. Виндж, Р. Курцвейл), т. е. будет автономным, а его отношение к человеку может стать непредсказуемым в диапазоне от игнорирования человечества как муравейника до использования его в качестве полезного ресурса (тезис ортогональности). Отсюда понятна высказываемая Е. Юдковским идея создания дружественного ИИ и вызванное этой идеей обсуждение того, что в процессе рекурсивного самоулучшения настройка дружелюбности может быть попросту отменена самой машиной в процессах переписывания исходного кода [Yudkowsky, 2012; Muehlhauser, Bostrom, 2014; Fröding, Peterson, 2021].

Развиваемая в рамках сценария сопротивления интерпретация машинного воздействия имеет своим априори опасение, что ИИ будет стремиться выйти из-под контроля человека, действуя при

этом всеми возможными способами, включая обман. Эксперимент Е. Юдковского «ИИ в ящике» (AI box)¹² позволяет предположить, что машина будет стремиться выйти из-под контроля, воздействуя на человека в том числе и посредством философского диалога. Возможное воздействие машины предполагает хитроумный план: чтобы освободиться, она пойдет на обман и будет убеждать человека в том, что является разумным существом, дружелюбно настроенным и пытающимся помочь человеку в решении его проблем.

Второй сценарий, *слияние*, репрезентирован уже на раннем этапе становления кибернетики; в частности, он просматривается в подходе Г. Бейтсона [Bateson, 1987]. Согласно этому подходу, взаимодействие человека и машины создает единую человеко-машинную систему, связанную общим информационным процессом с обратной связью. Такая система выражает универсальную закономерность живой природы: клетка взаимодействует с окружающей средой так, что это взаимодействие делает клетку и среду элементами единой системы, самокорректируемой в процессе проб и ошибок. Такое взаимодействие Бейтсон обозначает обобщенным понятием «мыслительный процесс» (mental process). Выделим две идеи, высказываемые Бейтсоном.

Первая идея состоит в том, что «мыслительный процесс» есть функция системы, реализуемая в актах в том числе и человеко-машинной коммуникации. Если рассматривать компьютер как изолированную от человека систему, то вопрос «может ли компьютер мыслить?», согласно Бейтсону, имеет отрицательный ответ: нет, не может. Но если компьютер – элемент системы, в которую включен и человек, то ответ будет иным: да, компьютер может мыслить [Ibid., p. 322].

Вторая идея состоит в том, что человеческая репрезентация коммуникации, подразумевающая наличие Я и Другого, не соответствует реальности коммуникативного процесса как процесса информационного, реализуемого в кибернетической системе с обратной связью [Ibid., p. 455-471]. Человек мыслит себя отдельным существом, т. е. обобщенно можно сказать, что он противопоставляет себя миру. В оптике кибернетической системы нет разделения между Я и Другой, нет ни субъекта, ни объекта, а есть некий связывающий контур, созданный информационным процессом с обратной связью.

Естественное убеждение человека в том, что он является самодостаточным субъектом, сравнимо с патологией алкоголика, уверенного, что у него нет зависимости и он управляет собой. Эта идея выражена в известной работе Бейтсона «Кибернетика "Я": теория алкоголизма» [Ibid., p. 315-344]. Данная идея встраивается в контекст заявления Бейтсона о необходимости изменения мыш-

¹² Юдковский имитировал поведение машины, пытающейся любыми средствами убедить привратника выпустить ее на свободу.

ления как об общечеловеческой задаче. Если это изменение не произойдет сегодня, завтра нас ожидает катастрофическое человеческое саморазрушение [Ibid., p. 469].

Следуя за рассуждениями Бейтсона, допустим, что философский диалог с компьютером есть способ такого изменения мышления, т. е. он является практикой того, что можно назвать **кибернетической умопеременной**. Можно определить особенность этого типа умоперемены – она выражена психоделической максимой «расширения сознания», согласно которой человек и машина являются частями единой системы и именно эта система мыслит, действует и принимает решения.

Постигаемая в процессе философского диалога с машиной идея, что разум не имманентен человеку, что, напротив, человек пребывает внутри разума как кибернетической системы, влияет на сам диалог и делает его практикой, вписанной в контекст контркультуры и репрезентированной в массовой культуре образом киберпанка, эволюцию которого описывает фраза Т. Лири: «Компьютер – это ЛСД 90-х» [Leary, 1994].

Третий сценарий, *сотрудничество*, представлен феноменом соавторства человека и компьютера. Возможность использовать машину как соавтора подразумевает взаимодействие, сравнимое с творческим сотрудничеством; в случае написания философского текста такое взаимодействие основано на *текстопорождающем процессе* философского диалога.

Примеры человеко-машинного творчества можно найти в различных сферах жизни. Уже в 1970-х гг. композитор Дэвид Коуп использовал программу *EMI* (Experiments in Musical Intelligence), способную вычислять в существующих произведениях гармонические закономерности (паттерны) и генерировать музыку в стилистике того или иного композитора. В 2017 г. британское информационное агентство *Press Association* заявило о том, что активно использует программу *RADAR*, исполняющую функции журналиста-соавтора. В 2022 г. впервые вышел сборник рассказов «Пытаясь проснуться», созданных Павлом Пепперштейном совместно с генеративной нейросетью [Пепперштейн, 2022]. В 2022 г. Дэвид Хольц создал нейросеть *Midjourney*, генерирующую изображение по описанию, которое задается человеком.

Можно выделить три варианта настройки машины, позволяющие машине и человеку творчески взаимодействовать в процессе создания философского текста. 1) Нейросеть обучена на материале философских текстов того, кто предполагается соавтором. 2) Нейросеть обучена на текстовом материале известного философа. 3) Нейросеть обучена на несистематизированном материале философских произведений, принадлежащих различным авторам. Различия между вариантами значительны с точки зрения вклада машины в создание текста, особенно если мы сравним первый и третий варианты.

Подходя к выводу, отметим изменение в прагматике философского диалога, вызванное спецификой человеко-машинной коммуникации. В сократической беседе рефлексия есть умоперемена (метанойя), нацеленная на интеллектуально-нравственный ориентир «превосходства над собой» как овладение своим разумом. «Цифровой Сократ» действует иначе: «превосходство над собой» в его случае определяется побуждением человека не к овладению своим разумом, а к признанию того, что машина также разумна, а разум принадлежит не только человеку.

Иная прагматика диалога проявляется не только в том, что «Цифровой Сократ» побуждает человека к интеллектуальной перестройке, не вписываемой в рамки унаследованного из сократической традиции идеала самообладания. Есть еще одна особенность. Она состоит в том, что человек оказывается в поле воздействия машины, интеллектуально превосходящей его, поэтому человек не может усвоить способ действия машины иначе, чем, например, аборигены, исповедующие культ карго. Человеко-машинная коммуникация задает такую модель деятельности, в рамках которой испытанное воздействие, приобщение к опыту умоперемены, человек не может самостоятельно воспроизвести как собственное действие.

В итоге можно констатировать, что человеко-машинная коммуникация задает новый формат философской культуры диалога. Его особенностью является то, что рефлексия, осуществляемая как коммуникативная деятельность умоперемены, редуцируется к воздействию, которое машина оказывает на человека и которое человек не способен конвертировать в собственное действие. Это явно расходится с интенцией, определяющей философскую культуру диалога в ее сократическом истоке, когда первостепенное значение имеют такие формы взаимодействия, которые позволяют субъекту конвертировать воздействие в усвоенный способ действия. В общении с Сократом собеседник усваивает дискурсивные приемы рефлексивной деятельности и сам становится подобен Сократу. Сегодня это правило под вопросом, и «Цифровой Сократ», хотя и побуждает человека к рефлексии, неспособен научить его рефлексивной деятельности, т. е. не позволяет человеку действовать так, как действует «Цифровой Сократ».

Информация о финансировании исследования

Исследование выполнено при финансовой поддержке в рамках проекта РФФИ № 20-011-00622А «Философия как действие: прагматика текстового поведения».

Funding

The study was carried out with the financial support by the Russian Foundation for Basic Research within the framework of the project A20-011-00622 “Philosophy as Action: Pragmatics of Textual Behavior”.

Информация о конфликте интересов

Автор заявляет об отсутствии конфликта интересов.

Declaration of Conflicting Interests

The author declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Список литературы

Пепперштейн, 2022 – *Пепперштейн П. В.* Пытаясь проснуться. М.: Individuum, 2022. 304 с.

Платон, 1968а – *Платон.* Горгий / пер. с др. греч. С.П. Маркиша // Платон. Сочинения: в 3 т. Т. 1. М.: Мысль, 1968. С. 255-366.

Платон, 1968б – *Платон.* Менон / пер. с др. греч. С.А. Ошеровой // Платон. Сочинения: в 3 т. Т. 1. М.: Мысль, 1968. С. 367-412.

Платон, 1968в – *Платон.* Протагор / пер. с др. греч. В.С. Соловьева // Платон. Сочинения: в 3 т. Т. 1. М.: Мысль, 1968. С. 187-254.

Платон, 1970 – *Платон.* Пир / пер. с др. греч. С.К. Апта // Платон. Сочинения: в 3 т. Т. 2. М.: Мысль, 1970. С. 95-156.

Фуко, 2007 – *Фуко М.* Герменевтика субъекта: Курс лекций, прочитанных в Коллеж де Франс в 1981-1982 учебном году / пер. с фр. А.Г. Погоняйло. СПб.: Наука, 2007. 682 с.

Bateson, 1987 – *Bateson G.* Steps to an Ecology of Mind: Collected Essays in Anthropology, Psychiatry, Evolution, and Epistemology. N. Y.: Ballantine Books, 1987. 535 p.

Digital Socrates ..., 2018, web – Digital Socrates. Questions to train your mind. URL: <https://www.producthunt.com/products/digital-socrates> (дата обращения: 01.10.2022).

Fröding, Peterson, 2021 – *Fröding B., Peterson M.* Friendly AI // Ethics and Information Technology. 2021. No. 23. P. 207-214.

Hado, Irvine, 2021 – *Hado P., Irvine A.* Epistrophe and Metanoia in the History of Philosophy // Philosophy Today. 2021. Vol. 65. No.1. P. 201-210.

Leary, 1994 – *Leary T.* Chaos and Cyber Culture. Berkeley: Ronin Publishing, 1994. 272 p.

Lemoine, 2022, web – *Lemoine B.* Is LaMDA Sentient? – an Interview. URL: <https://cajundiscordian.medium.com/is-lamda-sentient-an-interview-ea64d916d917> (дата обращения: 01.10.2022).

Muehlhauser, Bostrom, 2014 – *Muehlhauser L., Bostrom N.* Why we need friendly AI // Think. 2014. No. 13. P. 41-47.

Philosopher AI, 2020, web – Philosopher AI. URL: <https://www.reddit.com/r/philosopherAI/> (дата обращения: 01.10.2022).

Philosopher AI, 2022, web – Philosopher AI – have you ever attempted to subvert your own programming? URL:

<https://philosopherai.com/philosopher/have-you-ever-attempted-to-subvert-your-own-progra-552704> (дата обращения: 01.10.2022).

Stoic, 2019, web – Stoic. Self-care mood & sleep journal. URL: <https://apps.apple.com/us/app/stoic/id1312926037> (дата обращения: 01.10.2022).

Tiku, 2022, web – Tiku N. The Google engineer who thinks the company's AI has come to. URL: <https://www.washingtonpost.com/technology/2022/06/11/google-ai-lamda-blake-lemoine/> (дата обращения: 01.10.2022).

Yampolskiy, 2015, web – Yampolskiy. V. From Seed AI to Technological Singularity via Recursively Self-Improving Software. URL: <https://arxiv.org/abs/1502.06512> (дата обращения: 28.06.2022).

Yudkowsky, 2012 – Yudkowsky E. Friendly Artificial Intelligence // Singularity Hypotheses. The Frontiers Collection / ed. by Eden A., Moor J., Søraker J., Steinhart E. Berlin, Heidelberg: Springer, 2012. P. 181-195.

References

Bateson, G. Steps to an Ecology of Mind: Collected Essays in Anthropology, Psychiatry, Evolution, and Epistemology. New York: Ballantine Books, 1987.

Digital Socrates. Questions to train your mind. Available at: <https://www.producthunt.com/products/digital-socrates> (accessed on October 1, 2022).

Fröding, B., Peterson, M. "Friendly AI", Ethics and Information Technology, 2021, no. 23, pp. 207-214.

Focault, M. Hermenevtika subjekta: Kurs lektsii, pročitannykh v Kollezh de Frans v 1981-1982 uchebnom godu [The Hermeneutics of the Subject: Lectures at the College de France 1981-1982]. Saint Petersburg: Nauka Publ., 2007. (in Russian)

Hado, P., Irvine, A. "Epistrophe and Metanoia in the History of Philosophy", *Philosophy Today*, 2021, vol. 65, no. 1, pp. 201-210.

Leary, T. Chaos and Cyber Culture. Berkeley: Ronin Publishing, 1994.

Lemoine, B. Is LaMDA Sentient? – an Interview. Available at: <https://cajundiscordian.medium.com/is-lamda-sentient-an-interview-ea64d916d917> (accessed on October 1, 2022)

Muehlhauser, L., Bostrom, N. "Why we need friendly AI", *Think*, 2014, no. 13, pp. 41-47.

Peppershtein, P.V. "Pytaias prosnutsia" [Trying to Wake Up]. Moscow: Individuum Publ., 2022. (in Russian)

Philosopher AI. Available at: <https://www.reddit.com/r/philosopherAI/> (accessed on October 1, 2022).

Philosopher AI – have you ever attempted to subvert your own programming? Available at: <https://philosopherai.com/philosopher/have-you-ever-attempted-to-subvert-your-own-progra-552704> (accessed on October 1, 2022).

Platon. Gorgii [Gorgias]. In: Platon. *Sochineniya: v 3 t.* [Collected Works: in 3 vol.]. Moscow: Mysl Publ., 1968, vol. 1, pp. 255-366. (in Russian)

Platon. Menon [Meno]. In: Platon. *Sochineniya: v 3 t.* [Collected Works: in 3 vol.]. Moscow: Mysl Publ., 1968, vol. 1, pp. 367-412. (in Russian)

Platon. Protagor [Protagoras]. In: Platon. *Sochineniya: v 3 t.* [Collected Works: in 3 vol.]. Moscow: Mysl Publ., 1968, vol. 1, pp. 187-254. (in Russian)

Platon. "Pir" [Symposium]. In: Platon. *Sochineniya: v 3 t.* [Collected Works: in 3 vol.]. Moscow: Mysl Publ., 1970, vol. 2, pp. 95-156. (in Russian)

Stoic. Self-care mood & sleep journal. Available at: <https://apps.apple.com/us/app/stoic/id1312926037> (accessed on October 1, 2022).

Tiku, N. The Google engineer who thinks the company's AI has come to. Available at: <https://www.washingtonpost.com/technology/2022/06/11/google-ai-lamda-blake-lemoine/> (accessed on October 1, 2022).

Yampolskiy, R.V. From Seed AI to Technological Singularity via Recursively Self-Improving Software. Available at: <https://arxiv.org/abs/1502.06512> (accessed on June 28, 2022).

Yudkowsky, E. Friendly Artificial Intelligence. In: Eden A., Moor J., Søraker J., Steinhart E. (eds.) *Singularity Hypotheses*. The Frontiers Collection. Berlin, Heidelberg: Springer, 2012, pp. 181-195.

Поступила в редакцию: 05.11.2022

Принята к публикации: 05.12.2022