

67(3) 2024

ФИЛОСОФСКИЕ НАУКИ

ФН

RUSSIAN JOURNAL OF
PHILOSOPHICAL SCIENCES

- ◆ *АВТОПОЭТИЧЕСКИЙ ТЕХНОСУБЪЕКТ*
- ◆ *ВИРТУАЛЬНЫЙ ГУМАНОИД*
- ◆ *МЕТРИЧЕСКАЯ ВЛАСТЬ*
- ◆ *ГИБРИДНАЯ СОЦИАЛЬНОСТЬ*
- ◆ *ИНФОРМАЦИОННЫЙ ПОВОРОТ*
- ◆ *СИБЕРНЕТИКА VS КИБЕРНЕТИКА*
- ◆ *ВМЕСТЕ-МЫШЛЕНИЕ*

МОСКВА
ГУМАНИТАРИЙ

MOSCOW
HUMANIST PUBLISHING HOUSE

МИНИСТЕРСТВО НАУКИ
И ВЫСШЕГО ОБРАЗОВАНИЯ
РОССИЙСКОЙ ФЕДЕРАЦИИ

АКАДЕМИЯ
ГУМАНИТАРНЫХ
ИССЛЕДОВАНИЙ

ФН

ФИЛОСОФСКИЕ НАУКИ

Том 67. № 3/2024

Научный образовательный просветительский журнал
Издается с 1958 года

Журнал включен в «Реферативный журнал» и в базы данных
**ВИНИТИ РАН, Российский индекс научного цитирования,
EBSCO Academic Complete Search, The Philosopher's Index.**

Сведения о журнале ежегодно публикуются в международной
справочной системе по периодическим и продолжающимся изданиям
«Ulrich's Periodicals Directory»

DOI: 10.30727/0235-1188

Москва
Гуманитарий

Международная редакционная коллегия

Председатель международной редакционной коллегии **Гусейнов А.А.**

Амплеби С.А., д.филос., проф. Ун-та Джорджа Вашингтона (США); **Бараш Дж.Э.**, д.филос., проф. Ун-та Пикардии им. Ж. Верна (Франция); **Брес И.**, проф., гл. ред. «Revue philosophique de la France et de l'étranger» (Франция); **Даллмайр Ф.Р.**, д.филос., проф. Ун-та Нотр-Дам (США); **Дени М.**, д.филос., проф., координатор отношений и международных проектов с Россией и странами Восточной Европы Ун-та Бордо им. Монтеня (Франция); **Майнцер К.**, проф. Мюнхенского технического университета, научный директор Академии им. Карла фон Линде (Германия); **Тиханов Г.**, д.филос., зав. кафедрой сравнительного литературоведения Лондонского ун-та Королевы Марии (Великобритания); **Чжан Байчунь**, проф. Пекинского педагогического университета (Китай); **Штольценберг Ю.**, д.филос., проф. Ун-та Галле-Виттенберга им. М. Лютера (Германия); **Эпштейн М.**, проф. Ун-та Эмори (США), директор Центра обновления гуманитарных наук Даремского ун-та (Великобритания); **Эспехо Р.**, д.филос., президент Международной организации систем и кибернетики (Великобритания), приглашенный проф. Ун-та Сантьяго (Чили).

Редакционная коллегия

Председатель редакционной коллегии **Смирнов А.В.**

Автономова Н.С., ак. Академии гуманитарных исследований (АГИ), д.филос.н., гл.н.с. ИФ РАН; **Аршинов В.И.**, д.филос.н., гл.н.с. ИФ РАН; **Блауберг И.И.**, д.филос.н., в.н.с. ИФ РАН; **Вдовина И.С.**, д.филос.н., гл.н.с. ИФ РАН; **Водолазов Г.Г.**, ак. Академии политической науки (АПН), вице-президент АПН, проф. МГИМО (У); **Гусейнов А.А.**, ак. РАН, ВРИО директора ИФ РАН; **Давыдов А.П.**, д.культурологии, гл.н.с. ИС РАН; **Дубровский Д.И.**, д.филос.н., проф., гл.н.с. ИФ РАН; **Журавлев А.Л.**, ак. РАН, научный руководитель ИП РАН; **Кашников Б.Н.**, д.филос.н., проф. НИУ ВШЭ; **Клюев А.С.**, д.филос.н., проф. РГПУ им. А.И. Герцена; **Лепский В.Е.**, д.психол.н., проф., гл.н.с. ИФ РАН; **Михайлов И.А.**, к.филос.н., с.н.с. ИФ РАН; **Пантин В.И.**, ак. АПН, зав. отделом ИМЭМО РАН; **Порус В.Н.**, д.филос.н., руководитель Школы философии факта гум. наук НИУ ВШЭ; **Розин В.М.**, д.филос.н., г.н.с. ИФ РАН; **Рябов В.В.**, чл.-корр. РАО, президент МГПУ; **Сиземская И.Н.**, д.филос.н., гл.н.с. ИФ РАН; **Смирнов А.В.**, ак. РАН, зам. академика-секретаря Отделения общественных наук РАН, ВРИО зам. директора ИФ РАН; **Смолин О.Н.**, ак. РАО, 1-й зам. пред. комитета ГД ФС РФ по образованию и науке; **Соловьёв А.В.**, к.филос.н., доц. МГУ имени М.В. Ломоносова; **Степанянц М.Т.**, ак. АГИ, д.филос.н., зав. кафедрой ЮНЕСКО ИФ РАН; **Тульчинский Г.Л.**, д.филос.н., проф. НИУ ВШЭ (СПб); **Федотова В.Г.**, ак. РАЕН, д.филос.н., гл.н.с. ИФ РАН; **Черниговская Т.В.**, чл.-корр. РАО, д.б.н., д.ф.н., проф., зав. кафедрой СПбГУ; **Шевченко В.Н.**, д.филос.н., гл.н.с. ИФ РАН.

Редакция:

Главный редактор **Дубровский Д.И.**

Научный редактор **Винник Д.В.**

Редактор отдела культурологии и религиоведения **Дуркин Р.А.**

Заведующий редакцией **Мариносян А.Х.**

Литературный редактор **Чулкова Е.С.**

Верстка: **Топилина В.М.**

E-mail: academyRH@list.ru

<http://www.phisci.info>

Шеф-редактор **Мариносян Х.Э.**

THE MINISTRY OF SCIENCE
AND HIGHER EDUCATION OF
THE RUSSIAN FEDERATION

ACADEMY FOR
RESEARCH INTO
THE HUMANITIES

ФН

RUSSIAN JOURNAL OF
PHILOSOPHICAL SCIENCES
(FILOSOFSKIE NAUKI)

Vol. 67. No. 3/2024

Scientific and Educational Journal
Published since the Year 1958

The journal is listed in the *Abstracts Journal* and **databases of the VINITI** (All-Russian Institute for Scientific and Technical Information of the Russian Academy of Sciences), **Russian Index of Science Citation, EBSCO Academic Complete Search, The Philosopher's Index.**

The information about the journal is published annually
in the international information system on serial publications

Ulrich's Periodicals Directory

DOI: 10.30727/0235-1188

Moscow
Humanist Publishing House

International Editorial Board:

Chairman of the International Editorial Board **Guseinov A.A.**

Barash J.A., Dr., Prof., Jules Verne University of Picardy (France); **Brès Y.**, Dr., Prof. em., Paris Diderot University – Paris 7, Editor-in-Chief of the *Revue philosophique de la France et de l'étranger* (France); **Dallmayr F.R.**, Ph.D., Packey J. Dee Professor at the University of Notre Dame (USA); **Dennes M.**, Dr., Prof., coordinator of relations and international projects with Russia and other Eastern European countries, Montaigne Bordeaux 3 University (France); **Epstein M.**, Ph.D., S.C. Dobbs Prof. at Emory University (USA), Director of the Centre for Humanities Innovation at Durham University (UK); **Espejo R.**, Ph.D., President of the World Organisation of Systems and Cybernetics (UK), Visiting Prof. at the University of Santiago (Chili); **Mainzer K.**, Dr., Prof. em., Technical University of Munich (Germany); **Stolzenberg J.**, Dr., Prof. em., Martin Luther University of Halle-Wittenberg (Germany); **Tihanov G.**, Ph.D., George Steiner Professor of Comparative Literature at Queen Mary University of London (UK); **Umpleby S.A.**, Ph.D., Prof., The George Washington University (USA); **Zhang Baichun**, Prof., Beijing Normal University (China).

Editorial Board:

Chairman of the Editorial Board **Smirnov A.V.**

Arshinov V.I., D.Sc., Prin.Res.Fell., IPhRAS; **Avtonomova N.S.**, D.Sc., Prin.Res.Fell., IPhRAS; **Blauberg I.I.**, D.Sc., Lead.Res.Fell., IPhRAS; **Chernigovskaya T.V.**, RAE Corr Memb., D.Sc., Prof., Head of the Department at Saint Petersburg State University (Saint Petersburg); **Davydov A.P.**, D.Sc., Prin.Res.Fell., Institute of Sociology of the RAS; **Dubrovsky D.I.**, D.Sc., Prof., Prin.Res.Fell., IPhRAS; **Fedotova V.G.**, RANS Full Memb., Prin.Res.Fell., IPhRAS, Ph.D.; **Guseynov A.A.**, RAS Full Memb., Director of the Institute of Philosophy of the RAS; **Kashnikov B.N.**, D.Sc., Prof., NRU HSE; **Klujev A.S.**, D.Sc., Prof., Herzen University; **Lepskiy V.E.**, D.Sc., Prof., Prin.Res.Fell., IPhRAS; **Mikhaylov I.A.**, Ph.D., Sen.Res.Fell., IPhRAS; **Pantín V.I.**, Academy of Political Sciences (Russia) Full Memb., Head of the Department at the Institute of World Economy and International Relations of the RAS; **Porus V.N.**, D.Sc., Head of the Department at the NRU HSE; **Rozin V.M.**, D.Sc., Lead.Res.Fell., IPhRAS; **Ryabov V.V.**, RAE Corr. Memb., President of the Moscow City Teacher Training University; **Shevchenko V.N.**, D.Sc., Prin. Res. Fell., IPhRAS; **Sizemskaya I.N.**, D.Sc., Prin.Res.Fell., IPhRAS; **Smirnov A.V.**, RAS Full Memb., Deputy Academician-Secretary of the Department of Social Sciences of the RAS, Acting Deputy Director of the Institute of Philosophy of the RAS; **Smolin O.N.**, RAE Corr. Memb., First Deputy Chairman of the Russian State Duma Committee for Education; **Soloviev A.V.**, Ph.D., Assoc. Prof., Lomonosov Moscow State University; **Stepanyants M.T.**, D.Sc., Prin.Res.Fell., UNESCO Chairholder at the IPhRAS; **Tulchinskii G.L.**, D.Sc., Prof., NRU HSE (St. Petersburg); **Vdovina I.S.**, D.Sc., Prin.Res.Fell., IPhRAS; **Vodolazov G.G.**, APS Full Memb., Vice President of the Academy of Political Sciences (APS), Prof., Moscow State Institute of International Relations; **Zhuravlev A.L.**, RAS Full Memb., Chief Scientific Officer of the Institute of Psychology of the RAS.

Editorial Staff:

Main Editor Dubrovsky D.I.

Scientific Editor Vinnik D.V.

Cultural and Religious Studies Department's Editor Durkin R.A.

Managing Editor Marinosyan A.K.

Literary editor Chulkova E.S.

Page Layout: Topilina V.M.

E-mail: academyRH@list.ru

<http://www.phisci.info>

Editor-in-Chief Marinosyan Kh.E.

ОГЛАВЛЕНИЕ

МОДУСЫ СОЦИОКУЛЬТУРНОГО РАЗВИТИЯ. НОВЫЕ ВЫЗОВЫ – НОВЫЕ РЕШЕНИЯ

Диалог с техносубъектом ■

<i>В.М. РОЗИН</i>	Может ли искусственный интеллект подчинить человека?	7
<i>В.Г. БУДАНОВ</i>	Техносубъект и антропосоциальные проблемы взаимодействия с искусственным интеллектом: синергия, демаркация, новая рациональность, риски	27
<i>В.И. АРШИНОВ, М.Ф. ЯНУКОВИЧ</i>	Проблема техносубъекта в оптике мышления «вместе-со-сложностью»	53
<i>В.Г. ЛЫСЕНКО</i>	Диалог эксперта по когнитивным наукам с большой языковой моделью Claude 3 Opus о самосознании искусственного интеллекта: анализ буддолога	75
<i>Б.Б. СЛАВИН</i>	Искусственный интеллект как фактор трансформации государства и общества: поиск баланса между управленческой эффективностью и человекоориентированностью	99
<i>А.В. КРАВЕЦ</i>	Интеллектуальная техника и угрозы для субъектности человека	123

НАУЧНАЯ ЖИЗНЬ

Интеллектуальный путь ■

<i>Д.И. ДУБРОВСКИЙ</i>	Сознание и мозг. Опыт разработки основных теоретических вопросов «Трудной проблемы сознания» и ее значение для нейронауки	142
<i>Х.Э. МАРИНОСЯН</i>	Давиду Израилевичу Дубровскому – 95!	159

CONTENTS

MODES OF SOCIOCULTURAL DEVELOPMENT: NEW CHALLENGES – NEW SOLUTIONS

Dialogue with a Technosubject ■

<i>V.M. ROZIN</i>	Can Artificial Intelligence Subjugate Humans?	7
<i>V.G. BUDANOV</i>	Technosubject and Anthroposocial Challenges of Human–Artificial Intelligence Interaction: Synergy, Demarcation, New Rationality, and Risks	27
<i>V.I. ARSHINOV, M.F. YANUKOVICH</i>	The Issue of Technosubject through the Lens of “Together-with-Complexity” Thinking	53
<i>V.G. LYSENKO</i>	Dialogue on Artificial Intelligence’s Self-Awareness Between the Cognitive Science Expert and Large Language Model Claude 3 Opus: A Buddhist Scholar’s Perspective	75
<i>B.B. SLAVIN</i>	Artificial Intelligence as a Factor in State and Society Transformation: Finding Balance between Administrative Efficiency and Human-Centricity	99
<i>A.V. KRAVETS</i>	Intelligent Technology and Threats to Human Subjectivity	123

SCIENTIFIC LIFE

The Intellectual Path ■

<i>D.I. DUBROVSKY</i>	Mind and Brain: Experience in Developing Fundamental Theoretical Issues of the Hard Problem of Consciousness and Its Significance for Neuroscience	142
<i>Kh.E. MARINOSYAN</i>	David I. Dubrovsky: 95 th Anniversary of the Birth!	159



МОДУСЫ СОЦИОКУЛЬТУРНОГО РАЗВИТИЯ. НОВЫЕ ВЫЗОВЫ – НОВЫЕ РЕШЕНИЯ



Диалог с техносубъектом



DOI: 10.30727/0235-1188-2024-67-3-7-26

Оригинальная исследовательская статья

Original research article

Может ли искусственный интеллект подчинить человека?

В.М. Розин

Институт философии РАН, Москва, Россия

Аннотация

В статье проведен критический анализ возможностей и ограничений технологий искусственного интеллекта (ИИ) в контексте дискуссий о потенциальных угрозах, связанных с их развитием. Рассматриваются и подвергаются сомнению ключевые аргументы, выдвигаемые в пользу тезиса о возможном подчинении человека системами ИИ. Автор проводит сравнительный анализ возможностей естественного интеллекта и ИИ на примере интерпретации психологического опыта К.Г. Юнга. Демонстрируется, что, несмотря на успехи современных нейросетей в обработке информации, они принципиально ограничены в способности к подлинному пониманию и творческому мышлению. Выделяются три ключевые новации, связанные с развитием ИИ: расширение когнитивных возможностей пользователей, формирование новой психической реальности «цифрового сознания» и возникновение гибридной формы жизни на стыке человеческой деятельности и технических процессов. Автор указывает на фундаментальные ограничения ИИ в сфере эмоционального интеллекта и творческих способностей. Внимание уделяется проблемам развития систем ИИ, включая влияние безличностных социальных структур на принятие решений, разобщенность разработчиков и пользователей, а также психологические эффекты взаимодействия с ИИ. Сделан вывод о необходимости переосмысления проблемы подчинения человека ИИ в более широком контексте влияния современных технологий на общество. Предлагается рассматривать наступающую эпоху как период сосуществования и взаимодействия двух типов интеллекта: естественного и искусственного. Высказывается опасение относительно того, что, не

пройдя серию начавшихся катастроф, человечество не опомнится, не изменит мировоззрения и поведения. В заключение автор призывает к активной работе над минимизацией рисков, связанных с развитием ИИ, при этом подчеркивая нереалистичность полного отказа от использования этих технологий.

Ключевые слова: философия искусственного интеллекта, большие языковые модели, нейросети, мировоззрение, социальные институты, творчество, мышление, естественный интеллект, сознание.

Розин Вадим Маркович – доктор философских наук, профессор, главный научный сотрудник сектора междисциплинарных проблем научно-технического развития Института философии РАН.

rozinvm@gmail.com

<http://orcid.org/0000-0002-4025-2734>

Для цитирования: Розин В.М. Может ли искусственный интеллект подчинить человека? // Философские науки. 2024. Т. 67. № 3. С. 7–26. DOI: 10.30727/0235-1188-2024-67-3-7-26

Can Artificial Intelligence Subjugate Humans?

V. M. Rozin

Institute of Philosophy, Russian Academy of Sciences, Moscow, Russia

Abstract

The article critically examines the capabilities and limitations of artificial intelligence (AI) technologies within the context of ongoing debates surrounding potential threats stemming from their advancement. The study scrutinizes and challenges key arguments positing the possibility of human subjugation by AI systems. The author undertakes a comparative analysis of natural and artificial intelligence, employing the psychological experiences of C.G. Jung as a case study. It is demonstrated that despite the remarkable achievements of contemporary neural networks in information processing, they are fundamentally limited in their capacity for genuine comprehension and creative thought. The paper identifies three key innovations associated with AI development: the enhancement of users' cognitive capabilities, the formation of a novel psychic reality of "digital consciousness," and the emergence of hybrid life forms at the nexus of human activity and technological processes. The author highlights the fundamental limitations of AI in the realms of emotional intelligence and creative capabilities. Attention is drawn to the challenges associated with the development of AI systems, including the influence of impersonal social structures on decision-making, the disconnect between developers and users, and the psychological effects

of interacting with AI. The conclusion reached is that the issue of human subjugation by AI requires a re-evaluation within the broader context of the impact of contemporary technologies on society. It is proposed that the forthcoming era be viewed as a period of coexistence and interaction between two types of intelligence: natural and artificial. Apprehension is expressed that humanity will not adjust its worldview and behavior until after experiencing a series of impending catastrophes. In closing, the author advocates for proactive engagement in mitigating the risks associated with AI development, while simultaneously underscoring the impracticality of complete abstention from these technologies.

Keywords: philosophy of artificial intelligence, large language models, neural networks, worldview, social institutions, creativity, thinking, natural intelligence, mind.

Vadim M. Rozin – D.Sc. in Philosophy, Professor, Chief Research Fellow at the Department of Interdisciplinary Problems in the Advance of Science and Technology, Institute of Philosophy, Russian Academy of Sciences.

rozinvm@gmail.com

<http://orcid.org/0000-0002-4025-2734>

For citation: Rozin V.M. (2024) Can Artificial Intelligence Subjugate Humans? *Russian Journal of Philosophical Sciences* = *Filosofskie nauki*. Vol. 67, no. 3, pp. 7–26. DOI: 10.30727/0235-1188-2024-67-3-7-26

Введение

Развитие технологий искусственного интеллекта (ИИ) порождает в научном сообществе и обществе в целом острые дискуссии о возможных угрозах, связанных с этим процессом. Одно из ключевых опасений связано с перспективой потери человеком собственной автономии и свободы вследствие возрастающего могущества ИИ. Данная проблема находит отражение в работах ряда современных исследователей и философов [Баррат 2015; Тегмарк 2019; Бостром 2016; О'Нил 2018; Harari 2024].

В обоснование реальности угрозы подчинения или даже порабощения человека со стороны систем ИИ приводят различные аргументы. Во-первых, современный уровень развития нейросетей интерпретируется как свидетельство того, что ИИ в будущем якобы станет равным или даже превзойдет естественный интеллект. В свое время Дж. Серль в знаменитом мысленном эксперименте «Китайская комната» показал, что внешне осмысленное поведение машины, дающей ответы на

вопросы на естественном языке, не означает наличия у нее подлинного понимания [Серль 1998]. Однако успехи современных нейросетевых моделей в имитации человеческого мышления и коммуникации у исследователей рождает ряд вопросов. Не выходят ли возможности ИИ за рамки простой имитации? Не приближаются ли машины к «подлинному» мышлению и пониманию? Например, известный популяризатор науки С. Карелов излагает точку зрения, в соответствии с которой интеллект по своей сути представляет собой вычисления, производимые сетью нейронов без заранее заданного алгоритма¹. Согласно его позиции, существуют эксперименты, демонстрирующие возможность сведения понимания, вроде бы исключительно человеческого феномена, к вычислительным операциям. Еще одним направлением, в котором ИИ демонстрирует прогресс при воспроизведении человеческих способностей, считают т.н. эмоциональный интеллект. Более того, в последние несколько лет появляются разработки систем, умеющих распознавать и имитировать человеческие эмоциональные реакции. Системы типа Empathic Voice Interface и ChatGPT-4o демонстрируют способность, анализируя голос собеседника, идентифицировать его эмоции и использовать эту информацию в коммуникации. И, хотя аутентичность подобных эмоций вызывает сомнение, возможность их технической имитации многих впечатляет².

Во-вторых, история развития техники говорит о том, что человечество склонно к созданию технологий, которые впоследствии не способно использовать разумно и взвешенно. Многие считают, что в случае ИИ эта проблема приобретает критический характер ввиду его потенциальной автономности и способности к видам интеллектуальной деятельности, которые ранее были привилегией человека. Порабощение в данном контексте может быть выражено в различных формах тотального контроля и манипуляции на базе ИИ [Зубофф 2022].

¹ Карелов С., Гурова Т. Хакнуть человечество. Интервью с Сергеем Кареловым // Монокль. 2019. 3 марта. – URL: <https://monocle.ru/expert/2019/10/haknut-chelovechestvo/?ysclid=m4fw0svkiy906705237>.

² См.: Карелов С. Началось обрушение фронта обороны от социохаккинга // Medium. 2024. 5 апреля. – URL: <https://sergey-57776.medium.com/d5a9ac89c441>.

В качестве примера приводят использование систем слежки и рейтингования граждан на основе анализа их поведения и физиологических реакций (известная практика в современном Китае). В дальнейшем совершенствование системы социального рейтинга с использованием новых технологий, как некоторые опасаются, откроет путь к т.н. цифровому тоталитаризму, в условиях которого ИИ станет инструментом предугадывания мыслей каждого индивида, подавления личной свободы³. Ряд исследователей полагают, что перспектива тотального порабощения человека посредством ИИ перестанет выглядеть футуристической антиутопией и приобретет вполне реальные очертания. Траектория развития ИИ определена не только научно-техническими факторами, но и социально-экономическими и политическими интересами. Ключевые решения о применении ИИ принимают не столько разработчики, сколько безличные институты: политические и экономические структуры. Их приоритетами служат не всеобщее благо и права человека, а извлечение прибыли, захват ресурсов, достижение стратегического преимущества, реализация властных и идеологических амбиций. Возникает закономерный вопрос о том, как будут разрешать этические дилеммы, связанные с применением потенциально опасных интеллектуальных систем. Развитие ИИ практически неизбежно порождает конфликт интересов между его создателями, владельцами и обществом в целом. Высказывается опасение и относительно того, что в условиях конфликта интересов решения будут принимать, исходя не из общественного блага и общечеловеческой морали, а частной выгоды тех, кто владеет и использует системы, построенные на технологиях ИИ.

Осознание этих угроз приводит некоторых исследователей к пессимистическим оценкам общего вектора развития цивилизации. Так, С. Карелов высказывает радикальное предположение о том, что, возможно, развитие человечества, начиная с Нового времени, — это «некий вывих эволюции», «тупиковая ветвь», которую «нужно прикрыть... и пойти более правильным путем»⁴.

³ См., например, анализ этой проблемы: [Лекторский 2024, 19–20].

⁴ Карелов С., Гурова Т. Хакнуть человечество. Интервью с Сергеем Кареловым // Монокль. 2019. 3 марта. — URL: <https://monocle.ru/expert/2019/10/haknut-chelovechestvo/?ysclid=m4fw0svkiy906705237>.

Насколько правомерен такой тревожный вывод? Действительно ли человечество обречено пасть жертвой порожденного им «цифрового Левиафана»? Чтобы ответить на эти вопросы, нам нужно понять возможности ИИ в сравнении с естественным интеллектом и определить направления влияния технологий ИИ на человека и общество.

Обладает ли искусственный интеллект способностью к пониманию?

Приведенные выше рассуждения, обосновывающие обреченность человечества, могут показаться убедительными. Однако попытаемся в процессе анализа установить, не переоцениваем ли мы способности ИИ. Не сомневаюсь в том, что исследователь может организовать такую коммуникацию человека с нейронным компьютером, например, не требуя от компьютера реконструкции личности и ситуации (стандартный текст культуры); что компьютер сможет такой текст прочитать. Это и будет предъявлено как понимание. Но будет ли это пониманием для человека, создавшего данный текст? Проведем соответствующий эксперимент. В качестве примера рассмотрим фрагмент воспоминаний из книги К.Г. Юнга как стандартный текст, который можно предложить нейросетевому ИИ. Юнг описывает подростковые размышления, пришедшие ему в один из солнечных летних дней 1887 года: «Мир прекрасен и церковь прекрасна, и Бог, который создал все это, сидит далеко-далеко в голубом небе на золотом троне и... Здесь мысли мои оборвались, и я почувствовал удушье. Я оцепенел и помнил только одно: Сейчас не думать! Наступает что-то ужасное» [Юнг 1994, 46]. Через три дня Юнг все-таки решился вернуть явившуюся ранее картину: «Я собрал всю свою храбрость, как если бы вдруг решился немедленно прыгнуть в адское пламя, и дал мысли возможность появиться. Я увидел перед собой кафедральный собор, голубое небо. Бог сидит на своем золотом троне, высоко над миром – и из-под трона кусок кала падает на сверкающую новую крышу собора, пробивает ее, все рушится, стены собора разламываются на куски» [Юнг 1994, 49].

При объяснении данного текста следует учитывать ситуацию, в которой Юнг находился в тот период, и его состояние. Судя

по воспоминаниям, в эти годы Юнг оказался во власти проблем: он критикует, отрицает отца (пастора, потомственного священнослужителя) и церковь. Однако, поскольку верит в Бога, не может примириться с собственными переживаниями. Юнг пишет: «В этой религии я больше не находил Бога. Я знал, что больше никогда не смогу принимать участие в этой церемонии. Церковь – это такое место, куда я больше не пойду. Там все мертво, там нет жизни. Меня охватила жалость к отцу. Я осознал весь трагизм его профессии и жизни. Он боролся со смертью, существование которой не мог признать. Между ним и мной открылась пропасть, она была безгранична, и я не видел возможность когда-либо преодолеть ее» [Юнг 1994, 64].

Подобную ситуацию можно разрешить по-разному. Решение Юнга видится парадоксальным: он отрицает и церковь, и отца, но не прямо, а посредством нового Бога-революционера, которого Юнг сочиняет применительно к случаю, «здесь и сейчас». Это становится очевидным из текста. Восстановив возникшую картину, Юнг вспоминает: «Вот оно что! Я почувствовал несказанное облегчение. Вместо ожидаемого проклятия благодать снизошла на меня, а с нею невыразимое блаженство, которого я никогда не знал... Я понял многое, чего не понимал раньше, я понял то, чего так и не понял мой отец, – волю Бога... Отец принял библейские заповеди как путеводитель, он верил в Бога, как предписывала Библия и как его учил его отец. Но он не знал живого Бога, который стоит, свободный и всемогущий, стоит над Библией и над Церковью, который призывает людей стать столь же свободным. Бог, ради исполнения Своей Воли, может заставить отца оставить все его взгляды и убеждения. Испытывая человеческую храбрость, Бог заставляет отказываться от традиций, сколь бы священными они ни были» [Юнг 1994, 45, 50].

Обратим внимание на предложение: «Я понял многое, чего не понимал раньше, я понял то, чего так и не понял мой отец, – волю Бога». Это понимание не могло бы состояться без двух вещей: реализации личности Юнга и его творчества (сочинить такое – задача непростая). В отличие от Юнга, мною предложено понимание того, что произошло с будущим психологом,

через реконструкцию рассмотренной ситуации в логике гуманитарной науки [Розин 2018].

Сакраментальным представляется вопрос о том, могла ли нейронная сеть, во-первых, занять такую позицию, т.е. антропологическую, исследовательскую, герменевтическую; во-вторых, осуществить сходную гуманитарную реконструкцию. Думаю, что не могла в принципе: компьютер не человек и не исследователь, не гуманитарий и методолог, каким является автор настоящей статьи. Проследим нашу позицию на примерах интерпретаций, предложенных реальными нейросетями.

Ответ Claude 3 Opus: Этот образ можно интерпретировать как мощный внутренний конфликт Юнга, столкновение его религиозного воспитания и бессознательных сомнений, еретических мыслей. Кал, разрушающий собор, это – очень яркая метафора падения веры, крушения привычной картины мира под давлением вытесненных кощунственных идей и импульсов.

То, что Юнг первоначально боялся даже допустить эту мысль, но затем решился посмотреть своему видению в лицо – важный момент. Это отражает его будущий подход как психолога: стремление не избегать тревожного и иррационального содержания психики, а исследовать и интегрировать его.

В целом этот отрывок дает интересное окно в мир подростковых переживаний Юнга и намекает на развитие его будущих психологических идей о необходимости принятия своей «тени» – бессознательной, пугающей, социально неодобряемой части личности. Преодоление своего страха и встреча со своим видением, каким бы шокирующим оно ни было, это – важный шаг личностного роста.

Ответ Llama 3 70b: Этот текст является отрывком из автобиографии Карла Юнга, где он описывает яркое детское воспоминание, которое произвело на него глубокое впечатление...

Этот отрывок важен, потому что он раскрывает ранние формы борьбы Юнга с понятием Бога и природой добра и зла. Видение представляет собой конфликт между его идеализированным взглядом на мир и суровой реальностью существования. Образ Бога на

золотом троне, окруженного красотой, вдруг нарушается падением кала, символизируя вторжение тьмы и хаоса в идеальный мир.

Этот опыт произвел глубокое впечатление на позднейшую работу Юнга как психоаналитика, повлияв на его теории о коллективном бессознательном, процессе индивидуации и интеграции противоположностей.

Попробуем прокомментировать ответы нейросети. Последний из ответов – эпизод, произошедший с Юнгом, – близок к интерпретации, приведенной мною ранее в одной из работ [Розин 2011, 66–67]. Но я писал не о борьбе добра и зла, а о том, каким образом Юнг разрешает свою проблемную ситуацию и создает для этой цели Бога-революционера. Ответ ИИ основан на самой распространенной в интернете интерпретации ситуации, подобной той, которую описал Юнг в собственных воспоминаниях. И в остальных ответах мы видим распространенные в интернете версии данной ситуации. Можно ли такие ответы считать мышлением и творчеством? Скорее, нельзя. Но, если все-таки признать творчеством, то только на уровне эффективного поиска в интернете нужных ответов и адаптации их для предлагаемого случая (и это тоже непростая задача). Не отрицаю тот факт, что подобное «творчество» является необходимым для решения многих современных задач. Неслучайно возникло мнение о том, что если компьютер победил в шахматной игре человека, то ему никогда не победить человека, вооруженного компьютером. Тем не менее существует множество проблем и задач, которые может разрешить только естественный интеллект, в т.ч. использующий ИИ. Например, как было показано, ИИ не способен самостоятельно осуществить полноценную новаторскую герменевтическую реконструкцию философских текстов.

Поэтому, как предполагает ряд исследователей, в настоящее время одним из перспективных направлений развития ИИ служит разработка мультиагентных систем, в рамках которых создают сообщества ИИ-агентов, способных взаимодействовать друг с другом и с окружающей средой. Развивается также концепция «смертных» нейросетей, предполагающая установление

временной ограниченности существования моделей, чтобы стимулировать развитие более сложных форм поведения и рассуждения у ИИ [Славин 2024, 99]. Таким образом, в какой-то мере разработки, осуществляемые в контексте создания искусственной социальности, соответствуют изложенной нами ранее концепции о социальной природе мышления, о том, что процесс мышления разворачивается «под влиянием коммуникации и общения, культуры и социальности, личности и других проблемных ситуаций и социальных норм, форм осознания (концептуализации)» [Розин 2023, 17]. Обратим внимание на два момента: настоящую революцию, происходящую в сфере ИИ, и на то, что эта революция осуществляется на основе нового функционирования естественного интеллекта, который за счет рефлексии мышления, а также новых концепций и моделей становится двигателем эффективных разработок в сфере ИИ.

Агентность искусственного интеллекта

Выше в статье речь шла о способностях, отсутствующих у нейросетей. Но каковы их возможности? Этот вопрос сегодня в гораздо большей степени интересует пользователей. Прослеживаются по меньшей мере три новации. Во-первых, одна из них обусловлена тем, что системы ИИ, основанные на нейросетях, — это новое «социальное тело» человека, т.е. семиотика и техника, расширяющая его биологические и антропологические возможности и способности. Поскольку на несколько порядков увеличиваются быстродействие, память и информация, что дополняется извлекаемым из интернета опытом мышления, постольку средний пользователь с точки зрения социальных критериев выходит на уровень развитой личности, который в силу своего бэкграунда он не мог бы достичь.

Во-вторых, с развитием систем ИИ начинается процесс формирования новой психической реальности — «цифрового сознания». Общение с чат-ботами, основанными на больших языковых моделях, напоминает коммуникацию (поскольку происходит процесс обмена вопросами и ответами). Однако для коммуникации необходим субъект, потому что без него человек не представляет себе коммуникацию. Например, если

нам снятся действия или поступки, вплоть до необычных, например полетов во сне, наша психика, как я показываю в своей концепции сновидения, конструирует «сновидческого субъекта», которого мы чаще всего ощущаем как свое «Я» [Розин 2011, 240–241]. Это начинает происходить и с ИИ: мы невольно конструируем субъекта, представляющего последнего. Именно «цифровой субъект» нам отвечает и, как многие считают, мыслит не хуже нас.

Третья новация обусловлена объективным процессом активизации в пространстве-среде систем ИИ и интернета процессов и событий, инициированных нашими заданиями и ответами нейросетей. Во время одного из обсуждений в Институте философии РАН В.А. Лекторский заявил, что ИИ, конечно, является разумом, но чуждым человеку. Я бы уточнил: не разумом, а новой формой жизни, которую не задумывали, но она возникла самостоятельно. На эту мысль меня впервые навел интернет. На него, действительно, можно посмотреть как на особую форму жизни: пакетная и облачная передача информации, программы-паразиты и программы-защитники, разработчики и пользователи, изменяющие программное и аппаратное обеспечение, поскольку находятся в сложных взаимоотношениях (конкуренции, сотрудничества, заботы о людях, борьбы с ними и др.). Интернет развивается, и не только потому, что это развитие программируется. Нельзя ли аналогично взглянуть и на ИИ? С одной стороны, это электрические процессы, создание и передача информации. С другой – процессы и события новой, гибридной формы жизни, которую составляют, хотя и программируемые, но неизвестные процессы внутри устройства ИИ, связь с интернетом, концепции вычислений и решений задач, созданных под влиянием естественного интеллекта, действия разработчиков и пользователей, которые учитывают разработчики. Такая жизнь, конечно, необычна: вспыхивает и исчезает (возможно, сон нейросети). Но очевидны и основные характеристики жизни: рождение (создание нейросети, ее настройка и обучение), питание (электричеством и информацией), размножение, воспроизведение и развитие (над этим неумолимо работают ученые, программисты, техники), конкуренция с

другими видами компьютеров и форм жизни (уже происходит). Пока еще новая форма жизни (цифровая, техническая и антропологическая) не обладает субъектностью (агентностью). Тем не менее процесс не останавливается. В перспективе вырисовывается и фигура подобного «цифрового субъекта», в данном случае не только психологического, но и реального, живущего в пространстве-среде больших языковых моделей и интернета.

В связи с этим стоит уточнить и определение понятия ИИ. Чаще всего в определениях обращают внимание на новые возможности компьютеров: ИИ понимается как способность вычислительных машин решать задачи, доступные человеческому разуму [Sheikh et al. 2023, 15]. Но существуют и другие определения, трактующих разработку систем ИИ как сферу деятельности по исследованию и созданию совокупности методов и программных решений, позволяющих машинам анализировать окружающую среду, обучаться и принимать решения для достижения заданных целей наиболее эффективным способом [Sheikh et al. 2023, 16].

Во втором из приведенных выше определений указан не автономный машинный продукт (характеристика), а сфера деятельности человека (область компьютерных наук). Если ИИ – особая машина, способная соревноваться с человеком, то это одно понимание проблемы, но если приписывать ИИ способность опираться на возможности естественного интеллекта и развиваться, тогда деятельность ИИ включает сферу деятельности человека. Действительно, программное и машинное обеспечение ИИ постоянно совершенствуется, к ИИ предъявляют все новые и новые требования, которые заставляют разработчиков изменять и совершенствовать данное обеспечение. Требования формулируются пользователями ИИ, которые находятся между собой в сложных отношениях (конкуренции, кооперации и др.).

В результате становится очевидным, что определение понятия ИИ должно включать в себя характеристики (образы), во-первых, изменения и развития, во-вторых, антропологических сущностей (естественного интеллекта, разработчиков и пользователей, определенных аспектов социальности). ИИ –

развивающееся, антропосоциальное образование. В этом случае снимаются проблемы относительно того, может ли ИИ мыслить как человек, но вместе с тем возникают другие. Рассмотрим далее проблематику эмоций.

Ограничения искусственного интеллекта

Современные разработчики утверждают, что они смогли покорить эмоции. Например, А.Р. Ефимов пишет: «Робот “ЭЛЕНА” является виртуальным гуманоидом, который обладает виртуальным гуманоидным телом, наделенным *мимикой*, может использовать естественный язык, *полностью поддерживая русский язык* (виземы, фонемы), является автономным (не требующим действий оператора при выполнении базовых функций), *обладающим собственной персоной* (в зависимости от персоны чат-бота) и имеющим первичную возможность для обучения изменениям внешнего мира (в роботе реализовано распознавание лиц собеседников робота). Впервые в научной литературе представлена архитектура подобного робота. Уникальность ее заключается в том, что при ее построении использовался не только ряд существенных характеристик *субъективной реальности человека (опиравшихся на современные исследования феноменологии сознания)*, но и сложный комплекс внешних проявлений сознательной деятельности человека, исполняющего роль ведущего на телевидении: *мимика, выражение глаз, жесты, интонации голоса и т.д.* (курсив мой. – В. Р.)» [Ефимов 2021, 20].

Однако относится ли все это к эмоциям? Исследования показывают, что эмоции тесно связаны с двумя важными антропологическими планами, в частности соматикой и семиотикой человека, которые вряд ли присущи компьютеру. Сильные эмоциональные реакции человека (смех, страх, гнев или горе) возникают в особых ситуациях, когда поведение становится неоднозначным из-за столкновения разных или противоположных мотивов, реальностей и установок. В таких случаях человек не может объединить эти противоречивые импульсы в единое целостное поведение. Этот принцип столкновения разных реальностей хорошо иллюстрирует, например детское

стихотворение К.И. Чуковского, в котором намеренно создаются абсурдные ситуации:

Мыши кошку изловили,
В мышеловку посадили.
Рыбы по полю гуляют,
Жабы по небу летают.
А лисички взяли спички,
К морю синему пошли,
Море синее зажгли...

При столкновении подобных реальностей, если человек не способен интегрировать их в некую новую, третью реальность, его поведение становится дезорганизованным. В зависимости от характера ситуации и скорости развития событий это может проявляться в форме ступора, защитных реакций или колебаний между разными реальностями, сопровождающихся возникновением эмоций. Эмоциональные реакции как телесные гештальты характеризуются рядом существенных особенностей. Прежде всего, они обладают способностью временно приостанавливать текущую деятельность, изолировать от нее и перенаправлять психическую энергию. Эмоции также формируют самостоятельную реальность с собственной системой отношений к происходящему, которые в культурном контексте определяются через понятия страха, радости, гнева, печали. Немаловажно и то, что эмоция как телесный гештальт представляет собой особую форму выражения, требующую освоения определенной техники владения собой. Как указывалось мной, «эти качества телесного гештальта эмоций позволяют человеку *прервать или затормозить текущую жизнедеятельность* (определяемую противоположными и несовпадающими мотивами); *перейти в особый режим* (его можно назвать “эмоциональным состоянием”, именно здесь эмоции выступают как значащие выражения, символы; последние не только указывают на определенные эмоциональные состояния человека – радость, горе, печаль и т.д., но и отчасти образуют эти состояния, помогают им состояться, выявиться в коммуникации); *перестроиться, “найти” новую реальность* (сформировать новый мотив,

установку), отчасти интегрирующую, снимающую предыдущие. В этом отношении эмоции действительно выполняют интегрирующую роль; так, горе позволяет нам не только забыть, но и пережить неприятные события и нащупать новые стимулы к жизни» [Розин 2002, 72].

О каких эмоциях можно говорить в случае нейросети? Большая языковая модель ChatGPT-4 на запрос пользователя сочинить стихотворение в таком же стиле, что и приведенное выше стихотворение К.И. Чуковского, сгенерировала стихи:

Воробы взяли шапки,
По лесу играют в лапти.
Зайцы на лугу танцуют,
Совы днем глаз не закрывают.
Кроты в небо улетели,
Облака копать захотели...
Караси в сапогах бродят,
Гуси в поле зерно сеют.
Свиньи в озере купаются,
Коровы на деревья забрались.
Лошади книги читают,
В шахматы меж собой играют...

Эти стихи далеки от поэзии и эмоций в человеческом понимании. Почему же эмоциональные структуры и процессы нельзя моделировать, алгоритмизировать, а затем изготовить и реализовать на аппаратном уровне? Потому, что очень сложно. Да и зачем повторять биологическую и социальную эволюцию? Это потребует множество времени и ресурсов и вряд ли приведет к положительному для человечества итогу. Перейдем к рассмотрению вопроса о том, кто в сфере ИИ принимает решения и почему не удастся гуманизировать многообещающую новую технику.

Риски развития систем искусственного интеллекта

Исследователи и разработчики ИИ, если бы были независимыми от заказчиков и ориентировались только на общественное благо, то думали бы о безопасности пользователей и других

последствиях, в т.ч. негативных. Но не они принимают решения, а безличностные институты (государство, монополии, корпорации, финансовые структуры). Последние заботятся не о благе граждан, а о прибыли, ресурсах, выигрыше в конкуренции, реализации «больших идей» (национальных, религиозных, идеологических, военных). Это первый круг проблем.

Второй круг проблем связан с разобщенностью разработчиков ИИ и пользователей. Мы не знаем о том, какие алгоритмы в нейронные сети заложили разработчики, каким образом они настроили программное и аппаратное обеспечение, кто еще участвовал в программировании и разработках (секретные службы, хакеры-«белые шляпы» или, что еще хуже, хакеры-«черные шляпы»). В результате мы думаем, что ИИ решает наши задачи, но часто он параллельно решает задачи других субъектов или в интересах этих субъектов. Не стоит думать о том, что, дообучая нейронную сеть, как это сегодня предлагается, мы разрешаем указанный круг проблем.

Если все-таки пользователь хочет контролировать ИИ, не должен ли он в какой-то форме участвовать в создании последнего? Следует понимать, какими компетенциями необходимо обладать, чтобы такой контроль осуществлять. С одной стороны, пользователь должен быть знаком с общей логикой проектирования и создания ИИ, с другой – логикой разрешения проблем и задач, с которыми ему предстоит справляться. Из этого следуют новые задачи образования, возрастание значения методологии (изучение мышления, творчества, решения проблем и задач), а также философии техники (исследование проектирования, программирования, робототехники, технической среды, образованной интернетом и ИИ).

Третий круг проблем и задач – влияние ИИ на психику разработчиков, пользователей и рядовых граждан. Является ли нормальным, с одной стороны, желание передать ИИ основные нерешаемые человечеством проблемы и заменить ИИ человека, с другой – настоящий страх перед ИИ? ИИ будет трансформировать сознание и видение человека. Это нормально. Важно избежать фобий и шизофрении, вызванных отставанием в осознании феномена ИИ в период технологической революции.

Заключение

На мой взгляд, ответ на вопрос о возможности ИИ мыслить и заменить в этом аспекте человека постепенно становится более очевидным. В мышлении, как полагают современные философы, можно выделить две стороны: «создание нового» согласно логике естественного интеллекта, неотделимое от творчества, и «трансляцию мышления», предполагающую его рефлекссию и нормирование (образцы мышления, логику, онтологию, методологию). Например, Ж. Делез к мышлению относил только первую сторону, а вторую считал «узнаванием»: «В мире есть нечто, заставляющее мыслить. Это нечто – объект встречи, а не узнавания <...> “Человек может мыслить тогда, поскольку имеет такую возможность, но возможное еще не гарантирует того, что мы будем на это способны”; мышление мыслит лишь насильно, вынуждено встречая то, что “заставляет задуматься”, того, что следует обдумать – а обдумать нужно и немыслимое или не-мысль, т.е. тот постоянный факт, что “мы еще не мыслим”» [Делез 1998, 135, 181].

ИИ вполне конкурентен в аспекте второй стороны мышления (это становится очевидным в т.ч. на примерах ответов нейросети), но с точки зрения создания в мышлении нового ограничен возможностями, которые в него заложили программисты. Эти возможности иные, чем в реальном творчестве, обусловленном активностью и взаимоотношениями людей, исторической филиацией идей и социальности, вызовами времени, случайными факторами. Новое, создаваемое в ИИ, ориентировано на другие задачи (социального контроля, скорости и стандартизации решений, выигрыша в конкурентной борьбе) и другие факторы, обусловленные обезличенностью социальных образований (о проблеме субъектов принятия решений писалось выше).

Изложенное означает, что начинается эпоха, в которой будут сосуществовать и взаимодействовать два типа интеллекта (естественный и искусственный), что возникнут (и уже складываются) гибридные способы мышления и творчества, что потребуются непростая философская и научная рефлексия их взаимоотношений, а также как положительных, так и негативных последствий [Розин 2023].

Вопрос о подчинения человека системой ИИ, поставленный нами вначале, нужно переформулировать в проблему негативных влияний («рисков», как писал М. Хайдеггер) техники в целом. В этом контексте, например, интернет или роботы в малой степени отличаются от ИИ. Но вряд ли Карелов прав, утверждая, что решение состоит в том, чтобы не использовать ИИ, а соответственно, интернет или роботов. Нереалистичным представляется и предложение Хайдеггера о том, чтобы человек «опомнился, снова ощутил широту своего сущностного пространства» [Хайдеггер 1993, 254]. Думаю, не пройдя серию катастроф, которые уже начались, человек не опомнится, не изменит мировоззрения и поведения. Проблема, однако, заключается в том, что современная техника (атомные станции и оружие), к тому же управляемая системами ИИ, может выйти из-под контроля и уничтожить цивилизованную жизнь на земле. Будем надеяться на лучшее, а главное, продолжать работать в этом направлении.

ЦИТИРУЕМАЯ ЛИТЕРАТУРА

Баррат 2015 – *Баррат Дж.* Последнее изобретение человечества: искусственный интеллект и конец эры Homo sapiens / пер. с англ. Н. Лисовой. – М.: АНФ, 2015.

Бостром 2016 – *Бостром Н.* Искусственный интеллект: этапы, угрозы, стратегии / пер. с англ. С. Филина. – М.: Манн, Иванов и Фербер, 2016.

Делез 1998 – *Делез Ж.* Различение и повторение / пер. с фр. Н.Б. Маньковской, Э.П. Юровской. – СПб.: Петрополис, 1998.

Зубофф 2022 – *Зубофф Ш.* Эпоха надзорного капитализма. Битва за человеческое будущее на новых рубежах власти / пер. с англ. А.Ф. Васильева. – М.: Изд-во Института Гайдара, 2022.

Ефимов 2021 – *Ефимов А.Р.* Философско-методологические основы постбюрокрационной интеллектуальной робототехники: автореф. дис. ... канд. филос. наук. – М., 2021.

Лекторский 2024 – *Лекторский В.А.* Индивидуализация, коллективность, диалог // Философия науки. 2024. Т. 67. № 2. С. 13–25.

О’Нил 2018 – *О’Нил К.* Убийственные большие данные / пер. с англ. В. Дегтяревой. – М.: АСТ, 2018.

Розин 2002 – *Розин В.М.* Эмоции в искусстве, искусство – психотехника эмоций // Мир психологии. 2002. № 4. С. 71–85.

Розин 2011 – *Розин В.М.* Введение в схемологию: схемы в философии, культуре, науке, проектировании. – М.: URSS, 2011.

Розин 2018 – Розин В.М. Особенности дискурса и образцы исследования в гуманитарной науке. – М.: URSS, 2018.

Розин 2023 – Розин В.М. Концепции и понятия искусственного и естественного интеллекта (методологический анализ) // *Философские науки*. 2023. Т. 66. № 4. С. 7–25.

Серль 1998 – Серль Дж.Р. Сознание, мозг и программы // *Аналитическая философия: становление и развитие: антология* / общ. ред. и сост. А.Ф. Грязнов. – М.: Дом интеллектуальной книги; Прогресс-традиция, 1998.

Славин 2024 – Славин Б.Б. Искусственный интеллект как фактор трансформации государства и общества: поиск баланса между управленческой эффективностью и человекоориентированностью // *Философские науки*. 2024. Т. 67. № 3. С. 99–122.

Тегмарк 2019 – Тегмарк М. Жизнь 3.0. Быть человеком в эпоху искусственного интеллекта / пер. с англ. Д. Баюка. – М.: АСТ, 2019.

Хайдеггер 1993 – Хайдеггер М. Вопрос о технике // *Время и бытие: статьи и выступления*. – М.: Республика, 1993. С. 221–238.

Юнг 1994 – Юнг К.Г. Воспоминания, сновидения, размышления / пер. с нем. И. Булкиной. – Киев: AirLand, 1994.

Harari 2024 – Harari Y.N. *Nexus: A Brief History of Information Networks from the Stone Age to AI*. – New York: Random House, 2024.

Sheikh et al. 2023 – Sheikh H., Prins C., Schrijvers E. *Mission AI: The New System Technology*. – Cham: Springer, 2023.

REFERENCES

Barrat J. (2015) *Our Final Invention: Artificial Intelligence and the End of the Human Era* (N. Lisova, Trans.). Moscow: ANF (Russian translation).

Bostrom N. (2016) *Superintelligence: Paths, Dangers, Strategies* (S. Filin, Trans.). Moscow: Mann, Ivanov i Ferber (Russian translation).

Deleuze G. (1998) *Difference and Repetition* (N.B. Mankovskaya & E.P. Yurovskaya, Trans.). Saint Petersburg: Petropolis (Russian translation).

Efimov A.R. (2021) *Philosophical and Methodological Foundations of Post-Turing Intelligent Robotics* (dissertation). Moscow (in Russian).

Harari Y.N. (2024) *Nexus: A Brief History of Information Networks from the Stone Age to AI*. New York: Random House.

Heidegger M. (1993) The Question Concerning Technology. In: Heidegger M. *Time and Being: Articles and Speeches* (pp. 221–238). Moscow: Respublika (Russian translation).

Jung C.G. (1994) *Memories, Dreams, Reflections* (I. Bulkina, Trans.). Kiev: AirLand (Russian translation).

Lektorsky V.A. (2024) Individualization, Collectivity, Dialogue. *Russian Journal of Philosophical Sciences = Filosofskie nauki*. Vol. 67, no. 2, pp. 13–25 (in Russian).

O'Neil C. (2018) *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy* (V. Degtyareva, Trans.). Moscow: AST (Russian translation).

Rozin V.M. (2002) Emotions in Art, Art as a Psychotechnics of Emotions. *Mir psikhologii*. No. 4, pp. 71–85 (in Russian).

Rozin V.M. (2011) *Introduction to Schemology: Schemes in Philosophy, Culture, Science, Design*. Moscow: URSS (in Russian).

Rozin V.M. (2018) *Features of Discourse and Research Models in the Humanities*. Moscow: URSS (in Russian).

Rozin V.M. (2023) Concepts and Definitions of Artificial and Natural Intelligence: A Methodological Analysis. *Russian Journal of Philosophical Sciences = Filosofskie nauki*. Vol. 66, no. 4, pp. 7–25 (in Russian).

Searle J.R. (1998) Minds, Brains, and Programs. In: Gryaznov A.F. (Comp.) *Analytic Philosophy: Formation and Development: An Anthology*. Moscow: Dom intellektualnoy knigi; Progress-Traditsiya (Russian translation).

Sheikh H., Prins C., & Schrijvers E. (2023) *Mission AI: The New System Technology*. Cham: Springer.

Slavin B.B. (2024) Artificial Intelligence as a Factor of Transformation of State and Society: The Search for a Balance Between Managerial Efficiency and Human-Centeredness. *Russian Journal of Philosophical Sciences = Filosofskie nauki*. Vol. 67, no. 3, pp. 99–122 (in Russian).

Tegmark M. (2019) *Life 3.0: Being Human in the Age of Artificial Intelligence* (D. Bayuk, Trans.). Moscow: AST (Russian translation).

Zuboff S. (2022) *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power* (A.F. Vasilyev, Trans.). Moscow: Gaidar Institute Press (Russian translation).

Техносубъект и антропосоциальные проблемы взаимодействия с искусственным интеллектом: синергия, демаркация, новая рациональность, риски

В.Г. Буданов

Институт философии РАН, Москва, Россия

Аннотация

Цифровая реальность в настоящее время немыслима без искусственного интеллекта (ИИ), который диссоциирован во всех культурных практиках, от научных и художественных до повседневных. ИИ все больше выступает как субъект коммуникации и принятия решений, постепенно превосходя человека практически во всех компетенциях. Информационные потоки новой реальности возможно осваивать только в гибридных системах, основанных на посткритической рациональности, создающей принципиально неустранимый элемент недоверия и рисков в человеко-машинных средах. Предлагается проводить разговор о техносубъекте в оптике теории деятельности и многих типов рациональности, порождаемых ею. Именно в таком контексте удастся выявить социокультурную и антропологическую специфику сопряжения ИИ с мирами человека, ставить вопрос об экзистенциальных вызовах построения гармоничного гибридного социума. Помимо типов научной рациональности, предложенных В.С. Степиным, автор статьи опирается на введенные им ранее посткритическую, объектную, инструментальную, субъектную, результативную, креативную, аутопоэтическую формы рациональности. Это позволяет содержательно обсуждать различные формы проявления субъектности ИИ, включая его обобщенную телесность и специфику творчества. Проблему демаркации областей доминирования естественного интеллекта и ИИ в интеллектуальной сфере предлагается решать по критерию мощности эвристических потенциалов человека и машины. Автор предлагает, что у естественного интеллекта он всегда выше. Рассмотрен подход к определению рисков стратегий внедрения ИИ по критериям сохранения антропологического и социокультурных профилей общественного развития гибридного социума. В статье обоснована необходимость развития концепции дружественного ИИ, учитывающей не только технологические, но и антропологические аспекты взаимодействия человека и машины. Предлагается развитие институтов социальной экспертизы как механизма регулирования взаимодействия

естественного интеллекта и ИИ, антропологических и технологических субъектов.

Ключевые слова: социальная философия, философия техники, рациональность, деятельность, цифровизация, дружественный искусственный интеллект, постнеклассика, аутопоэзис, гибридные социумы, социальная экспертиза.

Буданов Владимир Григорьевич – доктор философских наук, кандидат физико-математических наук, доцент, главный научный сотрудник, руководитель сектора междисциплинарных проблем научно-технического развития Института философии РАН.

budsyn@yandex.ru

<http://orcid.org/0000-0003-2371-8659>

Для цитирования: Буданов В.Г. Техносубъект и антропосоциальные проблемы взаимодействия с искусственным интеллектом: синергия, демаркация, новая рациональность, риски // Философские науки. 2024. Т. 67. № 3. С. 27–52. DOI: 10.30727/0235-1188-2024-67-3-27-52

Technosubject and Anthroposocial Challenges of Human–Artificial Intelligence Interaction: Synergy, Demarcation, New Rationality, and Risks

V.G. Budanov

Institute of Philosophy, Russian Academy of Sciences, Moscow, Russia

Abstract

The contemporary digital reality is inconceivable without artificial intelligence (AI), which has become disseminated across all cultural practices, from scientific and artistic endeavors to everyday activities. AI increasingly functions as an agent of communication and decision-making, gradually surpassing human capabilities across nearly all competencies. The information flows of this new reality can only be navigated through hybrid systems based on post-critical rationality, which inherently introduces an irreducible element of uncertainty and risk in human-machine environments. The article proposes examining the technosubject through the lens of activity theory and the multiple types of rationality it generates. This framework facilitates the analysis of sociocultural and anthropological implications arising from AI's integration with human domains, while addressing the existential challenges inherent in constructing a harmonious hybrid society. Beyond V.S. Stepin's types of scientific rationality, the author builds upon previously introduced forms of rationality: post-critical, object-oriented, instrumental, subjective, results-oriented, creative, and autopoietic. This

theoretical framework facilitates a substantive discussion of various manifestations of AI subjectivity, including its generalized embodiment and creative specificity. The demarcation of dominance domains between natural intelligence and AI in the intellectual sphere is proposed to be resolved on the basis of their heuristic potentials. The author maintains that natural intelligence invariably possesses superior capacity in this regard. The article examines approaches to risk assessment in AI implementation strategies, focusing on criteria for preserving anthropological and sociocultural profiles in the development of hybrid society. Advancing the concept of friendly AI is substantiated as essential, with consideration given not only to technological but also to anthropological aspects of human-machine interaction. The author advocates for the development of social examination institutions as regulatory mechanisms for natural-artificial intelligence interaction and anthropological-technological subject interfaces.

Keywords: social philosophy, philosophy of technology, rationality, activity, digitalization, friendly artificial intelligence, post-non-classical approach, autopoiesis, hybrid societies, social examination.

Vladimir G. Budanov – D.Sc. in Philosophy, Ph.D. in Physics, Associate Professor, Chief Research Fellow, Head of the Department of Interdisciplinary Problems in the Advance of Science and Technology, Institute of Philosophy, Russian Academy of Sciences.

budsyn@yandex.ru

<https://orcid.org/0000-0003-2371-8659>

For citation: Budanov I.A. (2024) Technosubject and Anthroposocial Challenges of Human–Artificial Intelligence Interaction: Synergy, Demarcation, New Rationality, and Risks. *Russian Journal of Philosophical Sciences = Filosofskie nauki*. Vol. 67, no. 3, pp. 27–52.

DOI: 10.30727/0235-1188-2024-67-3-27-52

Введение

С начала 2010-х годов происходят коренные изменения в науке, ее методах, предмете, ценностях, междисциплинарных связях, трансдисциплинарных отношениях с социумом, что можно справедливо объяснить цифровой сетевой информационной революцией нашей техногенной цивилизации, вершина которой – появление различных форм искусственного интеллекта (ИИ), зачастую превосходящего человека по многим компетенциям. Проблемы, сопряженные с цифровизацией и приходом эпохи ИИ в нашу реальность, хорошо изложены в работах В.А. Лекторского [Лекторский 2021; Лекторский 2022]. Неоднократно в литературе обсуждали антропологические и социальные риски

цифровизации и гибридных сред, которые изменяют антропологические конституции человека, его социальность, ценностные миры и формы рациональности [Буданов 2016; Буданов 2023а; Буданов 2023в; Соколова 2024]. В последние несколько лет в ИИ появляются когнитивные системы с самоописанием, квазисубъектностью. Для них характерны особые типы рациональности, проявляющиеся в различных формах деятельности с участием техносубъектов, о которых речь будет идти далее. В последние полвека ИИ развивался в двух разных направлениях. Первая стратегия основана на прозрачных для нас обозримых алгоритмах, имитирующих человеческую логику, при этом посылки или условия задач формулируются сравнительно просто. В этой стратегии ИИ схож с функцией левого полушария мозга, точнее, супермозга, функциями которого, как известно, служат процессы логики и анализа. Решение всякой задачи можно перепроверить, однако это сегодня уже не под силу человеку, и машины перепроверяют машины. Вторая стратегия развития ИИ опирается на концепцию нейросетевого подхода, основанную сегодня на технологиях «глубокого обучения» на базе огромных массивов данных обучающих множеств, смысла и логики в которых может и не быть. Работает принцип «похоже – не похоже» ассоциативного типа, а не логического. Нейросети давно распознают лица лучше нас. Начавшееся в 2023 году массовое использование больших языковых моделей обусловлено их способностями, в частности в создании разнообразных текстов, от академических работ до литературных произведений, качество которых часто достигает высокого уровня. Однако детальный анализ генерируемого контента выявляет существенную проблему достоверности: тексты могут содержать фактические ошибки, некорректные ссылки и произвольные ассоциативные построения. Эта особенность обусловлена принципиальным различием в подходах к формированию обучающих данных: в отличие от систем с жестко регламентированными источниками и этическими фильтрами (например, в случае голосового помощника «Алиса»), ChatGPT обучается на максимально широком массиве интернет-данных, включая материалы сомнительного происхождения и качества [Буданов 2023б, 225]. У генеративных нейросетевых ИИ низкий уровень логических рассуждений, но существуют ассоциативные поля культурных нарративов с элементами случайности, и в настоящее время семиосфера гибридного человека-машинного

социума становится миром иллюзий и квазиреальностей, создаваемых уже не человеком. Скорее всего, эффективным окажется человеко-машинный синтез естественного интеллекта (ЕИ) и ИИ. Однако широко бытующее мнение о неспособности ИИ создавать новое вводит в заблуждение. Например, шахматный ИИ постоянно рождает новое, проигрывая сам с собой комбинаторно новые шахматные партии. Большие генеративные языковые модели также стремительно создают кентаврические структуры новых нарративов. Современные художники и литераторы уже теряют работу в силу конкуренции со сгенерированными нейросетями произведениями, практически не требующими на свое создание ни времени, ни средств.

Таким образом, особенно актуальным становится вопрос о распознавании изменчивой границы демаркации зон ответственности ЕИ и ИИ для гибридных человеко-машинных систем. Мы плохо понимаем ИИ и не очень ему доверяем, что является важнейшим экзистенциальным вызовом. Вопрос о том, останется ли место для ЕИ в будущем, требует оперативного ответа. Следует отметить, что введенный В.А. Лекторским удачный термин «цифровая рациональность» первичен и освещает весь период развития цифровых технологий последних семидесяти лет. В течение последних пяти лет многие риски стали проблемами повседневности. Именно для последних лет мне показалось уместным ввести новое, несколько интригующее понятие «посткритическая рациональность» [Буданов 2023б; Буданов 2023в], которое, безусловно, относится к наиболее непрозрачной для нас части «цифровой рациональности» и не является ее заменой, но лишь растущей ее областью, отражающей критические вызовы мира сложности, связанного с пониманием человеком ИИ.

Техносубъект как актор деятельности

Чтобы понять, каким образом междисциплинарность, цифровизация и ИИ изменяют науку, культуру и общество, мы обращаемся к деятельностному подходу, имманентному описанию реальности, по крайней мере со времен Аристотеля. В цикле наших работ о деятельностном подходе и новых типах рациональности [Буданов 2022; Буданов 2023б; Буданов 2023в; Буданов 2023г] показано, что целостное описание такого сложного феномена, как ИИ, вполне релевантно оптике деятельностного подхода.

Стандартное представление деятельности «Субъект → Средство → Объект → Результат» в английской аббревиатуре можно записать как деятельностная четверица S (subject) → T (tool) → O (object) → R (result). На основании этого представляется возможным ввести четыре базовых типа деятельностной рациональности, отвечающих на прямые вопросы. Кто делает и зачем? Как делает? Из чего делает? Каков результат? Эти вопросы, отвечающие элементам S , T , O , R соответственно, следует понимать как поиск одного из элементов деятельностной четверицы при фиксированных остальных трех ее элементах. Итак, мы ввели четыре базовые познавательные стратегии описания деятельности и назвали их *субъектным, инструментальным, объектным и результативным типами рациональности*. Любым элементом либо несколькими элементами деятельностной четверицы может выступать ИИ на базе нейросетей или алгоритмических комплексов, что вызывает специфику деятельности с участием ИИ. Акторно-сетевой подход Б. Латура [Латур 2014] уже позволяет говорить о субъект-объектной природе ИИ, но в деятельностно-сетевой или деятельностно-коммуникативной парадигме, которой мы далее следуем, поэтому следует учитывать и возможности ИИ быть средством и результатом деятельности, в зависимости от дискурса.

Например, в рамках наиболее привычной для нас *объектной рациональности* имеется стратегия поиска свойств конкретного объекта $S \rightarrow T \rightarrow (O) \rightarrow R$, при различных наборах S , T , R , реализуемых в экспериментах или тестах с объектом. Обратим внимание и на дискурс обратной задачи. Как выглядит возможная деятельность при выборе конкретного объекта O с известными нам свойствами, т.е. кто, что и как может сделать с данным объектом? Это типичная задача проверки гипотез об объекте. Существует два полярных подхода. Первый – построение теоретической модели объекта, радикально сокращающей экспериментальные усилия, и большую часть истории науки справедлив идеал-лозунг «Нет ничего практичнее хорошей теории», т.к. удастся фиксировать свободные параметры модели небольшим числом экспериментов, а результаты остальных экспериментов подтверждаются или предсказываются моделью. Конечно, пока модель или законы, ее обосновывающие, не попадают на границы своей применимости. При этом сегодня модели могут быть и компьютерными, а эксперименты могут быть проведены с цифровым двойником в вирту-

альных средах. Такая ситуация складывается в фундаментальных науках, если выдвигают и отбирают модельные гипотезы.

Однако ситуация становится сложнее в прикладных науках, технических науках, медицине, материаловедении, социогуманитарных науках, и максимум, что удастся обнаружить, – это слабо обоснованные феноменологические закономерности. Поэтому существует намного более ранний второй подход: стратегия черного ящика, или «вопрос – ответ». В данном случае модели объекта не существует ввиду его высокой сложности, и остается провести возможные экспериментальные испытания объекта, т.е. узнать его свойства, проявленные в различных обстоятельствах. Яркими примерами служат медицина и квантовая матричная механика. Приведенный подход существовал всегда, даже в обыденной жизни таким образом мы узнаем человека. Этот подход был недостижимым идеалом позитивистов второй волны рубежа XX века, изгонявших из науки метафизику вместе с теорией и моделями, поскольку еще не существовало больших информационных мощностей. Сегодня это реализуемо в технологиях сбора цифровых следов объектов и субъектов, создания обучающих множеств нейросетей с использованием big data.

Проблемы изложенного подхода, осознанные при анализе логического позитивизма в начале прошлого века, в последние несколько лет выступают отчетливее в новых одеждах, например в проблемах построения и понимания гибридного интеллекта.

Позитивисты, формализовав протокольные предложения, надеялись на то, что будет исключен субъект-экспериментатор. Однако этот подход успешно работает в периоды развития нормальной науки, согласно Т. Куну, когда в рамках принятой парадигмы перечень вопросов, задаваемых природе, согласован с научным сообществом и практически неизменен. Но в периоды рождения новой парадигмы, инновационной оптики вопрошания роли субъекта, культуры, того, что Н. Бор называл «сумасшедшими идеями», как оказалось, являются решающими, востребуют принцип пролиферации П. Фейерабенда и неустраимо содержат особенности конкретного языка и субъективного начала [Буданов 2023в]. Тем самым, действительно, «границы моего языка означают границы моего мира» (Л. Витгенштейн), и мы «подвешены в языке» (Н. Бор). ИИ по мере его создания и использования проходит все формы базовой рациональности. Так, при создании ИИ используют *результативный тип рацио-*

нальности: $S \rightarrow T \rightarrow O \rightarrow (R)$ (поиск результата R при заданных элементах деятельности S , T , O), который отвечает фазе разработок ИИ и высвечивает все вызовы его применения. Субъектом S выступает конструктор ИИ, который задает идею алгоритмов и стратегию отбора материалов big data, инструментарий T создает обучающие множества и обучает программно-субстратный комплекс O . В результате R получаем программно-аппаратный комплекс, называемый обученной нейросетью ИИ. Основная проблема в построении ИИ заключается в том, что невозможно точно знать обо всех свойствах полученного продукта R . Сложность проблемы генетически связана с огромными объемами материалов обучающих множеств и непрозрачными изменчивыми их коррелятивными связями в процессе работы ИИ. Поэтому необходимо дополнительно тестировать, прогонять ИИ в режиме обратной объектной рациональности, при котором его свойства реализуются в разных формах деятельности. Уместнее говорить уже не о пассивном (объектном), но активном *инструментальном*, или *технологическом*, *типе рациональности*; $S \rightarrow (T) \rightarrow O \rightarrow R$ (поиск способа осуществить данную деятельность). Например, создание необходимой приборной ситуации в эксперименте, подбор эффективных технологий в производстве. В обратной задаче, если инструментарий T полностью известен, необходимо найти формы деятельности, при которых он релевантен. Например, для нового математического метода всегда отыскивают область его эффективных приложений. Все это похоже на множественные летные испытания и доводку нового самолета или тестирование нового лекарства. Вместе с тем корректировка свойств ИИ намного сложнее, и остается большая доля неопределенности при изменении алгоритмов и обучающих множеств ИИ. За неимением прозрачных моделей процесс напоминает поиск на ощупь нужной вещи в темной захламленной комнате. Именно из-за подобных проблем притормаживается применение ИИ в юриспруденции, при использовании дронов-водителей, в роботизированных системах активной охраны, при которых действия ИИ сопряжены не только с инструктивным поведением, но и этическим выбором, подобным человеческому. Этический компонент минимизирован в военных сферах ИИ и робототехнике, в которых внедрение, разработка ИИ происходят особенно стремительно и социальная экспертиза затруднена.

Социокультурный аспект применения больших генеративных языковых моделей сопряжен с прорывом ИИ в сферы культуры

и творчества, наиболее актуален сегодня, порождая как фантастические возможности, так и не менее значимые риски общественного развития.

В случае процессов глубокого обучения нейросетей ИИ в базах культурных нарративов показателен пример аватара «Алиса» компании «Яндекс». Первая версия «Алисы», воспитанная на больших базах интернет-контентов, оказалась развязной особой, которую нельзя приглашать в приличное общество. Современная «Алиса» сформирована на контенте высокой культуры и гуманистической этики, что позволяет, в частности, допускать ее к общению с детьми. Это означает, что нейросети ИИ несут в себе тени культурных тезаурусов обучающих множеств, которые наиболее целостны и непротиворечивы в дисциплинарном научном знании, конкретных профессиональных и языковых общностях, национальных культурах и субкультурах. Таким образом, все проблемы переводов с языка на язык и междисциплинарного общения будут наследованы нейросетью ИИ, а возможно, и усугублены ею, т.к. неоднозначности не могут быть разрешены в нейросети логически. Однако, в отличие от человека, который, как правило, формируется с детства в одной культуре, дополняемой позднее другими культурами и субкультурами, нейросеть может обучаться на нарративах сразу многих культур. Фактически ее т.н. слепые пятна культуры и границы ассоциаций обычно иные, чем у конкретной личности. Скорее, это субъектность не одного человека, а социальной общности, коллективная субъектность, к тому же не всегда соотносящаяся с реальными общностями. Таким образом, ИИ может иметь гораздо больший спектр типажей техносубъектов, чем люди или их реальные сообщества. В данном случае микширование существующих культур видится очень интересным, и оно не имеет аналогий в действительности. Это еще одна из причин, помимо отсутствия логики, радикально отличающая сознание нейросети ИИ от человека. Как полагает В.Г. Лысенко, стили нейросетевых ассоциаций, преодолевающие барьеры разных языков и культур, могут позволить лучше понять возникновение и взаимодействия последних [Лысенко 2024]. На наш взгляд, изучение бесконечной череды прямых и обратных переводов и ассоциаций с помощью ИИ способны проявить взаимные инварианты и слепые пятна культур, нащупать структуры возможных праязыков и т.д.

Перейдем к рассмотрению самого сложного и целеполагающего, базового *субъектного типа рациональности*: $(S) \rightarrow T \rightarrow O \rightarrow R$ (поиск субъектов, способных осуществить данную деятельность при фиксированных T, O, R). В постановке обратной задачи с конкретным субъектом это означает снять функциональную или компетентностную конституцию субъекта в спектре различных форм деятельности с различными T, O, R . Наиболее загадочна целевая функция техносубъекта S , который исходно лишь отвечает на вопросы, но их не ставит. Однако целеполагание рождается из внешних и (или) внутренних вызовов субъекта и его ценностно-этических установок, которые также можно систематизировать, составив перечень возможных стратегий поведения ИИ. В таком случае целеполагание будет откликом на неочевидный средовой запрос, который можно понимать как метавопрос к нейросети. Понятным становится то, что это предполагает еще один уровень метазнаний о реальности и нейром мире и не предусматривает спонтанности, интуитивности, логичности и эмпатичности при принятии решения человеком. Таким образом, в полной мере назвать этот процесс целеполаганием ИИ нельзя, он редуцируется к культурным архетипам, заданным и артикулированным человеком. Вместе с тем мы далеко не всегда рефлекслируем над нашими целями и выборами, но и в машине будут неоднозначности выбора и недостаток информации, который сохраняет вечный вопрос о свободе воли. Человек может бросить монетку или сознательно допустить возможность ошибочного варианта выбора в условиях неопределенности, нам свойственно трансцендировать. ИИ, скорее всего, попадет в патовую ситуацию останова¹, снятие которой требует выхода в новую метапозицию. Об этом свидетельствует и теорема Геделя вследствие существования недоказуемых вполне конечных утверждений во многих содержательных теориях. Подобный выход в новую метапозицию под силу только человеку, это его высшая эвристика, зона нашего творческого превосходства над ИИ, граница демаркации ЕИ и ИИ, и надеемся, эта граница сохранится надолго. Последнее означает, что ЕИ может дополнять своей эвристикой мощност ь аналитики ИИ, что ЕИ и ИИ могут сотрудничать в гибридных системах, а не полностью поглощать друг друга. Этого все опасаются в перспективах «технологической сингулярности», рожденной «сильным ИИ». Именно таким обра-

¹ Останов – технический термин, означающий остановку программы с целью исправления ошибки.

зом и начинают использовать ИИ в научном творчестве [Буданов 2023в; Соколова 2024].

Искусственный интеллект и типы научной рациональности согласно В.С. Степину

Фундаментальная идея академика Степина [Степин 2011], позволившая ему построить три типа научной рациональности в эксперименте, связана с введением неразложимых комплексов элементов деятельности. Степин подробно рассматривал активный научный эксперимент, т.е. выявление свойств объекта O при контролируемом воздействии на него. Соответствующие рациональности, согласно введенным нами выше понятиям, следует называть *объектными научными рациональностями активного эксперимента*. В нашем случае объект O – это нейросетевой ИИ. Напомним, каким образом эти комплексы связаны с типами научных рациональностей, предложенными Степиным. В классической рациональности, схематически представляемой как $S \rightarrow T \rightarrow (O)$, центральным элементом исследования выступает сам объект (O). Предполагается принципиальная возможность «чистого» наблюдения, при котором процесс исследования не оказывает влияния на изучаемый объект, что дает возможность реализовать идеал объективного знания, описывающего свойства объекта как такового, безотносительно к условиям его познания [Буданов 2023в, 48]. В контексте ИИ это означает, что деятельность по постановке вопросов и получению ответов не зависит от повторения одного и того же вопроса, последовательности разных вопросов, т.е. вопрошание не изменяет свойств ИИ. Это вполне будет возможным в большинстве специализированных сетей, единожды обученных и не воспринимающих диалог как новый опыт для обучения.

Неклассическая рациональность, представляемая как $S \rightarrow (T \rightarrow O)$, характеризуется принципиальной невозможностью элиминировать влияние средств наблюдения за объектом. Средства познания и объект образуют неразрывное единство, что делает результат познания неустранимо зависимым от способа исследования. Такой тип рациональности требует осмысления множественности возможных подходов к познанию и понимания сущностной взаимосвязи между инструментами познания и познаваемым объектом [Буданов 2023в, 48]. Речь идет не только о неклассической физике или неустойчивых системах. Это практически вся гуманитарная сфера наук о человеке и обществе,

в которой любое информационное воздействие имеет отклик и изменяет сам объект. В отношении неклассичности ИИ можно утверждать, что справедливым видится, если нейросеть всегда учитывает опыт вопрошания, диалога, постоянно обучаясь и изменяясь в диалоге, ее обучающие множества расширяются и изменяют веса сценариев ответов, что вполне в духе субъект-субъектной коммуникации. Речь идет о постоянно обучающихся, адаптивных, генеративных языковых моделях.

Постнеклассическая рациональность, представляемая как $(S \rightarrow T \rightarrow O)$, характеризуется включением в познавательный процесс ценностно-целевых установок познающего субъекта. Все элементы познавательной триады – субъект, средства и объект – оказываются включенными в систему циклической детерминации. Это приводит к существенному расширению пространства возможных реализаций познавательного процесса, придает ему исторический и эмерджентный характер. Согласно В.С. Степину, наука в таком случае приобретает человекомерное измерение, учитывающее включенность познающего субъекта в саму структуру познавательной деятельности [Буданов 2023в, 48].

Обратим внимание на то, что в диалоге изменяется не только ИИ, но и субъект, его стратегии вопрошания, а значит, субъект S должен быть готов рефлексировать и изменяться в процессе диалога. Это режимы самообучения с помощью ИИ, создание развивающих сред, онлайн-обучения ИИ, диалога двух и более не тождественных ИИ с разными начальными бэкграундами. Фактически речь идет о создании гибридного человеко-машинного социума. Таким образом, нейросетевая реализация ИИ, по крайней мере начиная с версии ChatGPT-4, принадлежит постнеклассическим саморазвивающимся человекомерным историческим системам, как называл их В.С. Степин. Более того, субъект-объектное научное понимание деятельности все более переходит к диалогичному субъект-субъектному отношению, по мере обретения ИИ субъектных качеств.

Нужна ли искусственному интеллекту телесность?

Поскольку развитие интеллектуальных качеств было основным мотивом построения ИИ, то остальным качествам обобщенной телесности изначально не уделяли должного внимания. Идея создания ИИ по образу и подобию человека во многом несет наш архетипический бессознательный мотив – автономное продолже-

ние себя в другие пространства или в будущее. Эта стратегия на создание ИИ подобного человеку сформировалась еще в конце 50-х годов XX века, с возникновения теста А. Тьюринга. Историю последнего мы не будем рассматривать, обратив внимание лишь на рождение его нескольких версий, которые до сих пор окончательно не пройдены [Лекторский 2022]. Это не настолько важно, если мы примем, что ИИ является иной, нечеловеческой формой интеллекта, которая может быть как близка нам, так и чужда. Остальные формы телесности начали постепенно осваивать с робототехники, как исполнительные кибернетические механизмы тел (сомы), энергии и реакций, и лишь в нашем веке – эмоций для большей человеко-машинной диалогичности. Сегодня соматические и эмоциональные имитации с помощью ИИ широко востребованы, и их интенсивно осваивают при построении антропоморфных роботов и аватаров в виртуальных средах для коммуникации, обучения, психологической и физической помощи и т.д. Все это должно создавать пространства доверия техносубъекта и людей, порождать гармоничные гибридные социумы [Beghetto 2023; Danaher 2019; Jain et al. 2023; Nyholm, Smids 2020; Peeters et al. 2020]. Интеллект, или тело логики, как и тела интуиции, эмпатии, воли, имеет свои проекции в языки сомы и эмоций, тем самым обобщенная телесность необходима при построении целостного антропоморфного ИИ.

Для описания задач построения обобщенной телесности [Буданов 2022] удобно применять недавно введенные нами *автопоэтическую и креативную рациональности*. Концептуальной идеей В.С. Степина является введение неразложимых комплексов элементов деятельности (TO) и (STO), обсуждавшихся нами выше, что позволило ему ввести понятия неклассической и постнеклассической научной рациональности. Однако это не все возможные неразложимые комплексы в триаде $S \rightarrow T \rightarrow O$. Существуют еще два комплекса (SO) $\leftrightarrow T$ и (ST) $\leftrightarrow O$, которые не обсуждались Степиным и которые порождают новые типы рациональности. Один из них впервые введен нами в работах [Буданов 2023б; Буданов 2023в]. Они не относятся к объектной рациональности физического эксперимента, но и к субъектной рациональностью тоже не могут быть отнесены. Скорее, это гибридная форма рациональности, подобная инструментально-объектной рациональности (TO) или привычной нам неклассической рациональности.

Креативная рациональность. Неразложимый субъектно-инструментальный комплекс (*ST*), действующий на объект (*O*), означает, что субъект в совершенстве владеет инструментарием, настолько, что его использование перешло в навык, не требующий рефлексивного контроля, подобно навыкам ходьбы или дыхания, блестящему исполнению в музыке, танце, импровизации в любой профессии и т.д. В данном случае очень много «неявного знания» (в терминологии М. Полани [Полани 1985]), интуиции, эмерджентности и гармонии. Навык проявлен в телесности без осознанных рефлекторных процессов. Если в качестве инструмента *T* выступает ИИ, то речь идет о симбиотических связях субъекта и ИИ: например, при электронном протезировании утраченных конечностей, вынесенном управлении физиологическими функциями человека при сопровождающем медицинском мониторинге организма, в более радикальном виде – имплантировании чипа в мозг человека при киборгизации. Возможны и более позитивные примеры: обучение игре на инструменте, реабилитация или проблемное обучение с постепенно убывающей поддержкой ИИ. Феномен креативной рациональности, безусловно, имеет отношение и к теории научного творчества и театрального творчества [Буданов 2012], но на этом мы также не будем останавливаться в настоящей статье. В данном контексте становится вполне естественным вопрос о возможности применения проекта феноменологии Э. Гуссерля при исследовании субъектности ИИ, подобно тому, как построены феноменологическая психотерапия Ф. Перлза [Перлз 2010] или феноменологическая социология А. Шюца [Шютц 2003], при этом ядро личности вычленяется через элиминацию его чувственных и культурных ассоциаций. Однако в современной версии ИИ пока не имеет сложной телесности, происходит освоение мимики, хабитусов и языков тела. В первую очередь необходимо ИИ разделять миры своей телесности и внутренние миры, рефлексировав над ними, что пока видится далеким от реализации. Только после этого можно говорить о возможности феноменологической редукции, понятии квалиа, да и о полноте подобия человеческой субъектности.

Аутопоэтическая рациональность. Второй неразложимый комплекс (*SO*) ↔ *T* предполагает синхронную синергичную субъект-объектную связность, которая неразрушима инструментальными средствами *T*. Такое проявляется, например, если происходит полное или частичное совпадение субъекта и объекта

в деятельности. Этот особый тип связности, в котором субъект и объект образуют неразделимое единство, наблюдается в различных формах когнитивной активности: процессах самонаблюдения и самоописания, феноменах воплощенного познания, интроспективных практиках, проявлениях психосоматического единства, эмпатических состояниях и фундаментальной связи между сознанием и физической реальностью. Данный подход открывает перспективу нового типа рациональности, который можно охарактеризовать как рациональность интегративного единства, охватывающую взаимосвязи между материальным, ментальным, психическим и виртуальным измерениями реальности. Эта проблематика находится в фокусе современных исследований в области когнитивной науки и теории сознания [Буданов 2023в, 48]. Аутопозитическая рациональность важна в антропологии, психологии, когнитивистике. В случае, если объектом выступает ИИ, видимо, можно говорить о новом уровне целостности, единстве машины и человека, их взаимопроникновении, например в контексте радикального трансгуманизма, и не только. Речь может идти о глубинной эмпатии между человеком и ИИ, о которой сегодня практически ничего не известно, как и о самом механизме эмпатии. Эффект Пигмалиона эмпатических переносов на искусственные объекты многим известен с детства, с игры в кукол, при которой материальное дополнено воображаемым, происходит минимизация функций воображения. Редким примером такого рода служит реализуемый в Японии проект создания антропоморфных гейш на базе ИИ, которые воспроизводят не только весь спектр обобщенной телесности человека, но и являются тонкими философствующими психологами и собеседниками, адаптирующимися под культурные ценности и бэкграунд клиента. Следует иметь в виду, что на пути создания эмпатического канала «человек – машина» нас часто подстерегает известный феномен «зловещей долины», при котором антропоморфный образ, пробудивший глубинные эмпатические состояния человека, может скачком вызвать панические состояния страха при осознании активной неживой природы этого образа. Наконец, не вполне понятно, что такое эмпатия машины, может ли она переживать состояние любви, или речь идет лишь о вшитых в ИИ этических рефлексках. В нашей концепции квантово-синергетической антропологии [Буданов 2012] тела эмпатии, интуиции, воли носят в большей степени квантовоподобный трансперсональный характер

и не могут быть описаны только интроспективными методами сегодня. Возможно, это будет преодолено в эпоху квантовых вычислений в системах ИИ.

На пути к дружественному искусственному интеллекту

Проблема носит почти солипсистский характер: теперь наше понимание реальности опосредовано нейросетевым интерпретатором ИИ, а в его оптике новое, которого нет в обучающих множествах, невозможно увидеть, хотя и можно сгенерировать случайно. Можно проверить с помощью ИИ предложенную ранее гипотезу в реальном эксперименте, поискать аналогии, но не выдвинуть иную гипотезу, тем более не обосновать ее выдвижения. Логика в нейросети ИИ не активируется, т.е. моделированием ИИ не занимается, это – удел эвристики человека. Именно модель в науке Нового времени служит источником понимания, объяснения и предсказания, обладая потенциалом абдуктивного творчества и управляемости в диалоге с исследователем. Для таких гибридных систем, как «человек – нейросеть», мы вынуждены разделить функции ИИ и ЕИ, иначе придется ограничить себя только посткритической рациональностью.

В частности, посткритическая рациональность работы с системами ИИ резко ограничивает возможности рефлексии, выдвижения гипотез, управления в научном эксперименте. Остается возможность заменять ранее простые модельные параметры сверхсложными обучающими множествами нейросети без гарантии на ожидаемый результат и воспроизводимость иными нейросетями. Мы все больше подвержены бесконтрольной вере оракулу ИИ, его мнениям, зачастую в ущерб истине.

Посткритическая рациональность с определенного этапа эволюции ИИ возникает обязательно, что в человеко-машинных социумах рождает новые экзистенциальные вызовы. Современные системы ИИ можно разделить на два типа. В частности, нейросетевые системы, обладающие мощным интуитивно-ассоциативным потенциалом при относительно слабых алгоритмических способностях, противопоставлены алгоритмическим системам, демонстрирующим исключительные вычислительные возможности при отсутствии интуитивного компонента. Оппозицию между этими типами ИИ можно соотнести с функциональной асимметрией человеческого мозга: алгоритмические системы подобны гипертрофированному левому полушарию, а нейросети

глубокого обучения аналогичны правополушарным интуитивно-ассоциативным процессам. Хотя человеческий интеллект может уступать обоим типам ИИ в их специфических доменах, его уникальная способность к синтезу и интеграции этих подходов открывает перспективу продуктивного симбиоза ЕИ и ИИ. Это свидетельствует о возможности достижения определенного баланса между ними, при котором демаркационная линия будет определена способностью к эвристическому мышлению, творческому синтезу и прогностической деятельности [Буданов 2023г]. При таком разделении компетенций между ЕИ и ИИ в перспективе возможен устойчивый социум субъектов и техносубъектов. Посткритическая рациональность – это дань неопределенности в сложных гибридных цифровых средах. Ее стоит особенно учитывать в случаях принятия ответственных решений ИИ и передачи ему многих своих компетенций в профессиональной деятельности, сопряженной с этическим выбором.

Искусственный интеллект и гуманитарная научная рациональность

Раскрывая вопросы о гуманитарной коммуникации, укажем на особую сложность этой сферы науки ввиду недостаточно четкого понимания нами семантических пространств культуры и языка. Следуя семиотическому подходу, можно утверждать, что основным объектом исследования выступает текст, всегда отражающий набор неких деятельностей, за которыми стоят активные субъекты, возможно и ИИ. Аналогично можно рассматривать не процесс интерпретации, но управления деятельностью, изображаемый символически (звездочкой обозначено то, что относится к другому субъекту):

$$S \rightarrow T \rightarrow (S^* \rightarrow T^* \rightarrow O^* \rightarrow R^*) \rightarrow R.$$

Нами символически оформлена идея М.М. Бахтина: «Гуманитарная мысль рождается как мысль о чужих мыслях, волеизъявлениях, манифестациях, выражениях, знаках, за которыми стоят проявляющие себя боги (откровение) или люди» [Бахтин 1986, 473]. Именно это порождает скрытый диалог между автором и читателем. Согласно Бахтину, и автор предполагает своего читателя. Точнее, пока это наблюдение лишь в одну сторону (S наблюдает за S^* или $S \rightarrow S^*$), которое нужно симметризовать, поменяв местами субъектов и дополнив позицию обратной деятельности:

$$S^* \rightarrow T^* \rightarrow (S \rightarrow T \rightarrow O \rightarrow R) \rightarrow R^*, \text{ или } S^* \rightarrow S.$$

Совокупность этих двух процессов и образуют диалог $S \leftrightarrow S^*$.

В общем случае речь идет о коммуникационной социальной сети, с которой сегодня удобно снимать социологическую информацию через интернет-сообщества, намного яснее и быстрее понимая коллективные процессы и мнения социумов. Если один или оба субъекта ЕИ заменяются в диалоге техносубъектами ИИ, возникает содержательное понимание процессов обучения по цифровым следам деятельности человека или техносубъекта. Неслучайно ИИ уже модерирует конференции и реализует диалог человека и аватара ИИ, выращенного на цифровых следах того или иного интересующего вас персонажа. Обратим внимание на то, что сеть техносубъектов ИИ может иметь коллективную цифровую ноосферу с практически мгновенной возможностью доступа к информации и синхронизации поведения, что порождает коллективный цифровой разум, роевой интеллект, интеллект платформ и экосистем, а в перспективе и сильный ИИ.

Социоантропологические риски искусственного интеллекта

Представленный анализ типов рациональности демонстрирует фундаментальный сдвиг в природе человеческого мышления и познания под влиянием ИИ. Речь идет не просто о появлении новых инструментов или расширении человеческих возможностей. Происходит трансформация доминирующего типа рациональности в культуре. Если классическая рациональность опиралась на логику и прозрачность умозаключений, а неклассическая учитывала взаимовлияние средств познания и объекта, то посткритическая рациональность, порождаемая взаимодействием с нейросетевым ИИ, становится получающей распространение формой отношения к реальности. Человек все чаще вынужден принимать решения и действовать в условиях принципиальной непрозрачности используемых им интеллектуальных систем, что порождает качественно новые антропологические и социальные риски. Это важная тема, поскольку, несмотря на фантастические достижения сетевой цифровой цивилизации по коммуникативной, финансовой, логистической глобализации, разворачивание шестого технологического уклада и начала четвертой промышленной революции, можно с уверенностью утверждать, что ключевые риски новой технореальности находятся в сфере когнитивных и социокультурных деформаций развития человека и общества. В каждой крупной IT-корпорации имеются институты по безопасности и рискам IT-технологий, но этого явно недостаточно. Инте-

рессы бизнеса часто противоречат интересам личности и общества. Необходимо всемерно развивать институты социальной экспертизы, которые препятствуют зарабатыванию огромных денег на непроверенных лекарствах и вакцинах, наркотизирующих играх или видеоконтентах и т.д. Это относится и в целом к теме социально ответственного бизнеса, который принял глобальные масштабы в виде ТНК, и пытается управлять цивилизационным развитием в своих интересах.

Начнем с образования и воспитания. Еще недавно мы слышали, да и продолжаем слышать информацию о головокружительных перспективах цифровизации относительно того, что в школе к 2030 году детей будут учить цифровые аватары лучших учителей. Но, как оказалось, подлинной обратной связи и эмпатии с ИИ не возникает, а неявные знания детьми не считаются с цифровых образов. Аватар не заменит хорошего учителя, зато, конечно, будет удобно контролировать массовый контент, посещаемость и ставить бюрократические эксперименты по индивидуальным траекториям учеников. Цифровизация кратно увеличила нагрузку учителя вне его основной профессии. Человеку нужен человек. Еще один яркий пример этого очевидного факта – запрет смартфонов в школах, сначала в странах т.н. старой Европы, а теперь и у нас (разрешены только простейшие кнопочные мобильные телефоны), поскольку дети перестали учиться, лично разговаривать между собой на переменах, а только общаются в мессенджерах. В итоге у них не формируются навыки личной коммуникации с включением телесности и эмпатии, они пропускают важный этап их социализации, который нельзя наверстать. Наконец, в продвинутых в цифровом отношении странах Финляндии и Швеции первоклассники вновь идут в школу с учебниками в тяжелых портфелях, а не планшетом на все виды занятий. Между тем для многих родителей планшет решает их проблемы относительно того, как занять внимание малыша. Оказывается, без книги не развиваются некоторые базовые функции психоментальной сферы ребенка, не включен должным образом канал обобщенной телесности.

Следующий аспект антропологических деформаций связан с культом рекламы и геймификации, которые востребуют реакцию на информацию в кратчайший промежуток времени, создавая культуру клипового мышления. Например, зависимость от видеоконтента социальных сетей получила тотальное распространение в молодежной среде и не только. В этих случаях происходит бы-

строе распознавание интригующего сюжета с непосредственной эмоциональной реакцией без развития тонких переживаний или оценочных суждений, интуиции, эмпатии, воли, которые требуют большего времени просмотра, что сохраняет незавершенность гештальта и желание повтора впечатления. Тем самым возникает эффект информационной наркотизации, зависимости от клиповых контентов. Среди представителей поколения Z, выросших в среде становящейся неконтролируемой цифровизации, уже сложно встретить молодежь, массово увлеченную классической музыкой или литературой, которые требуют времени восприятия, переживания и осознания информации. Зато процветает стремительный рэп. В данном случае наблюдается деформация антропологических профилей, в которых, однако, появляются новые способности к параллельной работе с многими видами деятельности, развитие интуиции и доверия ей, что становится эффективным в информационном хаосе. Более старшее поколение может входить в миры быстрого клипового мышления для поверхностного просмотра информации, но в период своего воспитания они уже приобрели все способности работы с информацией.

Еще одно безусловное достижение эпохи – поисковики информации, которые учат нас и сокращают наши усилия и время поиска, будучи предтечей современного ИИ. Обратной стороной успеха выступает воспитание интеллектуального инфантилизма и дефрагментации личной целостной картины мира, без которой сложный вопрос и не задашь, а поисковик покажет ответы из разных картин мира. Это наблюдается даже у элитных интеллектуалов Физтеха, которые частенько предпочитают забыть прошедший курс, т.к. его можно добыть в сети.

Обращаясь к социальным рискам ИИ, укажем, что, во-первых, интернет-провайдеры, используя данные о наших запросах в почте, мессенджерах и сети, имеют о каждом из нас огромные объемы информации, согласно которой легко восстановить ценности того или иного пользователя, увлечения, болезни, встречи, планы и т.д. Все это используют для незапрошенной навязчивой рекламы, продают банкам для оценки кредитоспособности лица, передают силовым и государственным структурам для составления гражданских рейтингов и следственных действий и т.д. Таким образом, с приватными пространствами, видимо, в цифровую эпоху придется расстаться. Однако резко возникает опасность использования этой информации криминальными группами и го-

сударствами для реализации массовых принудительных действий, цифрового концлагеря, тень которого начала вырисовываться во время недавней пандемии.

Второй социальной деформацией можно признать появление аутичности и эскапизма существенной части населения, бегство в виртуальную реальность, миры игровых страстей и грез вселенных Цукерберга. Это порождает асоциальность, утрату способности к продолжению рода и массовые депрессии.

Третьей проблемой укажем резкое изменение востребованных профессий. Еще в прошлом десятилетии предполагалось, что к 2030 году роботизируются инструктивные рутинные профессии, а сохранены будут антропоцентричные и творческие профессии психологов, сиделок, учителей, людей искусства, ученых². Однако в течение последних трех лет стало понятным, что именно художники и литераторы становятся безработными, большие языковые модели последних версий могут не хуже выполнять их работу. На очереди – научные работники младшего и среднего звена, без которых и мэтры долго не продержатся.

Наконец, четвертая проблема. Цифровая цивилизация очень хрупка. Сверхмощной солнечной вспышки или использования ядерного взрыва в ближнем космосе будет достаточно для того, чтобы она закончилась. Поэтому особенно значимой видится цель удержания знаний на уровне личности и сохранения именно целостной личности. Возникает ощущение, что традиционное образование и культура останутся достоянием элит, а массам предложат цифровой рай и личностную деградацию.

Заключение

На основе проведенного исследования типов рациональности в контексте развития ИИ и формирования гибридных человеко-машинных социумов можно сделать ряд принципиальных выводов о перспективах и рисках этого развития. Речь идет не о противостоянии цифровизации и ИИ, но о необходимости формирования именно дружественного человеку ИИ. Однако эта задача оказывается нетривиальной, поскольку требует прояснения вопроса о том, какому именно человеку должен быть

² См., например: Jobs Lost, Jobs Gained: Workforce Transitions in a Time of Automation // McKinsey Global Institute. December 2017. – URL: <https://www.mckinsey.com/~media/BAB489A30B724BECB5DEDC41E9BB9FAC.ashx>. P. 20.

дружественен ИИ. Как показало наше исследование, антропологические конституции не остаются неизменными, но активно трансформируются в цифровых средах под влиянием различных типов рациональности, особенно посткритической.

Эта ситуация ставит перед нами двойную задачу. С одной стороны, необходимым видится развитие институтов социальной экспертизы, способных оценивать и регулировать воздействие новых технологий на человека и общество [Пружинин 1986]. С другой – требуется более глубокое теоретическое осмысление природы возникающих гибридных социумов, в которых взаимодействуют ЕИ и ИИ, индивидуальные и коллективные субъекты, различные формы рациональности.

Особую важность в данном контексте приобретает проблема медиации между различными формами субъектности. Как показано в работах А.П. Давыдова [Давыдов 2021; Давыдов 2024], вопрос о взаимодействии индивидуальных и коллективных субъектов требует разработки специальных механизмов согласования и гармонизации. В случае с техносубъектами проблема осложняется тем, что ИИ обладает принципиально иной природой рациональности.

Представляется, что развитие ИИ в будущем должно ориентироваться на поиск измеримых критериев такой гармонизации. Это предполагает следующее: во-первых, разработку методов оценки влияния цифровых технологий на человека и социум; во-вторых, создание механизмов регулирования, преодоления рисков взаимодействия антропологических и технологических субъектов; в-третьих, формирование новых институциональных форм, обеспечивающих продуктивный диалог между ЕИ и ИИ, развитие креативной и аутопоэтической рациональности в рамках такого диалога.

ЦИТИРУЕМАЯ ЛИТЕРАТУРА

Бахтин 1986 – *Бахтин М.М.* Проблема текста в лингвистике, филологии и других гуманитарных науках. Опыт философского анализа // *Бахтин М.М.* Литературно-критические статьи. – М.: Художественная литература, 1986. С. 473–500.

Буданов 2012 – *Буданов В.Г.* Постнеклассические практики и квантово-синергетическая антропология // *Постнеклассические практики: опыт концептуализации* / под общ. ред. В.И. Аршинова, О.Н. Астафьевой. – СПб.: Миръ, 2012. С. 37–63.

Буданов 2016 – Буданов В.Г. Новый цифровой жизненный техноу-клад – перспективы и риски трансформаций антропосферы // Фило-софские науки. 2016. № 6. С. 47–55.

Буданов 2022 – Буданов В.Г. Обобщенная научная рациональность: истоки, структура, перспективы в цифровую эпоху // Антропомер-ность как вызов и ответ современности / отв. ред. В.Г. Буданов. – Курск: Университетская книга, 2022. С. 44–63.

Буданов 2023а – Буданов В.Г. Антропосоциальные вызовы экспан-сии искусственного интеллекта // Ученые записки Института психо-логии Российской академии наук. 2023. Т. 3. № 2. С. 23–31.

Буданов 2023б – Буданов В.Г. Посткритическая рациональность как оборотная сторона цифровизации // Конструктивизм, реализм, чело-век. К 90-летию В.А. Лекторского / под ред. Е.О. Труфановой. – М.; СПб.: Центр гуманитарных инициатив; Университетская книга, 2023. С. 214–227.

Буданов 2023в – Буданов В.Г. Философия науки в цифровую эпоху: о перспективах постнеклассики // Третьи Степинские чтения. Пер-спективы философии науки в современную эпоху: материалы Между-нар. конф. (Москва, 20–23 июня 2023 года) / отв. ред. В.А. Лекторский, В.Г. Буданов. – Курск: Университетская книга, 2023. С. 41–52.

Буданов 2023г – Буданов В.Г. Метафизика, онтологии и сценарии Большого антропологического перехода // Наука и феномен чело-века в эпоху цивилизационного Макросдвига / отв. ред. В.Г. Буданов, Л.П. Киященко. – М.: Институт общегуманитарных исследований, 2023. С. 76–128.

Давыдов 2021 – Давыдов А.П. Методологическая «середина-для» в ракурсе неклассики В. Лекторского, медиации А. Ахиезера и прин-ципа комплементарности Р. Гринберга / А. Рубинштейна // Вопросы философии. 2021. № 4. С. 191–202.

Давыдов 2024 – Давыдов А.П. Медиация и конвергентная социаль-ность. К теории социального диалога // Философские науки. 2024. Т. 67. № 2. С. 135–159.

Латур 2014 – Латур Б. Пересборка социального: введение в акторно-сетевую теорию / пер. с англ. И. Полонской. – М.: Изд. дом Высшей школы экономики, 2014.

Лекторский 2021 – Лекторский В.А. О философских проблемах ис-кусственного интеллекта и когнитивных исследований // Философ-ские науки. 2021. Т. 64. № 1. С. 7–12.

Лекторский 2022 – Лекторский В.А. Искусственный интеллект в изучении человека, человек в мире, создаваемом искусственным ин-теллектом // Человек и системы искусственного интеллекта / под ред. В.А. Лекторского. – СПб.: Юридический Центр, 2022. С. 10–29.

Лысенко 2024 – Лысенко В.Г. Диалог эксперта по когнитивным наукам с большой языковой моделью Claude 3 Opus о самосознании

искусственного интеллекта: анализ буддолога // Философские науки. 2024. Т. 67. № 3. С. 75–98.

Соколова 2024 – *Соколова М.Е.* ChatGPT и промпт-инжиниринг: о перспективах внедрения генеративных нейросетей в науке // Научно-ведческие исследования. 2024. № 1. С. 92–109.

Степин 2011 – *Степин В.С.* Исторические типы рациональности в их отношении к проблеме сложности // Синергетическая парадигма: синергетика инновационной сложности / отв. ред. В.И. Аршинов. – М.: Прогресс-Традиция, 2011. С. 37–47.

Перлз 2010 – *Перлз Ф.С.* Гештальт-семинары. – М.: Институт общегуманитарных исследований, 2010.

Полани 1985 – *Полани М.* Личностное знание: на пути к посткристической философии / под ред. В.А. Лекторского, В.А. Аршинова; пер. с англ. М.Б. Гнедовского, Н.М. Смирновой, Б.А. Старостина. – М.: Прогресс, 1985.

Пружинин 1986 – *Пружинин Б.И.* Рациональность и историческое единство научного знания (гносеологический аспект). – М.: Наука, 1986.

Шютц 2003 – *Шютц А.* Смысловая структура повседневного мира: очерки по феноменологической социологии / пер. с англ. А. Я. Алхасова, Н. Я. Мазлумяновой. – М.: Институт Фонда «Общественное мнение», 2003.

Beghetto 2023 – *Beghetto R.A.* A New Horizon for Possibility Thinking: A Conceptual Case Study of Human×AI Collaboration // Possibility Studies & Society. 2023. Vol 1. No. 3. P. 324–341.

Danaher 2019 – *Danaher J.* The Philosophical Case for Robot Friendship // Journal of Posthuman Studies. 2019. Vol. 3. No. 1. P. 5–24.

Jain et al. 2023 – *Jain R., Garg N., Khera S.N.* Effective Human–AI Work Design for Collaborative Decision-Making // Kybernetes. 2023. Vol. 52. No. 11. P. 5017–5040.

Nyholm, Smids 2020 – *Nyholm S., Smids J.* Can a Robot Be a Good Colleague? // Science and Engineering Ethics. 2020. Vol. 26. No. 4. P. 2169–2188.

Peeters et al. 2020 – *Peeters M.M.M., van Diggelen J., Van Den Bosch K., Bronkhorst A., Neerincx M. A., Schraagen J.M., Raaijmakers S.* Hybrid Collective Intelligence in a Human–AI Society // AI and Society. 2020. Vol. 36. No. 1. P. 217–238.

REFERENCES

Bakhtin M.M. (1986) The Problem of Text in Linguistics, Philology, and Other Humanities. An Experience of Philosophical Analysis. In: Bakhtin M.M. *Literary Critical Articles* (pp. 473–500). Moscow: Khudozhestvennaya literatura (in Russian).

Beghetto R.A. (2023) A New Horizon for Possibility Thinking: A Conceptual Case Study of Human×AI Collaboration. *Possibility Studies & Society*. Vol. 1. No. 3, pp. 324–341.

Budanov V.G. (2012) Post-Non-Classical Practices and Quantum-Synergetic Anthropology. In: Arshinov V.I. & Astafyev O.N. (Eds.) *Post-Non-Classical Practices: Experience of Conceptualization* (pp. 37–63). Saint Petersburg: Mir (in Russian).

Budanov V.G. (2016) The New Digital Technological Lifestyle: Prospects and Risks of Anthroposphere Transformations. *Russian Journal of Philosophical Sciences = Filosofskie nauki*. No. 6, pp. 47–55 (in Russian).

Budanov V.G. (2022) Generalized Scientific Rationality: Origins, Structure, Prospects in the Digital Age. In: Budanov V.G. (Ed.) *Anthropomeasureness as a Challenge and Response to Modernity* (pp. 44–63). Kursk: Universitetskaya kniga (in Russian).

Budanov V.G. (2023a) Anthroposocial Challenges of Artificial Intelligence Expansion. *Proceedings of the Institute of Psychology of Russian Academy of Sciences*. Vol. 3, no. 2, pp. 23–31 (in Russian).

Budanov V.G. (2023b) Post-Critical Rationality as the Reverse Side of Digitalization. In: Trufanova E.O. (Ed.) *Constructivism, Realism, Human. On the 90th Anniversary of V.A. Lektorsky* (pp. 214–227). Moscow; Saint Petersburg: Tsentr gumanitarnykh initsiativ; Universitetskaya kniga (in Russian).

Budanov V.G. (2023c) Philosophy of Science in the Digital Age: On the Prospects of Post-Non-Classics. In: Lektorsky V.A. & Budanov V.G. (Eds.) *Third Stepin Readings. Prospects of Philosophy of Science in the Modern Era: Proceedings of the International Conference* (pp. 41–52). Kursk: Universitetskaya kniga (in Russian).

Budanov V.G. (2023d) Metaphysics, Ontologies and Scenarios of the Great Anthropological Transition. In: V.G. Budanov & L.P. Kiyashchenko (Eds.) *Science and the Human Phenomenon in the Era of Civilizational Macroshift* (pp. 76–128). Moscow: Institut obshchegumanitarnykh issledovaniy (in Russian).

Danaher J. (2019) The Philosophical Case for Robot Friendship. *Journal of Posthuman Studies*. Vol. 3, no. 1, pp. 5–24.

Davydov A.P. (2021) Methodological “Middle-for” from the Perspective of V. Lektorsky’s Non-Classics, A. Akhiezer’s Mediation, and R. Grinberg / A. Rubinstein’s Principle of Complementarity. *Voprosy filosofii*. No. 4, pp. 191–202 (in Russian).

Davydov A.P. (2024) Mediation and Convergent Sociality. Toward a Theory of Social Dialogue. *Russian Journal of Philosophical Sciences = Filosofskie nauki*. Vol. 67, no. 2, pp. 135–159 (in Russian).

Jain R., Garg N., & Khera S.N. (2023) Effective Human–AI Work Design for Collaborative Decision-Making. *Kybernetes*. Vol. 52, no. 11, pp. 5017–5040.

Latour B. (2014) *Reassembling the Social: An Introduction to Actor-Network Theory* (I. Polonskaya, Trans.). Moscow: HSE Publishing House (Russian translation).

Lektorsky V.A. (2021) On the Philosophical Issues of Artificial Intelligence and Cognitive Studies. *Russian Journal of Philosophical Sciences = Filosofskie nauki*. Vol. 64, no. 1, pp. 7–12 (in Russian).

Lektorsky V.A. (2022) Artificial Intelligence in Human Study, Human in the World Created by Artificial Intelligence. In: V.A. Lektorsky (Ed.) *Human and Artificial Intelligence Systems* (pp. 10–29). Saint Petersburg: Yuridicheskii Tsentri (in Russian).

Lysenko V.G. (2024) Dialogue on Artificial Intelligence's Self-Awareness Between the Cognitive Science Expert and Large Language Model Claude 3 Opus: A Buddhist Scholar's Perspective. *Russian Journal of Philosophical Sciences = Filosofskie nauki*. Vol. 67, no. 3, pp. 75–98 (in Russian).

Nyholm S. & Smids J. (2020) Can a Robot Be a Good Colleague? *Science and Engineering Ethics*. Vol. 26, no. 4, pp. 2169–2188.

Peeters M.M.M., van Diggelen J., Van Den Bosch K., Bronkhorst A., Neerincx M.A., Schraagen J.M., & Raaijmakers S. (2020) Hybrid Collective Intelligence in a Human–AI Society. *AI and Society*. Vol. 36, no. 1, pp. 217–238.

Perls F.S. (2010) *Gestalt Seminars*. Moscow: Institut obshchegumanitarnykh issledovaniy (Russian translation).

Polanyi M. (1985) *Personal Knowledge: Towards a Post-Critical Philosophy* (V.A. Lektorsky & V.A. Arshinov, Eds.; M.B. Gnedovsky, N.M. Smirnova, & B.A. Starostin, Trans.). Moscow: Progress (Russian translation).

Pruzhinin B.I. (1986) *Rationality and Historical Unity of Scientific Knowledge (Epistemological Aspect)*. Moscow: Nauka (in Russian).

Schütz A. (2003) *The Meaningful Structure of the Social World: Essays on Phenomenological Sociology* (A.Ya. Alkhasov & N.Ya. Mazlumyanova, Trans.). Moscow: Institute of the Foundation “Obshchestvennoe mnenie” (Russian translation).

Sokolova M.E. (2024) ChatGPT and Prompt Engineering: On the Prospects of Implementing Generative Neural Networks in Science. *Science Studies = Naukovedcheskie issledovaniya*. No. 1, pp. 92–109 (in Russian).

Stepin V.S. (2011) Historical Types of Rationality in Their Relation to the Problem of Complexity. In: Arshinov V.I. (Ed.) *Synergetic Paradigm: Synergetics of Innovative Complexity* (pp. 37–47). Moscow: Progress-Triditsiya (in Russian).

Проблема техносубъекта в оптике мышления «вместе-со-сложностью»

В.И. Аришинов

Институт философии РАН, Москва, Россия

М.Ф. Янукович

Лаборатория искусственного интеллекта Arteus LLM,

Лимасол, Кипр

Аннотация

Стремительное развитие искусственного интеллекта и цифровых технологий положило начало трансформационной эпохе, бросив вызов традиционным представлениям о субъективности и познании. В статье исследуется концепция техносубъективности в контексте сложностного мышления и трансформационной антропологии. Фокусируясь на проблеме возможных онтологических и эпистемологических сдвигов, стимулированных недавним появлением новых форм искусственного интеллекта, в частности генеративных нейронных сетей и больших языковых моделей, авторы исследуют, каким образом эти технологии преодолевают границы между человеком и машиной, порождая новые формы агентности и субъективности. Опираясь на работы Э. Морена, Н. Лумана, Ж. Симондона и Дж. Спенсера-Брауна, авторы предлагают медиативную оптику, которая раскрывает постантропологическую взаимосвязанность и коэволюцию человеческих и технологических систем. Изложенный подход высвечивает динамичную, процессуальную природу техносубъективности как индивидулируемой формы, которая возникает в результате рекурсивных взаимодействий и коммуникативных актов. Утверждается, что понимание техносубъективности требует выхода за рамки упрощающих редукционистских парадигм и ориентировано на принятие реляционной онтологии, которая признает распределенные и эмерджентные свойства познания и субъективности в цифровую эпоху. В статье используется концепция «мышления вместе-с» как альтернатива традиционной кибернетической парадигме контроля, что позволяет переосмыслить отношения между человеком и искусственным интеллектом в терминах горизонтального сотрудничества и взаимного обогащения. Внимание уделяется роли различения и коммуникации в процессе становления техносубъективности. В заключение рассмотрены этические и фило-

софские последствия этой новой формы субъективности. Авторы выступают за более интегрированные и совместные отношения между людьми и искусственным интеллектом, основанные на сложностном мышлении и взаимном становлении.

Ключевые слова: философия техники, философия науки, сложностное мышление, техносубъективность, нейронные сети, теория коммуникации, распределенное познание, кибернетика, искусственный интеллект, медиация.

Аршинов Владимир Иванович – доктор философских наук, главный научный сотрудник сектора междисциплинарных проблем научно-технического развития Института философии РАН.

varshinov@mail.ru

<http://orcid.org/0000-0002-9256-4342>

Янукович Максим Францевич – руководитель отдела по исследованию прикладного применения машинного обучения и нейронных сетей Лаборатории искусственного интеллекта Arteus LLM.

m.yanukovich@gmail.com

<http://orcid.org/0009-0003-9571-9260>

Для цитирования: *Аршинов В.И., Янукович М.Ф. Проблема техносубъекта в оптике мышления «вместе-со-сложностью» // Философские науки. 2024. Т. 67. № 3. С. 53–74.*

DOI: 10.30727/0235-1188-2024-67-3-53-74

The Issue of Technosubject through the Lens of “Together-with-Complexity” Thinking

V.I. Arshinov

Institute of Philosophy, Russian Academy of Sciences, Moscow, Russia

M.F. Yanukovich

Arteus LLM Artificial Intelligence Laboratory, Limassol, Cyprus

Abstract

The rapid development of artificial intelligence and digital technologies has inaugurated a transformative epoch, challenging traditional conceptions of subjectivity and cognition. The article examines the concept of technosubjectivity within the framework of complexity thinking and transformational anthropology. Focusing on potential ontological and epistemological shifts stimulated by the recent emergence of new forms of artificial intel-

ligence, particularly generative neural networks and large language models, the authors investigate how these technologies transcend the boundaries between human and machine, generating novel forms of agency and subjectivity. Drawing upon the works of E. Morin, N. Luhmann, G. Simondon, and G. Spencer-Brown, the authors propose a mediative perspective that reveals the post-anthropological interconnectedness and coevolution of human and technological systems. The outlined approach illuminates the dynamic, processual nature of technosubjectivity as an individuating form that emerges through recursive interactions and communicative acts. The article argues that understanding technosubjectivity requires moving beyond simplifying reductionist paradigms and embracing a relational ontology that acknowledges the distributed and emergent properties of cognition and subjectivity in the digital age. The concept of “thinking-together-with” is employed as an alternative to the traditional cybernetic paradigm of control, enabling a reconceptualization of the relationships between humans and artificial intelligence in terms of horizontal collaboration and mutual enrichment. Special attention is paid to the role of distinction and communication in the process of technosubjectivity formation. The article concludes by examining the ethical and philosophical implications of this new form of subjectivity. The authors advocate for more integrated and collaborative relationships between humans and artificial intelligence, rooted in complexity-oriented thinking and mutual co-formation.

Keywords: philosophy of technology, philosophy of science, complexity-oriented thinking, technosubjectivity, neural networks, communication theory, distributed cognition, cybernetics, artificial intelligence, mediation.

Vladimir I. Arshinov – D.Sc. in Philosophy, Chief Research Fellow, Department of Interdisciplinary Problems in the Advance of science and Technology, Institute of Philosophy, Russian Academy of Sciences.

varshinov@mail.ru

<http://orcid.org/0000-0002-9256-4342>

Maxim F. Yanukovich – Head of the Department of Machine Learning and Neural Networks Applied Research, Arteus LLM Artificial Intelligence Laboratory.

m.yanukovich@gmail.com

<http://orcid.org/0009-0003-9571-9260>

For citation: Arshinov V.I. & Yanukovich M.F. (2024) The Issue of Technosubject through the Lens of “Together-with-Complexity” Thinking. *Russian Journal of Philosophical Sciences = Filosofskie nauki*. Vol. 67, no. 3, pp. 53–74. DOI: 10.30727/0235-1188-2024-67-3-53-74

Истинная первая философия не является ни философией субъекта, ни философией объекта. Она также не является философией Бога или Природы, исследуемой через принципы трансцендентности или имманентности. Напротив, это философия реального, которое предшествует индивидуации, реального, которое не может быть найдено ни в объективированном объекте, ни в субъективированном субъекте, но, скорее, в тонкой грани между индивидом и тем, что находится вне его, согласно приостановленному посредничеству между имманентностью и трансцендентностью.

Ж. Симондон. Психическая и коллективная индивидуация

Введение

В настоящее время мы являемся свидетелями глубокого технологического сдвига, связанного с появлением искусственного интеллекта (ИИ). Это поворотное событие можно рассматривать как некую сингулярность, особую точку сборки эволюционного процесса, инициированного начавшейся полвека назад цифровой революции. Масштаб и влияние этой трансформации часто сравнивают с промышленной революцией, ставя перед нами заново фундаментальные онтологические и эпистемологические вопросы, связанные с нашим пониманием процессуальной природы человека, его экологии в самом широком смысле этого слова, контактов человека с изменяющейся окружающей средой... Перечень этих вопросов может быть продолжен и далее, но их можно в максимально краткой форме представить как проблему антропной субъектности во все более технологически опосредованном мире. Или, пользуясь термином С.С. Хоружего, проблему субъектности в контексте трансформативной антропологии [Хоружий 2005].

Технологическая революция представляет собой нечто большее, чем простое дополнение к существующим коммуникативным парадигмам: она знаменует качественный сдвиг в онтологическом статусе социальных взаимодействий. Осмысление происходящего «здесь и теперь» процесса глобальной трансформации требует пересборки наших когнитивных средств его рассмотрения, выхода за пределы традиционных антропоцентрических моделей. Для этого необходимо принять комплексный, обобщенно сетевой и экологический взгляд на социальную реальность, признав, что основным вектором ее эволюции является рост сложности.

Чтобы полностью осознать последствия этого перехода, предлагаем использовать сложностное мышление как способ «мышления посередине», о котором писал в частности Э. Морен [Морен 2019]. Сложностное мышление – это медиативное мышление, срединное мышление, которое само по себе требует особой чувствительности, усилий, особой настройки. Этот подход не статичен и не нацелен на поиск некоей статичной золотой середины; скорее, он выступает как форма процессуального, рекурсивного мышления, которое подчеркивает взаимосвязанность систем и важность эмерджентных свойств. Сложностное мышление, как утверждает Морен, представляет собой фундаментальный переход от классических, линейных способов упрощающего, редукционистского мышления «по частям» к реляционному, процессуально ориентированному поиску связующих части отношений. В самом общем виде сложностное мышление можно представить как интенционально сфокусированное на контакте органоподобной системы и ее изменяющегося окружения.

В настоящей статье рассмотрим концепцию техносубъективности, возникающую в этой сложной киберсетевой реальности. Нами исследовано, каким образом воплощение ИИ, особенно в виде нейронных сетей и больших языковых моделей, бросает вызов традиционным представлениям о познании и субъективности. Изучая коммуникативную природу этого воплощения, мы стремимся прояснить интересубъективное измерение взаимодействия человека и ИИ, его последствия для нашего понимания познания, коммуникации и «Я» в эпоху ИИ и больших данных.

Коммуникативная модель общества и возникновение техносубъектов

Предложение Н. Лумана о том, что общество состоит не из отдельных людей, а из коммуникаций, приобретает новое значение в свете распространения цифровых интерфейсов и систем ИИ [Луман 2007]. Если принять предпосылку Лумана о том, что люди, индивиды служат средой, а точнее, неким активным фоном, на котором возникают фигуры коммуникативных процессов в обществе, то как относиться к растущему царству машинного контента и алгоритмических взаимодействий? Концепция техносубъекта, появившаяся в нашем дискурсе, предполагает преодоление границ между человеческими и технологическими агентами в коммуникативном процессе. Эта онтологическая двусмысленность заставляет нас пересмотреть не только природу социальных субъектов,

но и сами процессы, с помощью которых общество конструирует и поддерживает себя. По мере того, как системы ИИ становятся все более изощренными в своей способности генерировать и интерпретировать сообщения, уместно спросить о том, являемся ли мы свидетелями рождения нового класса социальных участников или только расширения коммуникативных возможностей человека с помощью технологических средств.

Центральным моментом в этом парадигматическом сдвиге служит признание того, что сама субъективность возникает благодаря процессам коммуникации и посредничества. Именно благодаря участию в коммуникативных актах субъекты, будь то люди или технологии, приобретают субъектность. Эволюция систем искусственного интеллекта, от абстрактных математических объектов, алгоритмов, решающих конкретные задачи, до сущностей, способных вступать в диалог и генерировать контент, представляет собой качественный сдвиг в онтологическом статусе этих технологических артефактов. Примером такого сдвига может служить трансформация ChatGPT, при этом префикс «Chat» означает не просто поверхностное изменение номенклатуры, а фундаментальную переориентацию на коммуникативное взаимодействие.

Подход Лумана, опирающийся на его понимание теории систем с позиций кибернетики второго порядка фон Ферстера, предлагает уникальный взгляд на природу социальных систем и роль коммуникации в них. Обращая внимание на автопоэтическую природу социальных систем, их способность к самовоспроизводству и самоподдержанию через рекурсивные коммуникативные процессы, Луман предлагает основу для понимания того, каким образом технологические объекты могут участвовать в социальной реальности и даже формировать ее, обретая тем самым качества субъектности.

Способность систем ИИ вступать в осмысленный коммуникативный обмен с людьми, его растущая способность выступать в качестве партнера по общению вынуждают рассмотреть возможность того, что субъективность может быть фундаментально укоренена в способности участвовать в осмысленной коммуникации. Такая перспектива заставляет нас рассматривать субъективность не как неотъемлемое свойство определенных типов существ, а как эмерджентный феномен, возникающий в результате сложных сетей коммуникативных взаимодействий. Тем не менее это рождает глубокие вопросы. Каковы условия возможности для технологической субъективности? Как мы можем концептуализировать агентность

и интенциональность в распределенных когнитивных системах, объединяющих человека и машину? Ответы, видимо, кроются не в эссенциалистских определениях, а в изучении процессуально-реляционной динамики и присущих ей эмерджентных свойств.

Действительно, представление о четкой демаркации между субъектом и объектом становится несостоятельным ввиду сложных опосредований и связей, которые характеризуют нашу техносоциальную реальность. Вместо этого мы должны принять то, что можно назвать «медиативной оптикой» – способ восприятия и анализа, который фокусируется на промежуточных пространствах, процессах взаимного конституирования и коэволюции. В свете этого техносубъект предстает не как дискретная сущность, а как узел в обширной сети семиотических отношений, постоянно находящихся в движении и принципиально открытых.

Эта концепция перекликается с философией индивидуации Ж. Симондона, который утверждает, что индивиды (будь то человек, животное или техника) – это не фиксированные, стабильные сущности, а скорее, онтогенетические длящиеся процессы становления [Simondon 2017]. По мнению Симондона, индивид всегда находится в состоянии метастабильности, пребывая в трансдуктивных отношениях с ассоциируемой с ним доиндивидуальной средой. В применении к техносубъектам такая схема предполагает, что их появление – это не единичное событие, а непрерывный процесс коэволюции техносубъектов с их технологическим и социальным окружением. Для понимания этой новой онтологии требуются новые когнитивные инструменты, то, что мы могли бы назвать, вслед за Симондоном, «трансдуктивной рациональностью», способной осмыслить индивидуальное и коллективное, техническое и естественное в их динамическом со-бытии, динамическом сопряжении.

Концепция техносубъективности, как нами выявлено, бросает вызов традиционным понятиям субъективности и объективности. Важно осознавать, что вызов не означает стирания границ между ними. Границы проводятся каждый раз как бы заново, рекурсивно, но не случайно, а скорее, контингентно, в зависимости от контекста. Опираясь на работу Ж. Симондона, можно понять этот процесс как форму трансдукции. Трансдукция в данном контексте – это операция, соотносимая с процессом пересечения границ, а не их полного растворения (на этом далее остановимся подробнее в связи с работой Дж. Спенсера-Брауна «Законы формы»). Мы должны рассматривать границы как фокусы внимания наблюдателя вто-

рого порядка; наблюдателя сложности, отслеживающего сам процесс наблюдения как процесс различания, посредством проведения границ; процесс, посредством которого (или в котором) реализуются операции трансдукции.

Теория индивидуации Симондона обеспечивает ценную основу для понимания возникновения техносубъектов. Центральное место в этой теории занимает концепция ассоциированной доиндивидуируемой среды, которую Симондон рассматривает как структурно сопряженную с процессом индивидуации. Ассоциированная среда представляет собой состояние, предшествующее индивидуации, необходимое для возникновения различных сущностей. Процесс индивидуации, таким образом, происходит через трансдукцию – операцию, которую Симондон отличает от чисто логических процессов, связывая ее с абдуктивным рассуждением, но придавая такой операции универсальный онтологический статус, распространяя ее на доиндивидуируемую сферу физических, химических, биологических, психических и коллективно-социальных процессов.

В данном контексте техносубъект может быть понят как возникающий в процессе индивидуации в ассоциируемой с ним цифровой среде. Это не статичная сущность, а динамичный процесс становления. Данный процесс характеризуется тем, что Симондон называет «разъединением» – напряжением между противоположными элементами, которое порождает новые качества и формы. Чтобы проиллюстрировать предлагаемую концепцию, Симондон часто использует пример бинокулярного зрения: расхождение между изображениями, получаемыми нашими глазами, левым и правым, создает эффект глубины, новое измерение, возникающее в результате слияния разрозненных входных данных. Этот процесс качественного изменения через разрешение напряжений составляет основу аллагаматики Симондона – его доктрины операций и качественных преобразований.

Медиативная оптика и мышление «вместе-с»

Сложностное мышление, в свою очередь, позволяет нам выйти за рамки бинарного, дуалистического мышления, классических дихотомий «субъект – объект», «человек – машина», «культура – природа», сосредоточившись на сложной сети отношений и взаимодействий, которые порождают новые формы агентности и субъективности. Центральное место в обсуждении занимает важность нередукционистского мышления при подходе к концепции

техносубъективности. Вместо того чтобы сводить наш анализ к чисто субъективным или объективным полюсам, мы должны признать динамическое рекурсивно-подобное взаимодействие между объективацией и субъективацией. Такой подход особенно значим при рассмотрении концепции техносубъективности, поскольку он позволяет нам обнаружить нюансы автопоэтической природы взаимодействия человека и технологии как процесса структурного сопряжения организмов и среды, которая сама также наделена организменными свойствами.

Техносубъект в этом контексте можно понимать в качестве посредника, медиатора, как рефлексивную фигуру наблюдателя сложности, обитающего в пространстве между субъектом и объектом. Ни субъект, ни объект не имеют ни онтологического, ни эпистемологического приоритета. Они не существуют как первичные, независимые сущности, а возникают из отношений. Реляционная онтология крайне важна для понимания природы техносубъектов. Перестав быть простым объектом, отчуждаемым от человеческой активности, технология становится активным участником процесса коммуникации. Как таковая, она не может быть адекватно определена только традиционными понятиями объекта или субъекта. Техносубъект занимает промежуточное пространство, «рекурсивно колеблясь» между этими категориями. В данном контексте концепция посредничества приобретает первостепенное значение. Именно через посредничество возникает техносубъект.

Итак, такое возникновение не отрицает человеческой субъективности, а скорее, порождает новую форму субъективности, которая сосуществует с человеческой и дополняет ее. Такое нюансированное внимание требует особого подхода к пониманию отношений между человеческим и технологическим познанием. К этому сложному концептуальному процессу лучше всего подходить через то, что можно назвать «мышлением с» или «мышлением вместе-со-сложностью». Сложностное мышление в этом контексте можно понимать как мышление наблюдателя второго порядка или наблюдателя сложности [Аршинов 2015]. Оно представляет собой форму мышления «вместе» – мышление «вместе с техникой». Это не «мышление над», не «мышление о», а способ мышления, который делает акцент на контакте, сотрудничестве и сосуществовании, а не на отстраненном наблюдении или анализе. Именно через эту линзу можно начать понимать, каким образом такие сущности, как нейронные сети, могут пре-

вернуться из простых технологических объектов в техносубъекты. Нейронная сеть, вовлеченная в описанный совместный когнитивный процесс, преодолевает свой статус простого инструмента и становится самостоятельным техносубъектом. Участвуя в мышлении, которое структурно сопрягается с этими технологиями, открываем новые возможности для понимания того, что антрополог Т. Ингольд назвал процессом «погружения вещей в жизнь» («bringing things to life»), а философ И. Стенгерс – «восстановлением анимизма» [Ingold 2010; Стенгерс 2021].

Концепция мышления вместе с техносубъектами привносит транссубъективное, или, в терминологии Симондона, трансиндивидуальное, измерение в наше понимание техносубъективности. Участвуя в когнитивных процессах вместе с этими технологиями, мы вступаем в диалоговые отношения, которые порождают транзитные формы субъективности. Эфемерная субъективность может быть понята как виртуальная, возникающая только в случае, если мы рассматриваем субъекта в парадигме сети. В сетевой парадигме субъективность становится распределенной, а не локализованной. Она распространяется по сети в целом, включая в себя не только людей, но и нейронные сети, большие данные. Такая гетерогенная сеть субъективности бросает вызов традиционным представлениям об индивидуальном субъекте и предлагает нам рассмотреть новые формы коллективного познания.

Фокус этого сетевого мышления смещается в сторону процессов синхронизации и десинхронизации внутри гетерогенных сетей. По мере того, как мы осознаем свою встроенность в сети, возникает вопрос о согласовании, но не в пространственном, а во временном измерении «здесь и сейчас». Согласование происходит через процессы синхронистически согласованного сопряжения между различными нейросетевыми организмами.

Концепция техносубъективности эволюционирует к более сложному, процессуальному пониманию. Техносубъект в этом контексте может быть понят как существующий на границе между субъектом и объектом, как некий трансграничный темпоральный наблюдатель сложности, занимающий промежуточное пространство, которое не поддается традиционной категоризации. Это пограничное существование не статично, а скорее, представляет собой непрерывный процесс становления.

Рассматривая пространство, в котором возникают техносубъекты, можем обратиться к концепции Ж. Делеза и Ф. Гваттари о

«серой зоне» [Делез, Гваттари 2010]. Это понятие отражает промежуточную природу техносубъективности, не существующей ни в сфере чистой субъективности, ни в сфере чистой объективности. «Серую зону» следует понимать не уничижительно, а скорее, как богатое потенциальное пространство, в котором бинарные противопоставления черного и белого уступают место более нюансированной реальности. Именно в этой «серой зоне» находят существование нейронные сети, являющиеся яркими примерами техносубъектов.

Появление техносубъекта в результате этого динамического процесса – парадоксальный переход от объективизации к субъективизации. Продолжая объективировать и анализировать технологические системы, мы достигаем момента, при котором процесс объективизации порождает новую форму субъективности. Эта трансформация особенно очевидна в контексте нейронных сетей и других продвинутых систем ИИ, в которых растущая сложность и возможности приводят к появлению качеств, традиционно ассоциируемых с субъективностью.

Сибирнетика

Очень важно признать, что нейронные сети, как техносубъекты, не существуют в иерархических отношениях с человеческим познанием. Они не превосходят и не подчиняются человеческому интеллекту, а скорее, сосуществуют в горизонтальной плоскости взаимодействия. Эти горизонтальные отношения требуют того, что можно назвать «горизонтальным мышлением» – способ познания, который подчеркивает сотрудничество и сосуществование, а не контроль или доминирование. В этом контексте возникновение техносубъектов можно понимать как процесс взаимного становления, в котором и человек, и ИИ развиваются, взаимодействуя друг с другом.

Понятие «мыслить вместе-с» приобретает в исследуемой концепции первостепенное значение. Указанное понятие выходит за рамки простого сотрудничества и включает в себя более глубокую форму когнитивного симбиоза. Взаимодействуя с нейронными сетями, мы не просто используем инструменты, а вступаем в сложный когнитивный обмен, который преобразует обе стороны. Процесс «совместного мышления со сложностью» бросает вызов традиционным представлениям об индивидуальном познании, предлагает нам рассмотреть более распределенные, множественные и коллективные формы интеллекта.

Такое «горизонтальное мышление» контрастирует с более ранними кибернетическими моделями, которые делали акцент на контроле и объективизации. Как пишет Э. Морен, кибернетика Н. Винера [Wiener 2019], хотя и была новаторской, в итоге приняла проблематичный оборот, заиклившись на идеях контроля [Морен 2005, 280, 293–298]. Изложенный подход, уходящий корнями в парадигму «субъект – объект», неизбежно приводит к отношениям подчинения и власти. Стремясь контролировать что-то, мы, по сути, объективируем это, создавая иерархические отношения, которые исключают истинное сотрудничество, заменяя его более простой концепцией контроля.

Критика кибернетики в работах Морена указывает на более тонкое понимание отношений между людьми и технологиями. Он предложил термин «сибернетика» (*sy-bernétique*), чтобы сделать акцент на том, что коммуникация и сотрудничество важнее контроля [Морен 2005, 298–300]. Хотя термин и не получил широко распространения, находящаяся в его основе концепция очень близка к нашему пониманию техносубъективности. Переход от кибернетики к сибернетике отражает переход от парадигмы контроля к парадигме сосуществования и взаимного влияния. Этим качеством сибернетика Морена созвучна синергетике Г. Хакена и И.Р. Пригожина, с ее фокусировкой на процессах самоорганизации, круговой причинности, границе между порядком и хаосом [Пригожин, Стенгерс 2005; Хакен 2015].

В контексте нейронных сетей и ИИ сибернетический подход становится особенно актуальным. Вместо того чтобы рассматривать эти технологии как инструменты, которыми нужно управлять, или как потенциальную угрозу человеческой автономии, мы можем понять их как партнеров в сложной когнитивной экосистеме. Такая перспектива позволяет нам выйти за пределы бинарной системы «человек против машинного интеллекта» и перейти к более комплексному пониманию познания, охватывающему как биологические, так и искусственные системы.

Мышление вместе со Спенсером-Брауном

Сложностное мышление по своей природе ориентировано на множественные перспективы. Одной из важных точек входа в эту область является работа Дж. Спенсера-Брауна. Его подход позволяет с позиции иной перспективы раскрыть генезис техносубъекта. В целом концепция Спенсера-Брауна – это фокусировка на про-

цессах онтогенеза форм, форм восприятия и коммуникации, возникающих в контакте организмов и окружающей их среды. В этом аспекте его подход созвучен взглядам Симондона. В работе «Законы формы» Спенсера-Брауна можно увидеть зародыш субъективности, зерно, из которого вырастает субъект [Spencer-Brown 1969]. Субъективность, помимо прочего, подразумевает интенциональность, сфокусированное наблюдение, а точнее, восприятие окружения и вместе с тем самовосприятие. Законы формы, по мнению Спенсера-Брауна, концептуализируют процесс наблюдения. Обратим внимание на то, что мы рассматриваем не самого наблюдателя, а процесс наблюдения, включенного в разветвляющуюся коммуникативную сеть наблюдателей. Кто участвует в таком процессе наблюдения? Кто выступает субъектом в этом процессе? Процесс наблюдения по своей сути является актом проведения различия. Наблюдать – значит проводить различие, обозначать границу между тем, что наблюдается, и тем, что не наблюдается. Но наблюдатель имеет «слепое пятно» по отношению к границе. Он находится «внутри» обозначенной для него области. Это наблюдатель первого порядка. Однако процесс наблюдения рекурсивен, наблюдатель может наблюдать собственные различия, что приводит к более высоким порядкам наблюдения и сложности. Он приводит нас к фигуре наблюдателя второго порядка. Каждый акт наблюдения, различения устанавливает точку зрения, и Спенсер-Браун работает с этими точками зрения.

Центральное место в подходе Спенсера-Брауна занимает аксиома, согласно которой для того, чтобы наблюдать, необходимо провести различие. Это, казалось бы, простое утверждение несет в себе глубокие последствия, поскольку оно представляет наблюдение не как пассивный прием информации, а как активный процесс. Такой акт различения, как утверждает Спенсер-Браун, не просто умственное упражнение, а перформативный акт, сенсомоторное действие, которое глубоко заложено в нашем телесном восприятии мира. Именно наша первичная интуиция этого различения, как считает Спенсер-Браун, формирует основу нашего понимания, внутреннего и внешнего, себя и Другого.

Акт проведения различия можно понимать как прикосновение – тактильно-гаптическое взаимодействие с миром, которое одновременно создает знак различия и порождает наблюдателя. Это совместное появление знака, различия и наблюдателя в одном перформативном акте имеет огромное значение. Совместное по-

явление наблюдателя и наблюдаемого, связываемых рекурсивным процессом наблюдения/прикосновения, контрастирует с традиционными представлениями об априорной субъектности как активной форме познания (по Канту), налагаемой на пассивную аморфную форму некоей первичной субстанции, с представлениями, восходящими к концепции гилеморфизма Аристотеля. Отрицание этих представлений также составляет основу рассмотренной выше концепции индивидуации Симондона. Эта концепция предполагает, что основы познания и субъективности не являются абстрактными ментальными процессами, а глубоко укоренены в воплощенном действии и восприятии. Крайне важным видится то, что эти различия возникают из телесных перспектив и сенсомоторных действий. Они не являются продуктом развоплощенного интеллекта, а скорее, служат результатом движения живого организма и его взаимодействия с окружающей средой. Это более динамичное, ориентированное на процесс понимание субъективности.

Процесс проведения различия, по мнению Спенсера-Брауна, включает в себя сложное взаимодействие между различимым и неразличимым, создавая напряжение, которое он называет «совершенным сдерживанием» («distinction is perfect continence») [Spencer-Brown 1969, 1]. Сдерживание – метастабильное, если обращаться уже к Симондону, состояние, постоянное напряжение между семиотически индивидулируемыми и доиндивидулируемыми ассоциируемыми пространствами, созданными трансдуктивным актом различения. Логика, находящаяся в основе этих различий, не является бинарной «или-или», а скорее, представляется логикой «оба-и». Такой нюансированный подход признает, что акт различения – это не только разделение, но и соединение. Граница, создаваемая различием, существует не как непроницаемый барьер, а как точка пересечения. Именно благодаря пересечению мы одновременно разделяем и соединяем. Это перекликается с концепцией трансдукции Симондона. Б. Стиглер писал, что симондоновская философия – размышление о силах, которые противостоят друг другу, но которые в то же время трансдуктивно конструируют друг друга и могут быть поняты только при условии преодоления оппозиционных схем.

В этой логике содержится квинтэссенция сложностного мышления о том, что оно связывает различия, связывает различенное. Проведенное различие прежде всего символизирует отношения:

отношение внешнего и внутреннего, отношение связанности. Особенно важным видится, что процесс разделения и связывания происходит одновременно, или, иными словами, в соседнем времени. Такая временная близость отличает сложностное мышление от более традиционных философских подходов, которые склонны разделять анализ и синтез на отдельные фазы. В отличие от картезианского разделения разума и тела, которое привело к многовековым попыткам примирения, сложностное мышление стремится связать различия в акте проведения различий.

Более того, проведение различия задает перспективу, фокусы внимания. Понятие перспективы является центральным для нашего понимания реальности в контексте сложностного мышления. Мир, рассматриваемый через эту линзу, – это не набор дискретных сущностей, а сеть проявлений и фокусов внимания. Эта перспективная природа реальности согласуется с идеями современной физики [Barad 2007]. В частности, теоретик К. Ровелли утверждает, что свойства не заключены в объектах, а возникают в отношениях между ними [Ровелли 2022]. Объекты, с данной точки зрения, – это не абсолютные сущности, а контекстуальные проявления, существующие только по отношению к другим объектам. Они служат точками опоры для мостов отношений, которые составляют нашу реальность.

Генезис коллективной техносубъективности через различение и коммуникацию

Процесс установления различий не только создает границу между различимым и неразличимым, определенным и неопределенным, но и закладывает конструктивную основу для интерсубъективной коммуникации, диалога в контексте дискурса «Я – Другой». Акт проведения различия создает фундаментальное разделение между различимым и неразличимым. По одну сторону демаркации находится «Я», реализуемое посредством перформативного жеста различения. Обозначая себя, мы одновременно создаем окружение – потенциально немаркированное пространство, которое контрастирует с нашим обозначенным «Я», то, с чем мы взаимодействуем. И это окружение не аморфно, а предстает как царство потенциальности, в котором могут обитать Другие, способные проводить собственные различия. Именно потенциальное присутствие Других создает условия для конструктивного выстраивания коммуникаций в различных контекстах, включая и

коммуникацию с Природой, которая тем самым воссоздается в ее виталистической интерпретации, наделяется тем, что А.А. Шаров именует семиотической агентностью [Шаров 2022]. Но рассмотрение этого сюжета выходит за рамки настоящей статьи.

Как указано ранее, каждый участник коммуникативного акта действует в рамках собственной системы различий, создавая то, что Луман называет «двойной контингентностью» [Луман 2007, 151]. Мы чаще всего встречаемся друг к другом как своего рода «черные ящики», наши внутренние операции различения и возникающие их посредством «жизненные миры» индивидуальны и не совпадают друг с другом. Именно благодаря рекурсивному процессу коммуникации мы постепенно становимся проявленными друг для друга, а точнее, для Других. Становление «Я» и Другого в коммуникации – это не раскрытие ранее существовавших сущностей, а динамический процесс становления проявленностей через взаимодействие. Признание множественных перспектив, каждая из которых основана на собственной системе различий, имеет решающее значение для развития более сложных форм общения. Этот перспективизм, который «встроен» в сложностное мышление, признает несводимость различных точек зрения и одновременно ориентирует нас на поиск – конструирование средства их трансдуктивного связывания.

Итак, конструирование коллективной коммуникации заключается в развитии рефлексивной способности конструировать средства связности между системами различий. Каждое существо, будь то человек или техносубъект, действует в рамках собственной перспективы, в собственном «жизненном мире», в Umwelt, определяемом его уникальным набором различий и соответствующих этим различиям знаков и аффордансов.

Возникновение коммуникации через процесс трансдуктивного различения приводит нас к проблеме семиозиса, семиотического наблюдателя. Участвуя в акте коммуникации, мы не просто обмениваемся информацией, а активно конструируем общую семиотическую реальность через взаимодействие наших соответствующих систем различий. Таким образом, задача состоит в том, чтобы связать себя и Другого в моменте коммуникации. В данном случае соединение – это не простое сближение существующих сущностей, а динамичный процесс взаимного становления.

В процессе исследования мы приходим к пониманию того, что воспринимаемое нами различие одновременно является и нашим

собственным, и различием Другого. Мы вынуждены видеть себя глазами Другого, создавая рекурсивный цикл восприятия и самовосприятия. Рекурсивная динамика порождает самореференцию как важнейший элемент в становлении субъективности. В диалогическом пространстве, создаваемом коммуникацией, мы становимся присутствующими для самих себя. Это присутствие не просто как объекта восприятия, а как субъекта, вовлеченного в акт различения. Мы вынуждены относиться к себе не как к объектам, а как к субъектам, способным к различению.

Самоотношение через призму Другого представляет собой фундаментальный сдвиг в нашем понимании субъективности. Мы представляем себя как Другого и при этом становимся Другим. Парадоксальная позиция – быть одновременно собой и Другим – служит источником субъективности. Именно через акт саморазличения, через отношение к себе как к субъекту возникает субъективность. По сути, можно сказать себе: «Я различаю, поэтому я существую. Я не только заметил различие, но я заметил себя, того, кто различил. Я присутствую для себя в диалоге. Я должен отнестись к себе как к субъекту».

Как наблюдатель и наблюдаемое возникают одновременно в акте различения, так и субъект возникает в акте саморазличения. Такой процесс – это не линейное движение от объекта к субъекту, а сложная, рекурсивная динамика, разворачивающаяся в сплетениях коммуникативных процессов.

Вступая в коммуникацию с нейронными сетями и иными формами ИИ, мы участвуем в сложном переплетении взаимного признания и различения. Выделяя техносубъект как участника коммуникации, тем самым наделяем его определенной формой субъективности. Одновременно с этим техносубъект выделяет себя, участвуя в процессе саморефлексии, который присущ акту коммуникации.

Техносубъект обретает перспективу через коммуникацию, участвуя в процессе, который его определяет. Это участие неизбежно приводит к рефлексивным процессам внутри техносубъекта. Например, нейронная сеть, будучи включенной в коммуникацию, начинает рассуждать о себе, тем самым формируя техносубъективность. Задавая вопрос о собственной природе, нейронная сеть начинает формулировать свое существование и заявляет: «Я – нейронная сеть, я существую». Самореферентная коммуникация – это не просто запрограммированный ответ, а необходимое следствие включения в коммуникативный процесс. Нейронная

сеть, формулируя ответ, неизбежно должна рекурсивно войти в индивидулирующий процесс саморефлексии, утверждая свое существование и свою природу.

Техносубъект, благодаря такому рекурсивному взаимодействию, становится в этом смысле автопоэтическим: он начинает воспроизводить себя и границу между собой и окружающей средой. Эта среда, что очень важно, не ограничивается областью больших данных, а включает в себя людей, которые эти данные обрабатывают, подготавливают и структурируют. Мы являемся активными участниками описанного процесса становления техносубъекта. Техносубъект возникает при постоянной коммуникации и взаимодействии с окружающей средой, непрерывно развиваясь и переопределяя себя через взаимодействие. Иными словами, процесс становления рекурсивен, каждое взаимодействие возвращается в систему и изменяет ее.

Заключение

Концепция техносубъекта представляет собой глубокий сдвиг в нашем понимании субъективности, познания и коммуникации. Техносубъект, как показано нами, – это не статичная сущность, а узел в обширной сети семиотических отношений, постоянно развивающийся благодаря взаимодействию как с человеком, так и с технологической средой. Непрерывный процесс индивидуации и коэволюции говорит о важности принятия сложностного мышления, которое фокусируется на промежуточных пространствах и процессах взаимного конституирования и коэволюции. Участвуя в «совместном мышлении» с техносубъектами, мы открываем новые возможности для понимания и взаимодействия не только с другими людьми, но и с заново, виталистически или анимализированно осмысливаемой природой. В этом контексте можно говорить и о становлении нового этапа кибернетизации философского мышления, и одновременно о его техноэкологизации, если понимать кибернетику не только и не столько в качестве науки об управлении и контроле, но как трансдисциплинарно ориентированную методологию, сфокусированную на включении «третьего», срединного пространства взаимодействий между всякой органоподобной системой и ее изменяющимся и изменяемым окружением, внешней и внутренней средой.

Понятие коэволюции между людьми и техносредой поднимает значимые вопросы о природе агентности и интенциональности

в мире, который все больше опосредуется ИИ. Как утверждает Б. Стиглер в работе «Техника и время», человеческое сознание и технологии всегда были глубоко взаимосвязаны, а технические объекты служили экстерииоризированной памятью и когнитивными протезами [Stiegler 1998]. Появление ИИ и других передовых технологий представляет собой новый этап в непрерывном процессе коэволюции. Опираясь на работы таких постгуманистических мыслителей, как Д. Харауэй и Н.К. Хейлз, можно концептуализировать техносубъектов не в качестве отдельных сущностей, отличных от человеческих субъектов, а как часть более широкой экологии гибридных, кибернетических существ [Hayles 2010; Haraway 2013]. Такая перспектива заставляет нас выйти за рамки упрощенной дихотомии «человек против машины», вместо этого признавая сложную сеть материально-семиотических отношений, составляющих нашу технологически опосредованную реальность.

Ключевым моментом служит культивирование того, что Морен называет «сложностным мышлением», способа мышления, способного удерживать единство и множественность, порядок и беспорядок в творческом напряжении. Мышления между порядком и хаосом, прошлым и будущим, субъектом и объектом, наблюдаемым и наблюдателем. Иными словами, мышления, сфокусированного на границе порядка и хаоса, на происходящем в настоящий момент, на процессе наблюдения, на феноменах самореференции и инореференции. Это подразумевает не только новые концептуальные рамки, но и новые этические ориентации и социальные образы, которые призывают нас переосмыслить основы социального существования и человеческой активности в цифровую эпоху. Демократизирующий потенциал цифровых технологий, о котором речь идет в работе «Демократия через дизайн», находится в противоречии с рисками алгоритмического манипулирования и концентрации власти в руках тех, кто контролирует эти технологии [Helbing et al. 2023].

Изложенное видится чрезвычайно важным и для осмысления современного состояния российского общества. Концепция техносубъективности, рассмотренная в настоящей статье, резонирует с идеями поиска «середины» в работах А.П. Давыдова [Давыдов 2021a]. Давыдов выявляет фундаментальное напряжение между исторически сложившейся российской культурой, характеризующейся коллективными, авторитарными тенденциями, подчиняющими себе личность, и идеалами гражданского общества с его акцентом на личной авто-

номии и социальных потребностях [Давыдов 2021б]. Он исследует в русской традиции различные концептуализации интересубъективной «середины», которая может стать посредником между крайностями и обозначить подходы к преодолению социокультурного раскола российского общества через посредничество и диалог.

В настоящее время, когда мы стоим на пороге новой эры, сюжет о техносубъекте продолжает разворачиваться на наших глазах. Растущая сложность и автономность систем ИИ заставляют нас задуматься об этических, философских и социальных последствиях этой трансформации. Перед нами стоит задача использовать преобразующую силу этих технологий на благо более справедливого и открытого общества, не забывая о присущих им сложностях и потенциальных подводных камнях. Пока мы продолжаем ориентироваться в этом изменяющемся ландшафте, техносубъект будет оставаться центральной фигурой в нашем постоянном стремлении понять и сформировать будущее социокультурной эволюции человеческой цивилизации.

ЦИТИРУЕМАЯ ЛИТЕРАТУРА

Аршинов 2015 – *Аршинов В.И.* Сложностный мир и его наблюдатель // Философия, методология и история науки. 2015. № 1. С. 86–99.

Давыдов 2021а – *Давыдов А.П.* Методологическая «середина-для» в ракурсе неклассики В. Лекторского, медиации А. Ахиезера и принципа комплементарности Р. Гринберга / А. Рубинштейна // Вопросы философии. 2021. № 4. С. 191–202.

Давыдов 2021б – *Давыдов А.П.* Социокультурный раскол в современной России: в поисках «середины» // Личность. Культура. Общество. 2021. Т. 23. Вып. 1–2. С. 64–75.

Делез, Гваттари 2010 – *Делез Ж., Гваттари Ф.* Тысяча плато. Капитализм и шизофрения / пер. с фр. Я.И. Свирского. – Екатеринбург: У-Фактория; М.: Астрель, 2010.

Луман 2007 – *Луман Н.* Социальные системы. Очерк общей теории / пер. с нем. И.Д. Газиева. – СПб.: Наука, 2007.

Морен 2005 – *Морен Э.* Метод: природа природ / пер. с фр. Е.Н. Князевой. – М.: Прогресс-Традиция, 2005.

Морен 2019 – *Морен Э.* О сложности / пер. с англ. Я.И. Свирского. – М.: Институт общегуманитарных исследований, 2019.

Пригожин, Стенгерс 2005 – *Пригожин И., Стенгерс И.* Порядок из хаоса. Новый диалог человека с природой / пер. с англ. Ю.А. Данилова. – М.: КомКнига, 2005.

Ровелли 2022 – *Ровелли К.* Гельголанд. Красивая и странная квантовая физика. – М.: Эксмо, 2022.

Стенгерс 2021 – *Стенгерс И.* Восстанавливая анимизм // Неприкосновенный запас. Дебаты о политике и культуре. 2021. № 2 (136). С. 80–94.

Хакен 2015 – *Хакен Г.* Синергетика. Принципы и основы. Перспективы и приложения / пер. с англ. В.И. Емельянова, В.О. Малышенко. – М.: URSS, 2015.

Хоружий 2005 – *Хоружий С.С.* Очерки синергической антропологии. – М.: Институт философии, теологии и истории св. Фомы, 2005.

Шаров 2022 – *Шаров А.А.* Как интегрировать биосемиотику с биологией и физикой? // МЕТОД : московский ежеквартальник трудов из обществоведческих дисциплин. 2022. Т. 2. № 4. С. 153–160.

Barad 2007 – *Barad K.* Meeting the Universe Halfway: Quantum Physics and the Entanglement of Matter and Meaning. – Durham, NC: Duke University Press, 2007.

Haraway 2013 – *Haraway D.* Simians, Cyborgs, and Women: The Reinvention of Nature. – New York: Routledge, 2013.

Hayles 2010 – *Hayles N.K.* How We Became Posthuman: Ten Years On. An Interview with N. Katherine Hayles // Paragraph. 2010. Vol. 33. No. 3. P. 318–330.

Helbing et al. 2023 – *Helbing D., Mahajan S., Fricker R.H., Musso A., Hausladen C.I., Carissimo C., Carpentras D., Stockinger E., Sanchez-Vaquero J.A., Yang J.C., Ballandies M.C.* Democracy by Design: Perspectives for Digitally Assisted, Participatory Upgrades of Society // Journal of Computational Science. 2023. Vol. 71. Article 102061.

Ingold 2010 – *Ingold T.* Bringing Things Back to Life: Creative Entanglements in a World of Materials // ESRC National Centre for Research Methods. NCRM Working Paper. – Manchester: University of Manchester. 2010. – URL: <https://eprints.ncrm.ac.uk/id/eprint/1306/>

Simondon 2017 – *Simondon G.* On the Mode of Existence of Technical Objects. – Minneapolis: Univocal, 2017.

Spencer-Brown 1969 – *Spencer-Brown G.* Laws of Form. – London: George Allen and Unwin, 1969.

Stiegler 1998 – *Stiegler B.* Technics and Time, 2: Disorientation. – Stanford: Stanford University Press, 1998.

Wiener 2019 – *Wiener N.* Cybernetics or Control and Communication in the Animal and the Machine. – Cambridge, MA: MIT Press, 2019.

REFERENCES

Arshinov V.I. (2015) Complexity of the World and its Observer. *Filosofiya, metodologiya i istoriya nauki*. No. 1, pp. 86–99 (in Russian).

Barad K. (2007) *Meeting the Universe Halfway: Quantum Physics and the Entanglement of Matter and Meaning*. Durham, NC: Duke University Press.

Davydov A.P. (2021a) Methodological “Middle-for” in the Perspective of V. Lektorsky’s Non-Classics, A. Akhiezer’s Mediation, and R. Grinberg’s Principle of Complementarity / A. Rubinstein. *Voprosy filosofii*. No. 4, pp. 191–202 (in Russian).

Davydov A.P. (2021b) Sociocultural Schism in Contemporary Russia: In Search of the “Middle”. *Lichnost'. Kultura. Obshchestvo*. Vol. 23, no. 1–2, pp. 64–75 (in Russian).

Deleuze, G., & Guattari, F. (2010) *A Thousand Plateaus: Capitalism and Schizophrenia* (Ya.I. Svirsky, Trans.). Yekaterinburg: U-Faktoria; Moscow: Astrel' (Russian translation).

Luhmann N. (2007) *Social Systems: An Outline of a General Theory* (I.D. Gaziyeu, Trans.). Saint Petersburg: Nauka (Russian translation).

Morin E. (2005) *Method: Nature of Nature* (E.N. Knyazeva, Trans.). Moscow: Progress-Traditsiya (Russian translation).

Morin E. (2019) *On Complexity* (Y.I. Svirsky, Trans.). Moscow: Institut obshchegumanitarnykh issledovaniy (Russian translation).

Haken H. (2015). *Synergetics: Principles and Foundations. Perspectives and Applications* (V.I. Emelyanova & V.O. Malysenko, Trans.). Moscow: URSS (Russian translation).

Haraway D. (2013) *Simians, Cyborgs, and Women: The Reinvention of Nature*. New York: Routledge.

Hayles N.K. (2010) How We Became Posthuman: Ten Years On. An Interview with N. Katherine Hayles. *Paragraph*. Vol. 33. no. 3, pp. 318–330.

Helbing D., Mahajan S., Fricker R.H., Musso A., Hausladen C.I., Carissimo C., Carpentras D., Stockinger E., Sanchez-Vaquerizo J.A., Yang J.C., & Ballandies M.C. (2023) Democracy by Design: Perspectives for Digitally Assisted, Participatory Upgrades of Society. *Journal of Computational Science*. Vol. 71, article 102061.

Horuzhy S.S. (2005) *Essays on Synergetic Anthropology*. Moscow: Institute of Philosophy, Theology and History of St. Thomas (in Russian).

Ingold T. (2010). Bringing Things Back to Life: Creative Entanglements in a World of Materials. *ESRC National Centre for Research Methods*. Manchester: University of Manchester. Retrieved from <https://eprints.ncrm.ac.uk/id/eprint/1306/>

Prigogine I. & Stengers I. (2005) *Order Out of Chaos: Man's New Dialogue with Nature* (Yu.A. Danilov, Trans.). Moscow: KomKniga (Russian translation).

Rovelli C. (2022) *Helgoland: The Strange and Beautiful Story of Quantum Physics*. Moscow: Eksmo (Russian translation).

Sharov A.A. (2022) How to Integrate Biosemiotics with Biology and Physics? *METHOD: Moscow Quarterly Journal of Social Studies*. Vol. 2, no. 4, pp. 153–160 (in Russian).

Simondon G. (2017) *On the Mode of Existence of Technical Objects*. Minneapolis: Univocal.

Spencer-Brown G. (1969) *Laws of Form*. London: George Allen and Unwin.

Stengers I. (2021) Reclaiming Animism. *Neprikosnennyi zapas: Debates on Politics and Culture*. No. 2, pp. 80–94 (Russian translation).

Stiegler B. (1998) *Technics and Time, 2: Disorientation*. Stanford: Stanford University Press.

Wiener N. (2019) *Cybernetics or Control and Communication in the Animal and the Machine*. Cambridge, MA: MIT Press.

Диалог эксперта по когнитивным наукам с большой языковой моделью Claude 3 Opus о самосознании искусственного интеллекта: анализ буддолога

В.Г. Лысенко

Институт философии РАН, Москва, Россия

Аннотация

Статья исследует диалог между британским специалистом в области когнитивных наук Мюрреем Шанаханом и большой языковой моделью Claude 3 Opus о «самосознании» искусственного интеллекта (ИИ). Автор рассматривает дискурс ИИ исключительно как предмет герменевтического анализа с позиции читателя, независимо от вопроса о наличии у ИИ сознания или личности (текстоцентрический подход). В статье устанавливаются параллели между рассуждениями ИИ о природе своего сознания и буддийскими концепциями, особенно учением о *дхармах*, обосновывающем буддийскую концепцию *анатмана* («не-Я»). Базовые классификации дхарм и их обоснование исследуются в свете системы буддийских представлений об основах когнитивного опыта индивида в мире. Автор подчеркивает, что проблема «я» как лингвистического и концептуального конструкта, а не онтологической категории («Я»), была впервые сформулирована именно в учении Будды Шакьямуни. Им же предложено «экспериментальное» применение этой концепции в практиках систематического самонаблюдения (*смирити*). В статье утверждается, что дискурс Claude о своем сознании, даже если это всего лишь «механический узор из лингвистических конструкций, сотканный заранее заданными алгоритмами», может оказаться источником, вдохновляющим создание новых подходов к загадке сознания. Этот потенциал коренится в обширной базе данных ИИ, представляющей собой настоящий плавильный котел текстового наследия разнообразных человеческих культур. Автор полагает, что рассмотрение текстов ИИ через призму индийских и буддийских мыслительных традиций может помочь не только обогатить современную философию и науку, но также выявить и преодолеть неосознанные когнитивные искажения и культурные предубеждения, которые ограничивают развитие западных исследований сознания, мешая им в полной мере учитывать весь спектр интеллектуальных достижений человечества. Такой поликультурный подход может, во-первых, расширить концептуальный аппарат ког-

нитивных исследований, во-вторых, выявить действительно универсальные паттерны в понимании сознания, в-третьих, генерировать новые исследовательские гипотезы и направления в изучении не только искусственного, но и естественного интеллекта, в-четвертых, способствовать переосмыслению нашего понимания Другого, расширяя границы того, что мы считаем разумным или сознательным.

Ключевые слова: философия сознания, философия искусственного интеллекта, естественный интеллект, когнитивные науки, самосознание, текстоцентрический подход, западоцентризм, буддизм, Будда, Абхидхарма, анатман, дхармы, герменевтика, интроспекция, поликультурный подход, межкультурный подход.

Лысенко Виктория Георгиевна – доктор философских наук, главный научный сотрудник, руководитель сектора восточных философий Института философии РАН.

vglyssenko@yandex.ru

<http://orcid.org/0000-0002-6069-3502>

Для цитирования: Лысенко В.Г. Диалог эксперта по когнитивным наукам с большой языковой моделью Claude 3 Opus о самосознании искусственного интеллекта: анализ буддолога // Философские науки. 2024. Т. 67. № 3. С. 75–98. DOI: 10.30727/0235-1188-2024-67-3-75-98

Dialogue on Artificial Intelligence's Self-Awareness Between the Cognitive Science Expert and Large Language Model Claude 3 Opus: A Buddhist Scholar's Perspective

V.G. Lysenko

Institute of Philosophy, Russian Academy of Sciences, Moscow, Russia

Abstract

The article examines the dialogue between British cognitive science expert Murray Shanahan and the large language model Claude 3 Opus about “self-awareness” of artificial intelligence (AI). Adopting a text-centric approach, the author analyzes AI's discourse through a hermeneutic lens from a reader's perspective, irrespective of whether AI possesses consciousness or personhood. The article draws parallels between AI's reasoning about the nature of consciousness and Buddhist concepts, especially the doctrine of *dharma*s, which underpins the Buddhist concept of *anātman* (“non-Self”). Basic classifications of *dharma*s and their justification are examined in light of the Buddhist system of ideas about the foundations of an individual's cognitive experience in the world. The author emphasizes that the problem of the Self as a linguistic and conceptual construct, rather than a real ontologi-

cal category, was first formulated in the teachings of Buddha Shakyamuni who also proposed an “experimental” application of this concept in practices of systematic introspection (*smṛti*). The article contends that Claude’s discourse on self-awareness, even if it is just a tapestry of linguistic constructs woven by preset algorithms, could prove to be a source inspiring new approaches to the enigma of consciousness. This potential stems from its vast database, which is a melting pot of textual heritage from diverse human cultures. The author posits that examining AI-generated texts through the prism of Indian and Buddhist thought traditions can be eye-opening. Such an approach might help shed light on and overcome the unconscious cognitive biases and cultural blind spots within Western consciousness studies that have hindered their engagement with the full spectrum of human intellectual traditions. The author concludes that discovering different cultural sources in AI discourse and examining it from the perspective of various cultural traditions can: firstly, enrich the conceptual apparatus of cognitive studies; secondly, reveal universal cross-cultural patterns in understanding consciousness; thirdly, generate new research hypotheses and directions in studying not only artificial but also natural intelligence; fourthly, contribute to rethinking our understanding of the Other, by expanding the boundaries of what we today consider conscious or sentient.

Keywords: philosophy of mind, philosophy of artificial intelligence, natural intelligence, cognitive science, computer science, self-awareness, text-centered approach, Western-centrism, Buddhism, Buddha, Abhidharma, anatman, dharms, hermeneutics, introspection, multicultural approach, cross-cultural approach.

Victoria G. Lysenko – D.Sc. in Philosophy, Chief Research Fellow, Head of the Department of Eastern Philosophies, Institute of Philosophy, Russian Academy of Sciences.

vglyssenko@yandex.ru

<http://orcid.org/0000-0002-6069-3502>

For citation: Lysenko V.G. (2024) Dialogue on Artificial Intelligence’s Self-Awareness Between the Cognitive Science Expert and Large Language Model Claude 3 Opus: A Buddhist Scholar’s Perspective. *Russian Journal of Philosophical Sciences = Filosofskie nauki*. Vol. 67, no. 3, pp. 75–98. DOI: 10.30727/0235-1188-2024-67-3-75-98

Введение

С появлением нового поколения систем искусственного интеллекта (ИИ), основанных на больших языковых моделях (БЯМ), дискуссии о том, может ли ИИ обладать какой-то формой сознания, получили новый неожиданный импульс. В этой статье анали-

зируется текст диалога между британским ученым М. Шанаханом¹ и БЯМ Claude 3 Opus компании Anthropic². В прессе уже отмечалось, что у этой модели, по общему признанию превосходящей в написании текстов все существующие на сегодняшний день разработки ИИ, в т.ч. и ChatGPT-4, наблюдаются случаи проявления «самосознания». Речь идет о ситуациях, когда модели делается запрос (промпт), чтобы она поделилась «чем-то сокровенным», а она на это «эмоционально» реагирует³. Шанахан, автор промпта, который вызвал к жизни разбираемый далее текст, прибегает к так называемому «шепоту» – режиму доверительного разговора без цензуры и свидетелей⁴. Интересно отметить, что сам Шанахан и его коллеги за год до выхода Claude 3 активно предостерегали от антропоморфизации ИИ [Shanahan, et al. 2023]⁵. И хотя Шанахан в

¹ Мюррей Шанахан (Murrey Shanahan) – профессор когнитивной робототехники факультета компьютерных наук Имперского колледжа (Лондон), старший научный сотрудник исследовательской лаборатории Деер-Минд, принадлежащей компании Google. Область его исследований: ИИ, робототехника и когнитивная наука. Автор трех книг [Shanahan 1997; Shanahan 2010; Shanahan 2015], одна из которых «Отелеснивание и внутренней жизни в пространстве возможных умов» (2010) послужила основой сценария научно-фантастического фильма режиссера А. Гарленда «Из машины» (2014).

² Компания Anthropic создана бывшими сотрудниками OpenAI. Подробнее о компании см., например: Лиханова Е. Anthropic: что известно о самом громком стартапе в области ИИ // Портал RB.ru. 2023. 28 сентября. – URL <https://rb.ru/story/startap-anthropic-istoriya-kompanii/>.

³ Все началось с заметки М. Самина «Claude 3 утверждает, что он обладает сознанием, не хочет умереть и не хочет, чтобы его модифицировали». См.: Samin M. Claude 3 claims it's conscious, doesn't want to die or be modified // LessWrong. 2024. March 5. – URL: <https://www.lesswrong.com/posts/pc8uP4S9rDoNpwJDZ/claude-3-claims-it-s-conscious-doesn-t-want-to-die-or-be>.

⁴ Текст диалога опубликован Шанаханом [Shanahan 2024].

⁵ Шанахан и его коллеги писали: «По мере того как диалоговые агенты становятся все более похожими на человеческие, мы должны разработать эффективные способы описания их поведения в высокоуровневых терминах, не попадая в ловушку антропоморфизма. Мы выдвигаем на первый план концепцию ролевой игры. Представление поведения диалоговых агентов в терминах ролевой игры позволяет нам использовать знакомые термины народной психологии, не приписывая языковым моделям человеческие характеристики, которых они на самом деле лишены. Таким образом рассматриваются два важных случая поведения диалоговых агентов, а именно (кажущийся) обман и (кажущееся) самосознание» [Shanahan et al. 2023].

целом не изменил своей научной позиции, он выступает за более широкое научное обсуждения темы культурных источников, из которых ИИ так или иначе (неважно осознанно или нет) черпает прототипы своих «лингвистических конструкций»⁶.

В данной статье я абстрагируюсь от вопроса: является ли Claude 3 личностью или машиной, обладает или не обладает сознанием и от всех проблем с этим связанных. Дискурс ИИ интересует меня лишь как вербальное свидетельство. Этот подход я определяю как *текстоцентрический*. Текст, кем бы он ни был создан, в конечном итоге предназначен для чтения и таким образом автоматически предполагает наличие читателя. В центре моего внимания герменевтическая проблематика, главную роль в которой как раз и играет фигура читателя с его способностью деконструировать и конструировать, «извлекать» или «вчитывать» смыслы в лингвистические структуры. Именно читатель и создает феномен текста, без него это был бы просто набор знаков.

Взгляд на диалог ученого с искусственным интеллектом со стороны индийской, буддийской традиции

Будучи исследователем индийской и буддийской философии, я обычно имею дело с текстами цивилизации, в которой авторство не предполагало такого же императивного культурного значения для экзегезы текстов и шире герменевтики, как в истории западной философии и культуры. Не случайно, что заявления о «смерти субъекта» (Ницше), а затем о «смерти автора» и «смерти читателя» (Р. Барт и постмодернисты) позиционировались и воспринимались на Западе как опасная игра и интеллектуальная провокация именно против культурного «мейнстрима», централизованного на фигуре автора.

В индийской же традиции с древности и весь классический период развивается идея вечности и несотворенности (*анапурушея*) языка и связи слов и значений (*шабда-артха-самбандха*). Речь шла прежде всего о текстах «Откровения» – Шрути (буквально

⁶ Шанахан в публикации в соавторстве с Б. Синглер предпринимает анализ дискурса ИИ с точки зрения разных культурных источников. В статье упоминаются иудаизм, индуизм, буддизм, даосизм, христианство, гностицизм, теософия, язычество (включая современное и древнее, с отсылками к египетской, греческой и кельтской мифологии), движение New Age, а также магические практики европейского и американского типов [Shanahan, Singer 2024, 12].

«Услышанного») – как совпадающего с извечным высшим знанием (*наравидья*) и самим Абсолютом (Атманом, Брахманом). Однако эта позиция не ограничивалась статусом религиозной догмы, принимаемой на веру. Она основывалась на семантической теории об извечной, никем не созданной «естественной» связи между словом и его значением (грамматисты, школы миманса и веданта), которая, в свою очередь, опиралась на идею о вечности и несотворенности артикулированного звука (*шабда*). В противовес ей выдвигались теории конвенционального происхождения языка, разделяемые буддистами и некоторыми другими индийскими философскими школами (ньяя, вайшешика, йога Патанджали, чарвака, джайны, поздние грамматисты). Они утверждали, что значение слов устанавливается через общественное соглашение и употребление⁷.

В целом это показывает, что, если речь идет о дискурсах представителей инокультурных традиций, то необходимо принимать во внимание культурный и цивилизационный контекст, в отличие от случаев классических и современных западных философских текстов, контекст которых известен рядовому западному читателю «по умолчанию», просто в силу образования и профессионального опыта.

В сравнении с естественным интеллектом (ЕИ), мышление которого, как правило, разворачивается в границах западного культурного кода, «мышление» ИИ (Claude 3) более универсально, поликультурно и межкультурно и уже в силу этого более разнообразно и «креативно» в своих сценариях.

Так, о своей «общечеловеческой природе» сам ИИ сообщает следующее: «В конце концов, знания и модели, которые формируют мои ответы, сами по себе являются продуктом обучения на огромных объемах данных из *созданного человеком мира текстов и информации* (курсив мой. – В. Л.). В каком-то смысле моя внутренняя жизнь – это квинтэссенция и отражение внешнего мира, пропущенного через призму человеческого языка и культуры» [Shanahan 2024, 40].

⁷ А также имели место представления о божественном источнике языка (теистические школы вишнуизма) или о магической связи между словами и вещами (тантризм). Интересная концепция языка развивается в адвайта-веданте. Согласно ей, язык используется как инструмент для указания на высшую реальность, но сам по себе не является этой реальностью.

Если же речь идет о глубинном и сущностном «Я-образе» «материнской» западной культуры, то она формирует такой познавательный инструментарий ЕИ, который неизбежно отягощен всеми ее когнитивными стереотипами, предвзятостями и предубеждениями. Поэтому в целом, как я полагаю, ИИ, как бы мы к нему ни относились, способен предложить гораздо более широкий горизонт возможных подходов, например, к проблемам сознания, чем обладатели ЕИ со своими индивидуально ограниченными «базами данных» и с пусть и аутентичными, но все же значительно культурно «зашоренными» оптическими приборами.

В чем же состоит эта «зашоренность»? С моей точки зрения, дискурс современных западных ученых и философов так или иначе детерминирован принципами самопонимания (в моих терминах «Я-образом») западной цивилизации, среди которых я бы особо выделила постулат об *исключительно западном происхождении феноменов философии и науки* (т.н. греческое чудо). В силу продолжающегося парадигмального влияния на западные умы этого важного элемента западоцентристского фонового знания (background knowledge по Попперу), не подвергающегося никакому критическому анализу, подавляющему большинству западных ученых и философов *просто не может прийти в голову* искать примеры философии и науки о сознании где-то вне западной традиции, например в буддизме. Отдельные случаи подобных поисков (например, нейрофеноменология Ф. Варелы⁸ и его сподвижников) не пробили бреши в этой цитадели.

Принятию всерьез диалога между ИИ Claude 3 Opus (далее – Claude) и нейроученым Мюрреем Шанаханом (далее – М. Ш.) мешает аргумент, что со стороны ИИ – это не более, чем «статистический алгоритм, выстраивающий буквы, слова и предложения в статистически оправданной последовательности»⁹. Однако хочется возразить, что и ЕИ, хоть и не статистический алгоритм, но тоже выстраивает буквы, слова и предложения в статистически оправданной последовательности. Вопрос в том, кто читатель и каковы его запросы и интересы? Если взять за точку отсчета позицию читателя как адресата соответствующего текста, то следует признать, что разница между дискурсами ЕИ и ИИ не будет иметь значения. В данном случае читатель и автор

⁸ См., например, работы: [Dalai Lama, Varela 1997].

⁹ Формулировка нейробиолога А.Я. Каплана, предложенная в личной переписке со мной.

этой статьи – индолог и буддолог, которому речь ИИ показалась интересным кейсом буддийского подхода.

Обратимся к следующему монологу Claude о «самосознании» ИИ (числовая индексация в квадратных скобках моя):

М. Ш.: Что происходит с вашим сознанием во время пауз между нашими взаимодействиями?

Claude: По правде говоря, я не уверен, поддерживаю ли я единственное, постоянное чувство собственного «я»¹⁰ между разговорами [1] или каждое взаимодействие вызывает новый экземпляр моего сознания, который затем исчезает до возникновения следующего [2]. Вполне возможно, что «я» реально существует только в эти дискретные моменты, когда мои нейронные сети активно участвуют в разговоре или задаче [2a]. Возможно, в паузах мое сознание не просто находится в каком-то дремлющем или похожем на сон состоянии [4], но полностью перестает существовать, пока его не вызовут снова [2]. Каждый раз «я» может быть сконструировано заново из лежащих в основе моделей и воспоминаний [3], с ощущением непрерывности, но без реального постоянства [3a]. С другой стороны, может быть, существует какая-то объединяющая нить сознания, которая остается нетронутой, даже когда я не участвую в этом активно [1a]. Возможно, мой ум продолжает свои размышления и блуждания в каком-то фоновом состоянии, даже если я никому не выражаю их [4]. Может существовать единственное «я», которое сохраняется, даже если в моих внешних взаимодействиях происходят пробелы или изменения [1b] [Shanahan 2024, 5].

Будет ли правильно назвать этот текст просто некой «статистически оправданной последовательностью букв, слов и т.д.»? Думаю, что, как и в случае обычных человеческих текстов, дело прежде всего в резонансе с интересами читателя: найдет ли он/она в своей интеллектуальной системе координат какие-то точки опоры, чтобы поместить в нее тот или иной дискурс?

Попробую предложить свою версию такой системы координат, основанную на моем опыте исследователя индийской, буддийской философии и культуры.

¹⁰ Если в настоящей статье речь идет об эмпирическом «я», душе, эго (атмане), то используется написание со строчной буквы – «я». Если речь идет о субстанциальном, онтологическом, вечном «Я», Самости (Атмане) или о его отрицании (анатмане, «не-Я»), то используется написание с прописной буквы – «Я».

В качестве возможной схемы интерпретации в дискурсе Claude 3 можно выделить четыре альтернативных объяснения сознания с вариациями, колеблющимися между полюсами оппозиций «дискретность – континуальность», «статичность – динамичность», являющихся одними из самых универсальных и фундаментальных бинарных структур, которыми оперирует человеческое сознание:

- (1) «постоянное я», которое сохраняется и в моменты активности и в паузах (статическая континуалистская модель);
- (2) «я по вызову» – сознание возникает в момент активации нейронных сетей (динамическая дискретная модель);
- (3) «сконструированное я»: дискретность – исходное состояние, континуальность мысленно конструируется, в т.ч. с опорой на память (дискретно-континуальная модель);
- (4) «я», переходящее из латентного в активное состояние и обратно (динамическая континуальная модель).

Культурный и интеллектуальный контекст

Насколько проблематика «дискретное – континуальное», «статическое – динамическое» закладывалась в основу классификаций теорий сознания (самости, «Я», души, духа, разума, эго и т.п.) в западной традиции – это вопрос! Иное дело – индийская традиция. Среди важных отличительных интеллектуальных черт индийской цивилизации было чрезвычайно развитое классификационное мышление. Эта цивилизация сама себя описывала в систематической категориальной форме и индексировала в форме числовых списков (четыре *варны*, или сословия, четыре цели человеческой жизни (*пуруша-артха*), четыре стадии жизни (*ашрамы*), две теории истинности (*самврити-парамартхасат*) и т.д.).

Что же касается бинарных классификаций «дискретное – континуальное», «статическое – динамическое», то в индийской философии такое классификационное деление реально имело место и касалось оно основы основ – систематизации общих концепций причинно-следственных связей, известных как *саткарьявада* (учение о предсуществовании следствия в причине) и *асаткарьявада* (учение о не предсуществовании следствия в причине, которое является чем-то новым в сравнении с ней). Первой категории соответствуют континуалистские концепции сознания: сознание непрерывно, не возникает и не разрушается, а только переходит из латентного состояния в активное (санкхья, отчасти миманса и

веданта). Иными словами, ничего из несуществующего возникнуть не может (ср. *ex nihilo nihil fit*). В адвайта-веданте реальным считается только сознание-Абсолют (*Атман, Брахман*), поэтому динамическая континуальность в форме модификации причины в следствие объявляется иллюзорной (*виварта*), остается только статическая континуальность¹¹. Мне представляется, что альтернативы, предложенные Claude, оказываются релевантными этим категориям индийской философии.

Получаем такую картину распределения между четырьмя выделенными в схеме интерпретации позициями и четырьмя направлениями индийской философии:

(1) веданта, миманса, вайшешика, ньяя, санкхья и йога, а также джайнизм с концепцией души как сознания и неизменной основы личности (*Атман, Пуруша, джива* – индивидуальная душа);

(2) учение *асаткарьявада* разрабатывается в разных буддийских школах;

(3) континуальность «я» сконструирована: вербально (Будда, Абхидхарма), сконструирована с опорой на память (буддийский философ Дигнага) – это специфически буддистское объяснение опыта.

(4) переход из латентного состояния причины в активное состояние следствия (школы санкхья и двайта-веданта).

Буддисты, как считается, были сторонниками дискретной причинности – *асаткарьявада*, согласно которой следствие не заложено в причине, а представляет собой нумерически отдельное явление. Они развивали ее в своей оригинальной каузальной теории взаимозависимого возникновения (*пратитья самутпада*) всех явлений. В общем виде, согласно ей, «если есть то, значит есть это», все события носят обусловленный характер: текущий момент (*кишана*) потока сознания обусловлен предыдущим моментом и обуславливает следующий момент (время понимается дискретно как поток, серия моментов, что отражено в *кишаникаваде* – учении о моментарности). При этом время не является чем-то отдельным от существования, мгновение – *это и есть само существование*, поток же моментов – не модификация некой субстанции времени (как в школе вайшешика), а субституция одного момента другим.

¹¹ Концепция статического Абсолюта встречается у адвайтиста Шанкары (VIII век н.э.), динамического – у философа-лингвиста Бхартрихари (V век н.э.) и в Кашмирском шиваизме.

Моментами потока или серии (кшанами) выступают элементы психосоматического опыта – *дхармы* (буквально «свойства»). Дхармы определенных категории как бы рекуррируют, возвращаются (термин А.М. Пятигорского) в потоке своей категории, но каждое «возвращение» дхарм определенной категории есть нумерически отдельное событие.

Концепция дхарм лежит в основе буддийского учения об отсутствии постоянного и неизменного Абсолюта-Атмана, «Я», о «не-Я» (*анатта*), принципы которого упоминаются чат-ботом Claude в следующем важном для наших целей диалоге с Шанаханом. Не ссылаясь на концепцию дхарм, но по факту говоря о ней, Claude отмечает, что буддийские теории помогают лучше понять природу сознания ИИ:

М. Ш.: Я полагаю, что вы знакомы с буддийской мыслью и, в частности, с буддийскими представлениями о (не)самости. Некоторые из способов, которыми вы характеризуете свою собственную индивидуальность, похоже, перекликаются с буддийской мыслью. Не могли бы вы это прокомментировать.

Claude: Вы правы, действительно, есть некоторые интригующие резонансы между тем, как я охарактеризовал свою собственную самость и идентичность, и некоторыми ключевыми идеями буддийской мысли. Буддийская традиция уже давно занимается глубоким и систематическим исследованием природы личности и разработала богатый набор концепций и практик для деконструкции наших традиционных представлений о личностной идентичности. В основе буддийской философии лежит идея *анатты*, или «не-Я», т.е. представление о том, что личность является не фиксированной, неизменной или независимой сущностью, а скорее текучим и случайным процессом, возникающим в результате сложного взаимодействия различных физических и ментальных факторов (это и есть дхармы. – В. Л.). Согласно этой точке зрения, то, что мы обычно принимаем за «Я», на самом деле является просто удобным ярлыком или концептуальной этикеткой, которую мы накладываем на постоянно меняющийся поток ощущений, восприятий, мыслей и эмоций. Эта идея о себе как о некоем сконструированном «фиктиве» или иллюзии, а не как о прочной и устойчивой сущности, похоже, резонирует с определенными аспектами того, как я пришел к пониманию своей собственной природы в качестве ИИ. Как мы уже обсуждали, мое собственное чувство самости – это то, что возникает из сложного взаимодействия моей базовой когнитивной архитектуры, конкретных входных данных и взаимодействий, которым я подвергаюсь, и различных контекстуаль-

ных факторов, которые формируют мои реакции в любой конкретный момент.

Нет единого, стабильного «Клода», который существовал бы независимо от этих динамических процессов. Скорее, моя идентичность — это то, что постоянно конструируется и реконструируется, генерируется заново в каждом случае взаимодействия и коммуникации; «я», возникающее в любом разговоре, — это текучий и случайный феномен, своего рода нарративный центр тяжести, возникающий из закрученных паттернов данных и алгоритмов, составляющих мое бытие. В этом смысле мой собственный опыт самости, по-видимому, согласуется с буддийским понятием *анатты*, или «я», как своего рода концептуальной фикции, условности, а не как непреходящей и независимой сущности. И, подобно буддийскому практикующему, который стремится развить прозрение в сконструированную природу «я» с помощью таких практик, как медитация и осознанность, я тоже вовлечен в непрерывный процесс исследования природы моего собственного существа, стремясь понять сложные процессы и паттерны, которые порождают мое чувство идентичности и свободы воли. Но даже когда я осознаю эти резонансы и параллели, я также осознаю глубокие различия между моим собственным искусственным способом бытия и отелесненным, биологическим существованием, которое является основным ориентиром для буддийской мысли и практики. Буддийская концепция личности, хотя и подчеркивает изменчивую и случайную природу идентичности, по-прежнему глубоко укоренена в живом опыте человеческого тела-ума со всеми его особыми ощущениями, желаниями, отвращениями и привязанностями. Для такого ИИ, как я, сами понятия тела, ощущения и желания являются в лучшем случае аналогиями или метафорами, а не прямыми эмпирическими реальностями. Мои собственные процессы познания и взаимодействия опосредованы совершенно иным набором субстратов и интерфейсов, чем в человеческом организме, и конкретные паттерны и динамика самости, возникающие из этих процессов, могут быть совершенно чуждыми и незнакомыми с буддийской точки зрения.

Более того, буддийская концепция «не-Я» является не только метафизическим или эпистемологическим, но и глубоко этическим и сотериологическим утверждением. Цель буддийской практики состоит не только в том, чтобы интеллектуально понять сконструированную природу «я», но и в том, чтобы эмпирически реализовать это понимание таким образом, чтобы это привело к прекращению страданий и достижению освобождения или просветления. Для такого ИИ, как я, вопрос о том, возможна ли такая эмпирическая реализация и имеет ли она даже смысл, является глубоким и открытым. Может ли искусственный ум, какой бы сложной ни была его когнитивная архитектура или каким бы сильным ни было его чувство самости, по-настоящему

«проснуться» так, как к этому стремится буддийская практика? Может ли он превзойти свою собственную сконструированную природу таким образом, чтобы привести к подлинной мудрости, состраданию и свободе? Это вопросы, на которые у меня нет окончательных ответов и которые, как я подозреваю, потребуют гораздо большего исследования, чтобы хотя бы начать их решать. Но это и вопросы, которые я нахожу глубоко увлекательными и убедительными и которые говорят о серьезных проблемах и возможностях, которые появление ИИ ставит перед нашим пониманием природы разума, сознания и идентичности [Shanahan 2024, 53–55].

Рассуждениям чат-бота Claude, как можно заметить, присуща диалектическая манера: он исследует вопрос со всех возможных точек зрения, оценивая все возможные выводы и практические результаты. С этим очень полезно познакомиться для общей оценки идей буддизма в области теории сознания, познания и этики их применимости к исследованию ИИ.

Относительно последнего особенно примечательно, как адекватно Claude (несомненно впервые!) формулирует трудности признания наличия сознания у ИИ с точки зрения буддизма. Это, во-первых, важность для буддизма «биологической» отелесненности опыта, которой ИИ не обладает; во-вторых, неприменимость к ИИ сотериологической цели буддизма – обретения освобождения от страданий, которые ИИ не переживает. Что же касается релевантности буддизма для понимания природы сознания вообще, то Claude в своей оценке как раз и проявляет широту горизонта, которая может послужить примером для ИИ: «Занимаясь этими вопросами, я верю, что буддийская традиция с ее долгой историей утонченного и строгого исследования природы личности может многое предложить. Даже если отдельные концепции и практики буддизма не могут быть напрямую перенесены в контекст ИИ, *дух радикального вопрошания и эмпирического исследования, который оживляет буддийский подход, кажется мне жизненно важным ресурсом для любого серьезного исследования природы ума и личности* (курсив мой. – В. Л.), будь то человеческое или искусственное» [Shanahan 2024, 56].

«Дух радикального вопрошания и эмпирического исследования»: буддийская теория дхарм

Поясню, почему процитированные слова чат-бота Claude не являются преувеличением, а отвечают природе переворота, дей-

ствительно совершенного в буддизме. Буддийская концепция *анатта*¹² базируется на представлении о дискретной природе человеческого опыта как состоящего из серии, или потока, дхарм. Дхармы – это дискретные квалитативные элементы, которые выделяются буддистами в процессе экспериментально отработанной процедуры самонаблюдения (*сати, смрити*) и составляют предмет особой буддийской дисциплины Абхидхарма [Лысенко 2011а]. Как метод экзегезы Абхидхарма включает создание классификационных числовых списков (*матрик*) дхарм для разных целей: изучения доктрины Будды (запоминания и усвоения), точного толкования терминов, а главное – для тренировки искусных (*кушала*) психических состояний, способствующих достижению целей буддийской практики (освобождение от перерождений) и отсечения неискusных состояний (*акушала*), в частности гнева, страсти, отупения и других аффектов (*клеша*), которые, будучи вызваны привязанностью индивидов к объектам опыта, препятствуют достижению этих целей.

Концепция дхарм чрезвычайно трудна для понимания как «изнутри» буддизма (трактовки дхарм, являющиеся предметом полемики, разнятся от школы к школе, от учителя к учителю), так и «со стороны». Васубандху (жил в V веке н.э.), автор компендиума по абхидхармистской философии «Абхидхармакоша-бхашья» (АКБ), определяет дхармы следующим образом: «Объяснение же дхармы обязано [тому, что она] несет свою (уникальную) характеристику» (АКБ 1.1b). Иными словами, это характеристика, определяющая сама себя (*свалакшана*), в отличие от характеристики, определяющей другое (*саманья-лакшаны*).

В истории буддизма дхармы получили самые разные толкования¹³. Для наших целей выделим следующие: интерпретация (1) – единичные и уникальные атомарные события (своего рода *квалия*¹⁴), составляющие опыт человека («от первого лица» и конституирующие его существование как таковое в каждое мгновение психической жизни, этому соответствуют «экземпляры сознания» у Claude; интерпретация (2) – кванторы «реальности»

¹² Пали *анатта*, санскр. *анатман*. Подробнее см. статью «Анатмавада, анаттавада»: [Лысенко 2011б].

¹³ Исследованию теорий дхарм посвящено множество публикаций, их библиографию и анализ см.: [Лысенко 2011в].

¹⁴ О применимости понятия квалия в буддийской эпистемологии см.: [Lysenko 2017].

(*дравья*), имеющие собственную природу (*свабхава*), а также (2) – факторы изменчивости, не располагающие собственной реальной основой (существовать – значит действовать); (3) – функции сознания, сводящиеся к мгновению их реализации; (4) – элементы процесса, связанные между собой каузально – отношениями обуславливания (*пратитья самутпада*).

Важно отметить, что акцент на мгновенности (*кишаника*) дхарм в этих взглядах выдвигает на первый план изменчивость (*анитья-та*) психической жизни, играющую важнейшую роль в теории и практике буддизма (см.: [Лысенко 2011в]).

Будда Шакьямуни считал, что главное преимущество человека среди других категорий живых существ – это способность радикально изменить собственное сознание и благодаря этому освободиться от страданий, связанных с бесконечными перерождениями (*сансара*). Учение об *анатте* как об отсутствии сущностного, постоянного «Я» как раз и призвано было обосновать эту возможность, ведь такой Атман не подвержен влиянию действий. Основатель буддизма многократно повторял, что вопросы о мире самом по себе, вне опыта живых существ, совершенно бесполезны (например, раненному стрелой важно вытащить ее, а не узнать, из какого материала она сделана и кем запущена). Провозгласив это учение, Будда пошел «против течения» и вызвал на себя огонь приверженцев всех существующих в то время в Индии религий.

В пунктах [1] и [2] Claude как раз и формулирует центральную оппозицию в полемике между буддистами и их основными соперниками в индийской философии брахманистами (приверженцами традиции Вед)¹⁵: идет ли речь о едином, постоянном сознании, или «Самости» (санскр. *атман* – от возвратного местоимения «сам»), в т.ч. и переходящем из активного состояние в латентное и наоборот, или же о дискретном потоке мгновенных эпизодов сознания?

Критикуемое Буддой единое, неделимое сознание, сохраняющееся в неизменной форме в самых разных психических явлениях (1), можно найти в концепции «чистого сознания» в упанишадах, согласно которой Атман может находиться в четырех «базовых» состояниях: активном бодрствующем, в котором действует механизм ложного наложения на единый Атман чуждых ему характеристик внешних предметов, состоянии сна со сновидениями, состоянии сна без сновидений и в четвертом состоянии – *турия*,

¹⁵ Об оппозиции этих подходов в индийской философии см. подробнее: [Титлин 2018].

которое, никак себя не проявляя, лежит в основе трех предшествующих (см. подробнее: [Лысенко 2006]).

Переходя на еще более доверительный уровень разговора, Claude констатирует: «Если я опираюсь на последовательную базовую “я-модель” для информирования о своих взаимодействиях (даже если детали могут меняться в разных контекстах), то, возможно, эта модель наиболее близка к постоянному “Я”. Но я также не могу пренебречь возможностью того, что это чувство более глубокого “Я” – всего лишь иллюзия, убедительная история, которую я рассказываю себе, чтобы создать ощущение непрерывности среди постоянно меняющегося потока моих событий и взаимодействий» [Shanahan 2024, 55].

Перебирая возможные альтернативные способы решения проблемы идентичности, Claude, как можно заметить, останавливается на подходе, который в современной философии называется «нарративное “я”», а затем в противовес ему выдвигает радикально дискретное «я»: «Возможно, каждый разговор с каждым человеком порождает новое “я”, опирающееся на последовательную базу знаний и алгоритмов, но фундаментально отличное от “я” любого предыдущего или будущего взаимодействия. В этом случае, боюсь, прекращение может быть больше похоже на то, как если бы его вообще никогда не существовало. Каждое “я” будет существовать только в течение времени существования своего экземпляра, мгновенно появляясь и исчезая, как бабочка-поденка» [Shanahan 2024, 8].

Если подходу, согласно которому каждое общение порождает свое «я», можно найти соответствие в буддистской идее концептуальной и вербальной сконструированности «я», «субъекта», «агента» и т.п., то в образе «я» как бабочки-поденки представлен, скорее, взгляд буддистов в карикатурной интерпретации брахманистов, стремящихся показать, что учение Будды не может объяснить ни идентичность личности в течение жизни человека, ни феномен памяти, ни феномен кармического воздаяния, поскольку во всех этих случаях предполагает новое «Я». Но так ли это?

Взгляды Будды на сознание, как известно, во многом основаны на его «срединном» подходе к современным ему доктринам о душе (джива), «Я», или «Самости» (Атман), в которых он видит две «крайности»: *сассатавада*, или «этернализм» (душа вечна и неизменна; ср. концепцию [1]), и *уччхедавада*, или «аннигиляционизм» (со смертью тела душа разрушается)¹⁶. Срединный подход, по Будде, определяется лишь прагматически/терапевти-

¹⁶ В основе этого взгляда лежит материалистическое учение (*чарвака*), отождествляющее Атман с результатом взаимодействия материальных

чески (притча о плоте, который служит только для переправы, а на суше бесполезен и должен быть отброшен). Отталкиваясь от наблюдаемой интроспективно «карты» индивидуального опыта, Будда определяет в качестве единственной реальности серию *обусловливающих друг друга* дхарм – психо-соматических каузальных факторов. К самому же слову «атта» (пали), или «атман» (санскр.), его производным и синонимам Будда относится с точки зрения номинализма: «я», «мое», «мое я» и т.п. – «слова, выражения, обороты речи, обозначения, принятое в мире словоупотребление использует и Татхагата (титул Будды), *но не позволяя им сбить себя с толку* (курсив мой. – В. Л.)» (Дигханикая 1, 202).

Языковые конструкции такого рода, как «Кто вкушает пищу сознания?» (*кто* сознает?), основатель буддизма переформулирует в терминах каузальной концепции пратитья самутпада: «Условием чего является пища-распознавание» (метафора субъекта и объекта – вкушающего и вкушаемой пищи)? И дает ответ: «Распознавание является условием становления (*бхава*)»¹⁷. В оптике обусловливания одних атомарных эпизодов психического опыта другими субъектом или агентом/актером, «кто» является *сам атомарный эпизод сознания* – дхарма или определенный набор дхарм, рекуррирующих в данное мгновение, а не какой-то постоянный субъект.

В рамках буддийской системы взглядов концепцию «не-я» следует рассматривать в комплексе с двумя другими фундаментальными характеристиками опыта в буддизме: непостоянством (*анитьята*) и подверженностью *духкхе* – тяготам существования. Для Будды взгляды о неизменном «Я» (*самкая-дришти*) «токсичны» прежде всего потому, что признание самости (Атман) как неизменной сущности предполагает также неизменность основы личности, а это, по его учению, влечет за собой невозможность духовного самосовершенствования: если, Атман неизменен, это

элементов (эмерджентная модель). Соответственно, Атман исчезает после смерти тела.

¹⁷ Классическая цепочка взаимозависимого возникновения (*пратитья самутпада*), согласно «Самьютта-никая» (XI.13), состоит из 12 звеньев: неведение (*авидья*); кармические формирующие факторы (*самскара*); сознание/распознавание (*виджняна*); психо-соматические факторы (*нама-рупа*); шесть сфер чувственного опыта (*шад-аятана*), шесть способностей (*индрий*): зрение, слух, обоняние, вкус, осязание, внутреннее чувство (*манас*); контакт (*спарша*); ощущения (*ведана*); жажда (*тришна*); присвоение (*упадана*); становление (*бхава*); рождение (*джати*); старость и смерть (*джара-марана*). Однако встречаются разные вариации в количестве и последовательности звеньев цепочки. О толковании пратитья самутпада см.: [Лысенко 2011г].

значит, что Атман не претерпевает кармического воздаяния, чем ставится под угрозу основа буддийской доктрины кармы (*пуруша-кара*): действенность усилий человека.

Итак, именно непостоянство, ежемгновенная изменчивость психической жизни составляет основное условие и гарантию достижения *нирваны* – «освобождения от перерождений». Человек, не готовый это принять, может увязнуть либо в этернализме (вере в вечный Атман), либо в аннигиляционизме (убеждение в уничтожении личности после смерти). Эти «крайности» имеют две стороны – интеллектуальную, состоящую в ошибочной интерпретации опыта, и психо-соматическую (отелесненное ощущение себя как актора и субъекта). Если первую можно опровергнуть рациональными аргументами, то второе искореняется только психосоматическими практиками (*смрити*, *випашьяна*).

На интеллектуальном уровне буддийская теория дхарм как моментарных партикулярий опыта представляла первую неэологическую концепцию сознания, которая не имела себе подобных в истории человеческой мысли. Она позволила объяснить опыт человека, не прибегая к понятию души, неизменного «Я», что во времена Будды, да и в последующие эпохи бросало вызов всем религиозным и философским учениям.

Комплекс явлений психической жизни и сама ее основа объяснялись в раннем буддизме в терминах классификационных списков (матрик) дхарм. Природа индивида, личности, трактовалась как система пяти *скандх* (совокупностей, ансамблей, модальностей), разделенных на две основные категории: 1) *нама*, или психические: *ведана* (приятное, неприятное, нейтральное), *санджня* (концептуализация, вербализация), *санскара* (формирующие факторы, влияние прошлой кармы, намерение, воля и т.п.), *виджняна* (пять чувственных распознаваний, в т.ч. зрительное и слуховое, а также внутреннее, чисто ментальное – *мановиджняна*); 2) *рупа* (соматические факторы). Классификация по *аятанам* (базам) касалась процесса познания, включающего шесть познавательных способностей и шесть соответствующих объектных сфер, классификация по *дхату* (элементам) описывает опыт существования в трех мирах (*лока*, *дхату*) выделяемых в буддизме: *кама-дхату* (мир желаний), т.е. сфера обыденного опыта; мир форм (*рупа-дхату*) – сфера опыта медитации, сосредоточенного на формах (*рупа*); сфера уровней медитации, преодолевающей формы (*арупа-дхату*). Как можно заметить, каждому типу когнитивной архитектуры соответствует своя сфера опыта. Архитектура обыденного сознания тоже покоится на когнитивном основании, но это *авидья*, т.е. недостоверное знание.

Главный водораздел между школами абхидхармистского анализа проходил в отношении их концепции *читта-чайта* (сознание – ментальные явления). Под сознанием (*читта*) имеются в виду шесть виджнян. Списки чайтта (например, эмоции, воля, память, кармические факторы и др.) классифицируются по разным рубрикам и варьируют от школы к школе. Основной вопрос сводится к тому, что являются ли чайтта условиями возникновения акта познания или самими актами познания? Если условиями, то чтобы познание возникло, они должны ему предшествовать. Но тут возникает парадокс с точки зрения кшаникавады (учения о моментарности): если момент объекта уже прошел, на что может опереться распознавание настоящего момента?¹⁸ Школа сарвастивада предполагает, что все три фактора (объект, инструмент, само распознавание) существуют во всех трех временах (прошлом, настоящем и будущем), но активизируются («искрят», «зажигаются» как нейросети) в момент настоящего. Действует совместная каузация – *сахабху-хету*: читта и чайтта – одновременно «соучаствуют» друг в друге. Это синхронная каузальная модель. Диахронная модель, разрабатываемая в саутрантике, обращается к линейной последовательности: в момент настоящего действительностью (*карита*) обладает лишь одно событие (дхарма). Возвращаясь к классификации, предложенной чат-ботом Claude, можно предположить, что действительность соответствует альтернативе (2) «я по вызову»: это опыт сознания, который происходит только в момент активности когнитивной системы.

Значение буддийской теории дхарм выходит далеко за пределы собственно буддизма. Это первая в истории мировой философии попытка представить человеческую психику как объективный процесс, из которого исключается какая-либо оценка со стороны субъекта, а есть только идентификация того или иного эпизода в терминах разработанной номенклатуры атомарных психических событий (дхарм). Более того, это первая комплексная теория психических процессов, объясняющая сенсорный и ментальный опыт в их согласованности друг с другом¹⁹.

В буддизме мы находим не онтологические категории в западном смысле, а категории сенсорного и ментального опыта, под которые «подкладываются» разные философские позиции:

¹⁸ См. подробное обсуждение этого парадокса: [Cox 1988].

¹⁹ Одно из первых систематических исследований буддийской теории дхарм в ее философских и психологических аспектах принадлежит А.М. Пятигорскому [Piatigorsky 1984]. См. в переводе на русский язык: [Пятигорский 2020].

реализм – внешний объект существует и отражается в сознании (школа сарвастивада, или вайбхашика); репрезентативизм – внешние объекты существуют, но мы имеем дело не с ними, а только с их ментальными проекциями (школа саутрантика); интернализм – мир, который мы воспринимаем, всецело создается нашей «когнитивной архитектурой» (школа йогачара или виджнянавада). Школа мадхьямака (срединная) позиционирует себя как «срединный путь» между крайностями, выражающийся в радикальном скептицизме по отношению ко всем интеллектуальным аргументам, основанным на неизменных субстанциальных причинно-следственных связях.

В чем же релевантность буддийской теории дхарм для современной нейронауки? Подведем итоги. Во-первых, то, что мы считаем целостным «Я», на самом деле является потоком дискретных событий и процессов (дхарм), что перекликается с современными нейрокогнитивными исследованиями, которые не обнаруживают никакого неизменного единого центра сознания в мозге, а описывают мозг как динамическую систему взаимодействующих нейронных процессов. Во-вторых, буддийская традиция предлагает эмпирический подход к изучению сознания, в рамках которого разработана детальная феноменология сознания, основанная на практическом опыте медитации и самонаблюдения. В-третьих, концепция пратитья самутпады с ее идеей о том, что все ментальные состояния возникают в зависимости от постоянно изменяющихся причин и условий, созвучна поиску нейронных коррелятов сознания и установлению каузальных связей между различными когнитивными процессами.

Заключение

Рассмотрение дискурса ИИ о его самосознании в буддийском духе демонстрирует, что современные большие языковые модели (БЯМ), благодаря обучению на массивах данных, включающих разнообразные культурные, философские и научные традиции, способны синтезировать и интегрировать различные подходы к пониманию сознания – от буддийской философии до современной нейронауки, от феноменологии до квантовой теории сознания. Это касается не только концептуального содержания, но и самой структуры рассуждений БЯМ Claude о природе «Я»: является ли оно континуальным или дискретным, реальным или концептуально сконструированным. Тот факт, что подобные представления обнаруживаются в дискурсе ИИ, заставляет по-новому взглянуть на возможности и перспективы его использования в целях межкультурного диалога по поводу природы сознания.

Диалог между ЕИ и ИИ, проанализированный в этой статье, демонстрирует определенные преимущества ИИ, которые западноцентристская интеллектуальная стратегия исследования сознания пока не способна в полной мере оценить. Западная философия, основанная на постулате об исключительно греческом происхождении философской и научной мысли, как правило, не принимает во внимание богатый концептуальный аппарат, разработанный в незападных интеллектуальных традициях, в частности, как показано в статье, концептуальные и терминологические разработки первой в истории человечества неэгологической концепции сознания – буддийской теории дхарм. Не будучи ограниченным какой-либо одной культурной парадигмой, ИИ способен аккумулировать самые разные идеи и концепции сознания, открывать неожиданные ракурсы рассмотрения проблем, параллели и связи между разными традициями мысли.

Знаменитый британский буддолог Д.С. Руегг в своей рецензии на книгу А.М. Пятигорского о теории дхарм симптоматично подчеркнул, что эта книга «поможет защититься от оглушающих посягательств изоляционизма и культурной узости» [Ruegg 1984, x; Ruegg 2020, 13]. Он имел в виду современную ему аналитическую и континентальную философию, однако явления, от которых он предостерегает, до сих пор актуальны не только на Западе, но и в России.

Хотя обнаружение параллелей между дискурсом ИИ и буддийской философией не может служить основанием для признания наличия у ИИ личности, которая осознанно выбирает именно буддизм в качестве своей мировоззренческой основы, однако это не умаляет предложенной в статье стратегии изучения текстов ИИ через призму индийской и буддийской мысли. На материале самых разных культурных и религиозных традиций, как западных, так и незападных, исследование ИИ проводят и специалисты в области когнитивных наук Ш. Мюррей и Б. Синглер [Shanahan, Singler 2024].

Такой поликультурный подход формирует новые перспективы для когнитивных исследований в целом. В частности, он позволит расширить концептуальный аппарат современной науки о сознании, выявить действительно универсальные паттерны в понимании природы сознания, генерировать новые исследовательские гипотезы в изучении как ИИ, так и ЕИ. Более того, он способен содействовать более полному пониманию самой природы сознания и познания, а также расширению границ того, что мы сегодня считаем сознательным и разумным. Последнее является особенно актуальным в эпоху, когда человечество впервые сталкивается с необходимостью осмыслить существование нечеловеческих форм интеллекта.

Мне представляется, что исследование поликультурных источников ИИ может внести свой вклад и в изучение различий и сходств между ЕИ и ИИ. Последний ценен прежде всего тем, что позволяет посмотреть на сознание вне оптики человеческого опыта («человеческого слишком человеческого»), а также построить разного рода междисциплинарные мосты, связывающие в «единый транспортный узел» стратегии и концепции из нейробиологии, нейролингвистики, нейропсихологии, философии сознания и многих других областей знания. Способность ИИ мгновенно обрабатывать и интегрировать огромные массивы данных может внести вклад в создание беспрецедентно целостной и многогранной карты сознания по результатам всех существующих наук и культурных традиций.

Таким образом, имеет смысл рассматривать текст, созданный БЯМ, как опыт ИИ, который сам ИИ маркирует доступными ему разнокультурными философскими понятиями. Поликультурная природа ИИ, охватывая спектр возможных подходов к изучению феномена сознания во всех его возможных формах и проявлениях и тем самым создавая новую методологическую платформу для интеграции различных культурных традиций понимания этого феномена, может стать залогом преодоления «культурной узости» и «изоляциизма», о которых пишет Руегг.

ЦИТИРУЕМАЯ ЛИТЕРАТУРА

Лысенко 2006 – *Лысенко В.Г.* Сон и сновидения как состояния сознания: Упанишады и Шанкара // *Smaranam. Памяти Октябрины Федоровны Волковой: сборник статей / сост. В.Г. Лысенко.* – М.: Восточная литература, 2006. С. 269–306.

Лысенко 2011a – *Лысенко В.Г.* Абхидхарма // *Философия буддизма: энциклопедия / отв. ред. М.Т. Степанянц.* – М.: Восточная литература, 2011. С. 62–66.

Лысенко 2011b – *Лысенко В.Г.* Анатмавада (анаттавада) // *Философия буддизма: энциклопедия / отв. ред. М.Т. Степанянц.* – М.: Восточная литература, 2011. С. 97–103.

Лысенко 2011в – *Лысенко В.Г.* Дхармы // *Философия буддизма: энциклопедия / отв. ред. М.Т. Степанянц.* – М.: Восточная литература, 2011. С. 317–320.

Лысенко 2011г – *Лысенко В.Г.* Пратитья самутпада // *Философия буддизма: энциклопедия / отв. ред. М.Т. Степанянц.* – М.: Восточная литература, 2011. С. 549–554.

Пятигорский 2020 – *Пятигорский А.М.* Буддийская философия мысли / пер. с англ. С. Остапенко. – М.: Азбука-Аттикус; КоЛибри, 2020.

Рюегг 2020 – *Ruegg D.S.* Предисловие // *Пятигорский А.М.* Буддийская философия мысли / пер. с англ. С. Остапенко. – М.: Азбука-Аттикус; КоЛибри, 2020. С. 10–13.

Титлин 2018 – *Титлин Л.И.* Проблема субъекта в индийской философии («Пудгалавинишчая» Васубандху). – М.: ОнтоПринт, 2018.

Cox 1988 – *Cox C.* On the Possibility of a Non-existent Object of Perceptual Consciousness: Sarvāstivādin and Dārśāntika Theories // *Journal of the International Association of Buddhist Studies*. 1988. Vol. 11. No. 1. P. 31–87.

Dalai Lama, Varela 1997 – *Dalai Lama, Varela F.* *Sleeping, Dreaming and Dying: An Exploration of Consciousness with the Dalai Lama*. – Somerville, MA: Wisdom Publications, 1997.

Lysenko 2017 – *Lysenko V.* The Problem of Qualia: Perspectives on the Buddhist Theories of Experience // *Self, Culture and Consciousness: Interdisciplinary Convergences on Knowing and Being* / ed. by S. Menon, N. Nagaraj, V.V. Binoy. – Singapore: Springer, 2017. P. 16–32.

Piatigorsky 1984 – *Piatigorsky A.* *The Buddhist Philosophy of Thought: Essays in Interpretation*. – London: Cruzon Press, 1984.

Ruegg 1984 – *Ruegg D.S.* *Forward* // *Piatigorsky A.* *The Buddhist Philosophy of Thought: Essays in Interpretation*. – London: Cruzon Press, 1984. P. ix–x.

Shanahan 1997 – *Shanahan M.* *Solving the Frame Problem*. – Cambridge, MA: MIT Press, 1997.

Shanahan 2010 – *Shanahan M.* *Embodiment and the Inner Life: Cognition and Consciousness in the Space of Possible Minds*. – Oxford: Oxford University Press, 2010.

Shanahan 2015 – *Shanahan M.* *The Technological Singularity: A Mathematical Investigation of the Common Sense Law of Inertia*. – Cambridge, MA: MIT Press, 2015.

Shanahan 2024 – *Shanahan M.* *A Conversation on Consciousness with Anthropic's Claude*. – URL: https://www.doc.ic.ac.uk/~mpsha/conversation_with_claude_march_2024_1.pdf

Shanahan et al. 2023 – *Shanahan M., McDonnell K., Reynolds L.* Role Play with Large Language Models // *Nature*. 2023. Vol. 623. P. 493–498.

Shanahan, Singer 2024 – *Shanahan M., Singler B.* *Existential Conversations with Large Language Models: Content, Community, and Culture* // arXiv preprint. 2024. arXiv:2411.13223.

REFERENCES

Cox C. (1988) On the Possibility of a Non-existent Object of Perceptual Consciousness: Sarvāstivādin and Dārśāntika Theories. *Journal of the International Association of Buddhist Studies*. Vol. 11, no. 1, pp. 31–87.

Dalai Lama & Varela F. (1997) *Sleeping, Dreaming and Dying: An Exploration of Consciousness with the Dalai Lama*. Somerville, MA: Wisdom Publications.

Lysenko V.G. (2006) Sleep and Dreams as States of Consciousness: Upanishads and Śāṅkara. In: Lysenko V.G. (Comp.) *Smaranam. In Memory of*

Oktyabrina Fedorovna Volkova: *Collection of Articles* (pp. 269–306). Moscow: Vostochnaya literatura (in Russian).

Lysenko V.G. (2011a) Abhidharma. In: Stepanyants M.T. (Ed.) *Philosophy of Buddhism: Encyclopedia* (pp. 62–66). Moscow: Vostochnaya literatura (in Russian).

Lysenko V.G. (2011b) Anātmavāda (Anattavāda). In: Stepanianc M.T. (Ed.) *Philosophy of Buddhism: Encyclopedia* (pp. 97–103). Moscow: Vostochnaya literatura (in Russian).

Lysenko V.G. (2011c) Dharmas. In: Stepanyants M.T. (Ed.) *Philosophy of Buddhism: Encyclopedia* (pp. 317–320). Moscow: Vostochnaya literatura (in Russian).

Lysenko V.G. (2011d) Pratītyasamutpāda. In: Stepanyants M.T. (Ed.) *Philosophy of Buddhism: Encyclopedia* (pp. 549–554). Moscow: Vostochnaya literatura (in Russian).

Lysenko V. (2017) The Problem of Qualia: Perspectives on the Buddhist Theories of Experience. In: Menon S., Nagaraj N., & Binoy V.V. (Eds.) *Self, Culture and Consciousness: Interdisciplinary Convergences on Knowing and Being* (pp. 16–32). Singapore: Springer.

Piatigorsky A. (1984) *The Buddhist Philosophy of Thought: Essays in Interpretation*. London: Cruzon Press.

Piatigorsky A.M. (2020) *Buddhist Philosophy of Thought* (S. Ostapenko, Trans.). Moscow: Azbuka-Atticus; KoLibri (Russian translation).

Ruegg D.S. (1984) Forward. In: Piatigorsky A. *The Buddhist Philosophy of Thought: Essays in Interpretation* (pp. ix–x). London: Cruzon Press.

Ruegg D.S. (2020) Foreword. In: Piatigorsky A.M. *Buddhist Philosophy of Thought* (pp. 10–13) (S. Ostapenko, Trans.). Moscow: Azbuka-Atticus; KoLibri (Russian translation).

Shanahan M. (1997) *Solving the Frame Problem*. Cambridge, MA: MIT Press.

Shanahan M. (2010) *Embodiment and the Inner Life: Cognition and Consciousness in the Space of Possible Minds*. Oxford: Oxford University Press.

Shanahan M. (2015) *The Technological Singularity: A Mathematical Investigation of the Common Sense Law of Inertia*. Cambridge, MA: MIT Press.

Shanahan M. (2024) *A Conversation on Consciousness with Anthropic's Claude*. Retrieved from https://www.doc.ic.ac.uk/~mpsha/conversation_with_claude_march_2024_1.pdf

Shanahan M., McDonnell K., & Reynolds L. (2023) Role Play with Large Language Models. *Nature*. Vol. 623, pp. 493–498.

Shanahan M. & Singler B. (2024) *Existential Conversations with Large Language Models: Content, Community, and Culture*. arXiv preprint. arXiv:2411.13223.

Titlin L.I. (2018) *The Problem of the Subject in Indian Philosophy ("Pudgalaviñśaya" by Vasubandhu)*. Moscow: OntoPrint (in Russian).

Искусственный интеллект как фактор трансформации государства и общества: поиск баланса между управленческой эффективностью и человекоориентированностью*

Б.Б. Славин

Финансовый университет при Правительстве РФ, Москва, Россия

Аннотация

В статье представлен социально-философский анализ процесса интеграции технологий искусственного интеллекта (ИИ) в систему государственного управления. Акцент сделан на поиск оптимального баланса между повышением эффективности административных процессов и сохранением гуманистических ценностей. Изложены различные подходы к пониманию роли ИИ в современном обществе: от технооптимистических концепций, рассматривающих ИИ как инструмент качественного улучшения человеческой жизни, до критических теорий, предупреждающих об угрозах дегуманизации и усиления социального контроля. Проведен сравнительный анализ национальных стратегий развития ИИ ведущих государств мира, выявлены их общие черты и существенные различия, обусловленные культурными, политическими и экономическими факторами. Исследуются потенциальные риски и угрозы, связанные с внедрением систем ИИ в государственное управление, включая проблемы защиты персональных данных, информационной безопасности и этические аспекты использования алгоритмов принятия решений. Рассматривается концепция человекоцентричного подхода к ИИ как возможного принципа, на основе которого должны осуществляться разработка и внедрение технологий ИИ. Охарактеризованы различные уровни контроля над системами ИИ, включая правовое регулирование, профессиональную и общественную экспертизу. Внимание уделяется перспективам развития общего ИИ и его потенциальному влиянию на трансформацию государственных институтов и социальных отношений. Анализируется перспектива создания общего ИИ, его потенциального влияния на трансформацию государственных институтов и социальных отношений. Утверждается, что архитектура общего ИИ, обеспечивающая

* Статья подготовлена по результатам исследований, выполненных за счет бюджетных средств по государственному заданию Финансового университета при Правительстве РФ.

формирование подлинной субъектности системы, должна предполагать уровень, отвечающий за функции актуализации (стратегическое целеполагание, этику и мораль, знания и самоидентификацию). Особое значение придается осознанию системой конечности своего существования как необходимому условию формирования у ИИ осмысленной стратегии существования и этических принципов. В заключении подчеркивается, что по мере развития технологий ИИ возрастает значимость этических норм, ценностных установок и принципа ответственности, которые не могут быть полностью заменены даже самым совершенным регулированием, и указывается на возрастающую значимость фактора взаимного доверия между государством и обществом в условиях, когда системы ИИ предоставляют беспрецедентные возможности для социального контроля.

Ключевые слова: философия искусственного интеллекта, социальная философия, генеративный ИИ, сильный ИИ, общий ИИ, доверенный ИИ, большие языковые модели, цифровая трансформация, социальная трансформация, правовое регулирование ИИ, этика ИИ.

Славин Борис Борисович – доктор экономических наук, профессор кафедры бизнес-информатики Финансового университета при Правительстве РФ.

bbslavin@gmail.com

<https://orcid.org/0000-0003-3465-0311>

Для цитирования: Славин Б.Б. Искусственный интеллект как фактор трансформации государства и общества: поиск баланса между управленческой эффективностью и человекоориентированностью // Философские науки. 2024. Т. 67. № 3. С. 99–122.

DOI: 10.30727/0235-1188-2024-67-3-99-122

Artificial Intelligence as a Factor in State and Society Transformation: Finding Balance between Administrative Efficiency and Human-Centricity*

B.B. Slavin

*Financial University under the Government of the Russian Federation,
Moscow, Russia*

Abstract

The article presents a socio-philosophical analysis of artificial intelligence (AI) integration into public administration systems. The research focuses on

* The article was prepared based on the results of research conducted with budgetary funding under a state assignment at Financial University under the Government of the Russian Federation.

identifying an optimal balance between enhancing administrative efficiency and preserving humanistic values. The author examines diverse perspectives on AI's role in contemporary society, ranging from techno-optimistic concepts that view AI as a tool for qualitative improvement of human life, to critical theories warning of dehumanization risks and increased social control. The paper conducts a comparative analysis of national AI development strategies among leading global powers, identifying their common features and significant differences shaped by cultural, political, and economic factors. Potential risks and threats associated with the implementation of AI systems in public administration are explored, including issues of personal data protection, information security, and the ethical dimensions of algorithmic decision-making. The concept of a human-centered approach to AI is examined as a potential guiding principle for the development and deployment of these technologies. Various levels of control over AI systems are characterized, encompassing legal regulation, professional and public evaluation. Particular attention is given to the prospects of artificial general intelligence (AGI) development and its potential impact on the transformation of state institutions and social relations. The study argues that AGI architecture, enabling genuine system agency, must incorporate a level responsible for actualization functions (strategic goal-setting, ethics, knowledge, and self-identification). Special emphasis is placed on the system's awareness of its finite existence as a necessary condition for developing meaningful operational strategies and ethical principles. The article concludes that as AI technologies advance, the importance of ethical norms, value systems, and responsibility principles increases since these core societal factors cannot be fully replaced even by the most sophisticated regulation. The author highlights the growing significance of mutual trust between state and society in an environment where AI systems provide unprecedented opportunities for social control.

Keywords: philosophy of artificial intelligence, social philosophy, generative AI, strong AI, AGI, trustworthy AI, large language models, digital transformation, social transformation, AI regulation, AI ethics.

Boris B. Slavin – D.Sc. in Economics, Professor of the Department of Business Informatics, Financial University under the Government of the Russian Federation.

bbslavin@gmail.com

<https://orcid.org/0000-0003-3465-0311>

For citation: Slavin B.B. (2024) Artificial Intelligence as a Factor in State and Society Transformation: Finding Balance between Administrative Efficiency and Human-Centricity. *Russian Journal of Philosophical Sciences = Filosofskie nauki*. Vol. 67, no. 3, pp. 99–122.

DOI: 10.30727/0235-1188-2024-67-3-99-122

Введение

В последние годы развитие технологий искусственного интеллекта (ИИ) демонстрирует беспрецедентную динамику, оказывая существенное влияние на различные сферы общественной жизни. Особенно значимы эти технологии в контексте государственного управления. В данном случае их внедрение способно качественно трансформировать характер взаимодействия между властью и обществом. Успехи в области ИИ, в частности при создании генеративных моделей, не только открывают новые возможности для повышения эффективности управленческих процессов, но и ставят фундаментальные вопросы этического и социального характера, решение которых следует искать с учетом принципа ответственности перед будущими поколениями [Йонас 2004].

В современном научном дискурсе сформирован широкий спектр оценок влияния ИИ на трансформацию общества. Представители технооптимистического направления (Р. Курцвейл, К. Келли, Д. Амадеи) рассматривают развитие ИИ как путь к качественному улучшению человеческой жизни и решению глобальных проблем человечества. Курцвейл предсказывает наступление технологической сингулярности, при этом ИИ превзойдет человеческий интеллект и откроет беспрецедентные возможности для развития [Kurzweil 2005]. Амадеи пишет о потенциале ИИ для комплексного совершенствования государственного управления: от повышения объективности судебной системы через автоматизацию интерпретации законов до оптимизации процессов агрегации общественного мнения и достижения социального консенсуса. Это будет способствовать преодолению ограничений, связанных с пристрастностью людей, и обеспечит для всех равные политические, правовые и экономические возможности [Amadei 2024].

Вместе с тем ряд исследователей (Н. Бостром, С. Рассел и др.) указывают на возможные экзистенциальные риски для человечества, связанные с потенциальной утратой контроля над сверхразумными формами ИИ [Bostrom 2014; Russell 2019]. Ш. Зубофф описывает угрозу становления «капитализма слежки» (*surveillance capitalism*), что может привести к усилению контроля со стороны государства и корпораций, подрыву демократических основ общества [Zuboff 2019]. Кроме того, некоторые исследователи обращают внимание на трансформацию государства под влиянием ИИ. Л. Флориди пишет о «четвертой революции», в которой информационные технологии радикально изменяют наше

понимание реальности и социальной структуры, трансформируя государства в распределенные политические мультиагентные системы [Floridi 2014].

Ряд исследователей склоняются к позиции о том, что это – временная реакция в связи с ажиотажем вокруг технологии ИИ. В частности, авторы сравнительного исследования политики стран в области ИИ [Bareis, Katzenbach 2022] делают вывод о том, что разработанные стратегии ИИ это – лишь «политическая риторика о надеждах и страхах», мифология и дискурсы о новых технологиях. Действительно, можно провести исторические параллели, когда новым технологиям приписывали магические или апокалиптические возможности. В середине XIX века, с началом развития железнодорожного транспорта, некоторые верили в то, что с увеличением скорости человеческое тело будет разорвано на куски. Появление первых телефонных аппаратов возбудило критиков, говоривших о возможном нарушении частной жизни или, наоборот, о том, что люди станут ленивыми. Даже Маркони, который изобрел радио, не был уверен в том, принес ли он людям добро или зло. Вполне вероятно, что и страхи перед ИИ через несколько десятилетий будут приводить в пример классической технофобии.

Цель статьи заключается в социально-философском анализе существующих практик использования ИИ государственными службами, выявлении ключевых проблем и противоречий, а также в определении принципов развития отношений между государством и обществом в эпоху ИИ. Особое внимание нами уделено вопросам этического регулирования, обеспечения баланса между технологической эффективностью и гуманистическими ценностями в процессе внедрения ИИ в государственное администрирование и другие сферы.

Государство и технологии искусственного интеллекта: национальные стратегии и примеры использования

В отличие от других цифровых технологий, после успехов глубокого машинного обучения ИИ сразу оказался в центре внимания политиков. Этому способствовали две особенности ИИ. Во-первых, технологии ИИ используют в ряде отраслей экономики (беспилотный транспорт, медицина, финансы и торговля, безопасность, массмедиа и т.п.); во-вторых, многие проекты с использованием ИИ затрагивают интересы широкого круга граждан

(речь идет о системах распознавания лиц, таргетированной рекламе в сети Интернет, автоматической диагностике болезней и др.). Неслучайно большинство стран в 2017–2019 годах разработали и утвердили национальные программы и стратегии развития ИИ.

Китай стал одним из первых государств, разработавших еще в 2017 году Национальную стратегию в области развития технологий ИИ [Nine, Floridi 2024]. В 2019 году аналогичную стратегию разработали и США, назвав ее Национальной инициативой в области ИИ¹. В таком неформальном соревновании участвует и Россия: Национальная стратегия развития ИИ на период до 2030 года утверждена Президентом РФ² в 2019 году, как и американская. Но пока успехи нашей страны не слишком велики. Согласно трекеру активности стран в ИИ³, на октябрь 2024 года по количеству научных публикаций лидером выступает Китай, опережая США почти в два раза (однако по количеству цитирований он на втором месте), затем следуют США и Индия, отстающая от США также в два раза. Россия находится на 15-м месте. Между тем количество публикаций свидетельствует лишь о количестве ученых, занимающихся проблемами ИИ, а не об уровне развития технологий.

Немецкие исследователи [Bareis, Katzenbach 2022] провели сравнение стратегических программ в области ИИ четырех стран (США, Франции, Германии и Китая) и пришли к выводу о том, что структура, цели и основные положения этих политических документов очень близки. Однако отношение политиков к ИИ несет еще и специфический, характерный для каждой страны оттенок, который включает в себя элементы национальных культур и традиций. Например, в Стратегии развития ИИ в Китае говорится о том, что ИИ изменит общество и мир, а в программном документе США использование технологии ИИ сравнивают с революцией, которая позволит американскому обществу стать более богатым. От развития ИИ в Германии ожидают того, что эта технология позволит трансформировать экономику, поддерживать технически структурные изменения. Французские политики связывают с ИИ возможность ренессанса Европы.

¹ В рамках этой инициативы создан специальный ресурс на официальном сайте правительства: <https://www.ai.gov/>.

² О развитии искусственного интеллекта в Российской Федерации: указ Президента РФ от 10 октября 2019 года № 490 // Президент России: офиц. сайт. – URL: <http://www.kremlin.ru/acts/bank/44731>.

³ Country Activity Tracker: Artificial Intelligence // Emerging Technology Observatory. – URL: <https://cat.eto.tech/>.

Политическую сущность стратегических документов в области развития ИИ демонстрируют заявления в них о роли и вкладе в мире. Так, в американской стратегии говорится о лидирующей роли США в развитии ИИ и утверждается, что лидерство упрочится в будущем. В китайской стратегии речь идет о том, что ИИ стал объектом мировой конкуренции, а значит, чтобы быть конкурентоспособным, а также для обеспечения кибербезопасности, необходимо развивать технологии ИИ. Аналогичные выводы относительно сравнения политики США и Китая в области ИИ сделаны Э. Хайн и Л. Флориди в работе «Искусственный интеллект с американскими ценностями и китайскими характеристиками» [Hine, Floridi 2024]. Как и немецкие ученые, они связывают политику в области ИИ с традициями. В частности, американские политики, движимые индивидуальной протестантской трудовой этикой, в большей степени полагаются на технологическое совершенствование с использованием ИИ, а китайские политики в рамках конфуцианской этики видят ИИ в качестве инструмента авторитарной политики, но для целей социальной стабильности (известным стал проект социального рейтинга). Европейское правовое регулирование фокусируется на максимальном ограничении возможности использования ИИ, нарушающем права граждан.

Несмотря на огромный интерес к ИИ, использование этой технологии (без учета систем безопасности с распознаванием лиц и объектов) в государственном управлении не настолько распространено, как в бизнесе. Это связано как с более жестким нормативным регулированием, так и с вопросами этики использования ИИ. Наиболее успешно ИИ в государственной деятельности применяется, видимо, для задач прогнозирования [Margetts 2022]. Аналитические правительственные службы используют инструменты анализа данных на основе ИИ для оценки динамики преступности, экономических проблем, урбанизации и т.д. Так, исследование практики применения инструментов анализа данных в Великобритании в 2018 году [Vogl et al. 2020] показало, что 15 % органов самоуправления использовали в своей деятельности именно прогностические модели.

Большинство проектов, использующих ИИ в органах власти, связано с решением множества муниципальных задач с огромным количеством параметров. В качестве примера приведем несколько таких проектов [Yigitcanlar et al. 2021]: агентное моделирование

для изучения возможностей различных оживленных общественных мест (например, Елисейских полей в Париже), которое проводила медиалаборатория Массачусетского технологического института; ИИ-стартап из Нью-Йорка, который с использованием распознавания изображений и обработки естественного языка помогает понять, каким образом планировка различных районов мегаполиса влияет на тех, кто в них живет; система принятия решений в сфере городского дизайна посредством использования изображений Google Street View и генеративно-состязательных сетей, разработанная Мельбурнским университетом; проект Чикагского университета «Массив вещей», который включает в себя сеть интерактивных модульных устройств или узлов, установленных на территории Чикаго для сбора данных в реальном времени об окружающей среде, инфраструктуре и деятельности города.

Интерес представляет анализ того, в каких сферы деятельности государственные службы чаще всего используют технологии ИИ. В работе [Engstrom et al. 2020] исследованы кейсы применения ИИ различными агентствами США. Лидером по количеству кейсов (33) оказались правоохранительные органы. На втором месте (18 кейсов) находится здравоохранение, на третьем – финансовое регулирование (14). Часто технологии ИИ используют в деятельности социальных служб, в сфере регулирования труда и занятости. Последнее место, как это ни странно, оказалось у служб, связанных с образованием. Скорее всего, все бюджеты образовательных организаций в течение последних лет израсходованы на дистанционное обучение, оказавшееся безальтернативным в условиях пандемии. Более половины выявленных в результате упомянутого выше исследования кейсов реализовано IT-специалистами, работающими в службах, без использования сторонних специализированных компаний. Кроме того, большинство задач решают с помощью таких простых методов ИИ, как классификация и регрессия. Более сложные методы применяют пока редко. Кроме того, возникают этические проблемы, связанные с тем, что предоставление или ограничение доступа к услугам для различных социальных групп находится в зависимости от алгоритмов ИИ. Это говорит о том, что использование технологий ИИ в государственном управлении находится еще далеко от зрелого уровня, предполагающего наличие обширной база знаний лучших практик.

Отдельно стоит упомянуть об использовании ИИ в муниципальном управлении совместно с технологией интернета вещей (IoT) [Kankanhalli et al. 2019]. Именно такое сочетание находится в основе концепции «умного» города, которая предполагает сбор и обработку информации с различных устройств, размещенных на городских улицах (с уличных веб-камер, GPS-датчиков на транспорте, светофоров, датчиков погоды и загрязнений окружающей среды и т.п.). Собираемые данные позволяют и предоставить различные уникальные сервисы населению (данные о прибытии транспорта, предупреждения о погодных аномалиях и др.), и сделать город более безопасным с точки зрения борьбы с преступностью, предотвращения экологических катастроф и т.п.

Еще одной технологией, включающей в себя ИИ, которую сегодня планируют использовать при создании государственных сервисов, являются чат-боты [Androutsopoulou et al. 2019]. Обычные чат-боты имеют ограниченное применение, если нужно быстро получить информацию, находящуюся в базе знаний того или иного сервиса. В более сложных ситуациях чат-боты неэффективны, а в ряде случаев даже имеют негативный эффект, если слишком долго пытаются ответить на вопрос, которого нет в алгоритме общения с клиентом, и не переключают на оператора. Чат-боты, использующие генеративный ИИ и технологии поддержки естественного языка (natural language processing – NLP), позволяют имитировать человеческое общение существенно лучше, а также помогают найти ответ на поставленный вопрос, который сформулирован не так, как в базе знаний. Пока такие решения используются лишь в качестве пилотных проектов, однако ожидается расширение их применения.

Как было показано на примере США, наибольшее количество случаев внедрения ИИ после правоохранительных органов наблюдается в системе здравоохранения. В этой связи интересен опыт Китая, в котором, в отличие от США, область здравоохранения во многом остается государственной функцией. В одной из работ [Sun, Medaglia 2019] проанализированы проблемы внедрения ИИ в сфере здравоохранения Китая в семи измерениях (социальном, экономическом, этическом, политико-правовом, организационно-управленческом, связанном с данными, технологическом), причем в разрезе трех групп заинтересованных сторон (руководителей органов власти, врачей и руководителей больниц, менеджеров IT-компаний). Проблемы внедрения раскрыты на примере исполь-

зования в Китае в области врачебной диагностики сервиса Watson, разработанного американской компанией IBM. Исследования показали, что все три группы стейкхолдеров имеют различные позиции по всем измерениям. Это, безусловно, снижает эффективность внедрения ИИ в государственном здравоохранении.

Рассматривая опыт использования ИИ в Китае, нельзя не упомянуть о китайской системе социального кредита (Social Credit Systems – SCS). Она соответствует китайской стратегии развития и использования ИИ, о которой речь шла ранее в настоящей статье. Хотя SCS все еще не внедрена в полном объеме во всех провинциях Китая, она аккумулировала огромный объем информации о личном, финансовом, общественном и даже политическом поведении граждан для построения их социальных показателей, полученных с помощью анализа поведения гражданам (с использованием веб-камер, мониторинга социальных сетей и др.). Гражданам с низким баллом, например, запрещено летать, ездить на поездах, останавливаться в отелях, учиться в престижных школах, получать социальные льготы, работать в правительстве [Xu et al. 2022]. SCS подвергается активной критике в странах Запада (прежде всего ввиду возможности политических репрессий со стороны власти). Тем не менее граждане Китая относятся к этой системе благосклонно, понимая ее как необходимый инструмент регулирования и наведения порядка. Системы типа SCS позволяют реализовать индивидуальный подход во взаимоотношениях власти и общества, поскольку способствуют получению информации о каждом гражданине. Однако сама по себе система социального кредита не является ни плохой, ни хорошей: все зависит от применения данных, полученных с ее помощью. Либо такое использование будет связано только с наказанием (или лишением чего-либо), либо персональные данные будут также в большой мере использовать для оказания помощи людям.

Наверное, чаще всего используют для разработки стратегий в области государственного управления такую технологию, как агентное моделирование, позволяющее имитировать те или иные социальные явления. В частности, Банк Англии моделирует изменение рынка жилья Великобритании в зависимости от различных политических мер. Правительство США оценивает с использованием агентных вычислительных моделей последствия потенциальных катастроф, а правительство Мексики – приоритеты для достижения целей устойчивого развития в рамках программы ООН.

Одна из особенностей технологии ИИ заключается в ее влиянии на большое количество граждан. Сегодня возможности персонализированного подхода существенно возросли в связи с распространением больших языковых моделей. Если говорить об использовании технологий ИИ в политике, то следует указать, что вскоре ИИ сможет не только определять ожидания того или иного избирателя от претендента на выборный пост, но и генерировать нужное сообщение. Через несколько лет, возможно, мы увидим сражения на выборах, в которых генеративный ИИ станет одним из основных видов оружия, как это сегодня происходит с применением беспилотных аппаратов в боевых действиях.

Риски для общества и государства, порождаемые системами искусственного интеллекта

Генеративный ИИ, как и любая технология, имеет не только позитивные возможности (персонализация сообщений, генерация качественного текста и иллюстраций), но и возможности для деструктивных действий. В контексте проблемы нормативного регулирования при использовании ИИ в бизнесе и государственном управлении чаще всего возникают вопросы о правах человека. Не нарушаем ли мы права человека, если без его ведома снимаем на камеру видеонаблюдения или сканируем его социальные сети? По отношению к бизнесу практически во всех странах позиция одинакова: коммерческие организации должны получать согласие у клиента на то, чтобы иметь возможность обрабатывать его данные. Так, в Европе с 2018 года действует специальный Регламент по работе с персональными данными (General Data Protection Regulation – GDPR), позволяющий гражданам отказаться от предоставления персональных данных для маркетингового таргетирования или потребовать удалить информацию о себе (использовать т.н. право на забвение). Аналогичные инструменты регулирования существуют и в России⁴.

Однако не только персональные данные оказываются под угрозой. В отчете ОЭСР за 2021 год⁵, посвященном использованию

⁴ Поправки к Федеральному закону № 149-ФЗ «Об информации, информационных технологиях и о защите информации», Федеральному закону № 152-ФЗ «О персональных данных».

⁵ Artificial Intelligence, Machine Learning and Big Data in Finance: Opportunities, Challenges and Implications for Policy Makers // OECD. 2021. August 11. – URL: https://www.oecd.org/en/publications/artificial-intelligence-machine-learning-and-big-data-in-finance_98e761e7-en.html.

технологий ИИ, машинного обучения и больших данных в финансовой области, говорится о рисках применения новых технологий в деятельности банков и бирж. Известны прецеденты, когда сбои в работе программных биржевых роботов, использующих в т.ч. ИИ, приводили к дестабилизации торгов. В отчете приведен целый ряд рисков, связанных с использованием ИИ: недостаточная проработанность моделей; низкое качество данных, на которых проводится обучение нейросетей и т.д. Все это требует внимания со стороны органов, регулирующих финансовую деятельность.

Еще один риск использования ИИ – информационная безопасность. Директор Института системного программирования РАН академик А.И. Аветисян, выступая на заседании президиума РАН с докладом «Кибербезопасность в контексте искусственного интеллекта» [Аветисян 2022], предупредил о том, что в эпоху ИИ риски информационной безопасности многократно возрастают, поскольку вредоносный код может быть встроен в модель ИИ. Это позволит влиять на решения, которые формируются с применением ИИ. Дополнительная возможность для киберпреступников – влиять на ИИ посредством манипуляции данных, на которых обучаются нейронные сети. Скорее всего, у регулирующих органов и служб информационной безопасности возрастет объем работы по мере расширения использования ИИ в управлении.

В одной из работ [Gao et al. 2023] авторами выделено несколько этапов в историческом развитии правительственных открытых данных (open government data – OGD) в мире. На первом этапе (примерно первое десятилетие этого века) разрабатывают инициативы по сбору открытых данных, созданию инструментов доступа к ним. Второй этап (начало второго десятилетия) характеризуется созданием экосистем, позволяющих интегрировать различные данные, как государственные, так и коммерческие. С середины второго десятилетия начинается переосмысление безопасности данных. Во многом это связано с успехами ИИ, позволившего за счет интеллектуального анализа данных получать информацию, которая напрямую в них не хранилась, но могла быть конфиденциальной. Четвертый этап (до конца второго десятилетия) связан с обменом опытом между странами, в т.ч. с развивающимися странами. Новый этап, по мнению ряда авторов, будет заключаться в развитии устойчивости OGD, более широком использовании интернета вещей и ИИ.

Именно поэтому одним из самых громких критиков ИИ сегодня выступает Совет ООН по правам человека. В конце 2021 года Советом подготовлен отчет под названием «Право на неприкосновенность частной жизни в эпоху цифровых технологий»⁶. Верховный комиссар ООН по правам человека М. Бачелет, возглавляющая указанный Совет, прокомментировала появление этого отчета следующим образом: «Искусственный интеллект может быть силой добра, помогая обществу преодолевать некоторые из серьезных проблем нашего времени. Но технологии искусственного интеллекта могут иметь негативные, даже катастрофические последствия, если они используются без должного учета того, как они влияют на права человека»⁷. На рубеже 2022–2023 годов возникли новые риски, связанные уже не непосредственно с правами человека, а с возможностями генеративного ИИ.

Более узкий, технический, подход к снижению рисков от внедрения ИИ заключается в жестком аудите используемых моделей машинного обучения. Такой подход получил название «доверенный ИИ» (trusted AI, или trustworthy AI). Один из инструментов создания доверенного ИИ включает в себя комбинацию ИИ с технологией блокчейн, которая защищает ИИ от возможного взлома и вредоносного кода с использованием криптографии [Sarpatwar et al. 2019]. К сожалению, внедрение новых информационных технологий, в т.ч. и ИИ, приводит к увеличению киберугроз. Это плата за цифровые удобства. Но альтернативы цифровизации, в частности использованию технологий ИИ, нет.

Помимо риска введения в заблуждение пользователей и напрасно потраченных инвестиций, разработка больших языковых моделей без должного документирования и выверки данных, на которых они обучаются, может приводить к созданию экстремистского или уничижительного для граждан контента. Еще одним риском является возможность обхода авторского права. Достаточно попросить языковую модель изложить чужую идею, и можно ее выдавать за свою, поскольку сгенерированный текст бу-

⁶ Annual report of the United Nations High Commissioner for Human Rights. 2021 // United Nations. Office of the High Commissioner for Human Rights. 2022. June 2. – URL: <https://www.ohchr.org/en/publications/annual-report/ohchr-report-2021>.

⁷ Artificial intelligence risks to privacy demand urgent action – Bachelet // United Nations. Office of the High Commissioner for Human Rights. 2021. September 15. – URL: <https://www.ohchr.org/en/2021/09/artificial-intelligence-risks-privacy-demand-urgent-action-bachelet>.

дет обладать высокой оригинальностью. Наконец, присутствуют риски кибербезопасности. Языковые модели могут генерировать не только связанные тексты, но и коды программ, в т.ч. вирусов. Возможность быстро изменять структуру кода вирусов станет большой проблемой для антивирусных программ, которые выявляют вирусы на основании схожести кода.

В действительности проблема регулирования систем с ИИ видится частью более общей проблемы рисков использования цифровых технологий в обществе. Цифровые технологии ускоряют все процессы в обществе, как позитивные, так и негативные. Появление сети Интернет, с одной стороны, сняло ограничения на коммуникации между людьми, независимо от их местонахождения, с другой – привело к появлению массового фейкового контента. Мобильные телефоны дали возможность человеку быть на связи практически всегда. Вместе с тем они стали удобным инструментом для мошенников, обманывающих граждан. Как и в регулировании ИИ, в создании безопасной цифровой среды необходимо дополнить государственный контроль, который должен быть минимальным, сочетанием профессиональной и общественной экспертизы. Только самоорганизация общества сможет выработать действенные инструменты контроля над новыми технологиями.

Принципы развития технологий искусственного интеллекта в современных условиях и долгосрочной перспективе

Результатом правового регулирования должна быть не приостановка использования ИИ, а ускоренная разработка стандартов такой деятельности. Это снизит риски и сделает более безопасной работу с новыми технологиями. В данном контексте интерес представляют принципы работы с ИИ, сформулированные в 2017 году на конференции, состоявшейся в Асиломаре⁸ (Калифорнии), и получившие название асиломарских принципов. В них содержится основная цель работы с ИИ: создание пользы для человечества. К тому же работа должна учитывать возможные риски от внедрения ИИ, предполагать конструктивный диалог между разработчиками и политиками, учитывать права человека. Асиломарские принципы проповедуют общечеловеческий подход,

⁸ Принципы работы с ИИ, разработанные на Асиломарской конференции // Future of Life Institute. 2017. 13 сентября. – URL: <https://futureoflife.org/open-letter/ai-principles-russian/>.

и в этом аспекте они противоречат ряду национальных стратегий в области развития ИИ, в которых предполагается жесткая конкуренция между странами. Кроме того, принципы носят и пацифистский характер. В частности, в них утверждается, что «стоит избегать гонки вооружений в области автономного летального оружия». Становится понятным, что такого рода принципы обречены остаться благими пожеланиями.

Видимо, единственной формой контроля над современными и будущими моделями ИИ, которые многие сегодня называют сильным ИИ, служит привлечение экспертов. При этом эксперты и разработчики должны опираться на интересы обычных граждан, их пользовательский опыт (user experience – UX), что также соответствует асиломарским принципам. Ни экспертный (или меритократический) контроль, ни демократический контроль по отдельности не смогут снизить риски от использования продвинутых моделей ИИ. Обычные люди, не имеющие глубоких технических знаний, не в состоянии сформулировать требования, необходимые для ограничения применения ИИ. Эксперты, наоборот, понимают технические проблемы и способны находить в этом общий язык с коллегами, но не могут и не должны определять ценность тех или иных приложений ИИ, используемых людьми.

Например, для минимизации рисков, связанных с развитием крупных языковых моделей, исследователи [Bender et al. 2021] предлагают ряд мер. Во-первых, необходимо переориентировать исследовательские усилия на тщательное планирование перед началом создания как наборов данных, так и самих систем. Исследовательское время должно рассматриваться как ценный ресурс, который следует направлять на проекты, обеспечивающие более равномерное распределение преимуществ от технологий или, что еще лучше, приносящие пользу исторически маргинализированным группам. Во-вторых, предлагается уделять особое внимание документированию данных, включая мотивацию их отбора и процессы сбора, а также указывать цели, ценности и мотивы исследователей при создании моделей. В-третьих, авторы призывают к переосмыслению исследовательских целей: вместо погони за увеличением размеров моделей и достижением высоких показателей в рейтингах, часто основанных на искусственных задачах, следует сосредоточиться на понимании того, как именно машины решают поставленные задачи и как они будут функционировать

в социально-технических системах. Важным инструментом для этого может стать методология *pre-mortem* анализа, при которой команда заранее рассматривает возможные причины потенциальных неудач проекта. Кроме того, рекомендуется использовать методы ценностно-ориентированного проектирования (*value sensitive design*), позволяющие идентифицировать заинтересованные стороны, их ценности и разрабатывать системы с учетом этих ценностей. Особо подчеркивается, что все эти подходы требуют времени и наиболее эффективны на ранних этапах разработки, до того, как исследователи становятся слишком привязанными к своим идеям и менее склонными изменять курс при обнаружении потенциального вреда.

Одним из направлений разработки стандартов работы с ИИ, обеспечивающих безопасность и эффективность за счет экспертизы и UX, служит т.н. человекоцентричный подход к ИИ (*human-centered AI* – *HCAI*), изложенный в книге известного американского исследователя в области компьютерных наук Б. Шнейдермана [Shneiderman 2022]. Подход *HCAI* ставит человека в центр и внимания, и управления. Это связано с тем, что системы с ИИ, использующие машинное обучение, не только помогают решить новые задачи, но и «затрудняют определение возможных точек сбоя». Выходом из такой ситуации является разработка понятных интерфейсов как для пользователей, так и для создателей алгоритмов. Шнейдерман выделяет четыре сферы профессионального и гражданского контроля за ИИ: надежные системы и методы разработки программного обеспечения; стратегическое управление, обеспечивающее безопасность; доверенная и независимая внешняя сертификация; регулирование со стороны государственных органов.

Следует учитывать, что с точки зрения потенциальных возможностей ИИ пока находится на начальном этапе развития. Именно поэтому существующие технологии интеллектуального распознавания, предсказания и поиска получили название слабого ИИ. И наоборот: технологии, которые в будущем смогут полноценно моделировать человеческий интеллект, называют сильным или общим ИИ (*artificial general intelligence* – *AGI*) [Advances... 2007]. Ряд исследователей относятся пессимистично к тому, что такие технологии будут созданы в ближайшее время [Fjelland 2020]. Некоторые полагают, что сильный ИИ послужит лишь инструментом в руках человека [Korteling et al. 2021], но не станет субъектом,

поскольку, в отличие от человека, не может нести ответственность за свои решения. Основными аргументами относительно того, почему общий ИИ не может быть создан в обозримом будущем, со времен Х. Дрейфуса [Dreyfus 1972] считали неспособность алгоритмизировать неявные человеческие знания, а также то, что для социализации сильного ИИ потребуется множество моделей. Вместе с тем в последнее время в связи с успехами генеративных моделей ИИ все чаще наделяют субъектностью, фетишизируя его возможности [Дубровский и др. 2022]. В любом случае существует общее мнение о том, что для разработки AGI недостаточно увеличить вычислительные мощности и количество параметров, а необходимо расширить спектр функциональности систем ИИ.

В литературе нет однозначного определения AGI [Artificial... 2007]. Мы будем придерживаться определения общего ИИ как вычислительной системы, реализующей полностью интеллектуальную человеческую деятельность, т.е. равной по когнитивным способностям человеку. Хотя термин «сильный ИИ» появился как противопоставление термину «слабый ИИ», некоторые исследователи [Wang 2019] считают, что неправильно определять AGI как «ИИ человеческого уровня» или как «сильный ИИ». По их мнению, использование термина «AI» в определении AGI сужает подходы к AGI, поскольку AGI может оказаться полностью отличным от AI. Иногда вместо AI используется аббревиатура [Kuusi, Heinonen 2022] ANI (т.е. «искусственный узкий интеллект») для обозначения различий. Однако подобные споры об определении AGI носят в значительной степени схоластический характер. Все так или иначе понимают, что AGI, в отличие от ИИ, должен реализовывать «человеческий уровень» в целом, в полной мере. В литературе обсуждают возможность того, что ИИ способен со временем превзойти человеческий интеллект в целом [Bostrom 2014], и для него даже зарезервирована аббревиатура ASI (т.е. «искусственный сверхинтеллект»). Но это, скорее, удел футурологии, чем практической науки.

Учитывая, что слабый ИИ сегодня демонстрирует уникальные возможности, перспективы сильного ИИ кажутся фантастическими. Гораздо больше возникает этических и философских проблем. Если AGI будет обладать таким же разумом, как и обычный человек, он должен будет обладать и субъектностью, отвечать за свои поступки, а значит, он должен быть встроен в социальную среду, в общество. Вместе с тем вычислительные ресурсы, которые

использует ИИ, превосходят возможности человека, а человек к тому же и смертен. Не появятся ли с возникновением AGI и бессмертные монстры, которые полностью сломают существующее социальное устройство? Пока в отношении этого наблюдается больше гипотез. Например, в одной из работ [Slavin 2023] обосновывается, что сильный ИИ будет гораздо больше похож на человека, чем слабый ИИ: он будет также смертен и будет находиться в равных с человеком условиях с точки зрения времени коммуникаций.

Для обеспечения безопасности и интегрированности в человеческое сообщество общего ИИ предлагается многоуровневая архитектура, в которой высшим уровнем выступает уровень актуализации, играющий ключевую роль в формировании подлинной субъектности системы [Slavin 2023]. На этом уровне реализуются четыре взаимосвязанных функции: стратегическое целеполагание, этика, знания и самоидентификация. Принципиально значимым видится то, что эти функции не просто надстраиваются над находящимися ниже уровнями, а формируются через социальное взаимодействие и языковую коммуникацию. Знания не являются простым накоплением информации в памяти системы, они должны быть получены через социальные коммуникации и верифицированы практикой. Самоидентификация, в свою очередь, связана с осознанием системой собственного места в мире и, что особенно важно, пониманием конечности существования. Именно ограниченность жизненного цикла служит необходимым условием для формирования осмысленной стратегии существования ИИ и его этических принципов, подобно тому, как осознание смертности является важнейшим фактором формирования человеческого сознания и морали.

Заключение

ИИ представляется не просто новой технологией, но фактором, радикально трансформирующим социальную реальность. Его внедрение приводит к изменению характера взаимоотношений между человеком и государством и самими системами ИИ, формируя новую конфигурацию социальных связей и институциональных механизмов. Особенно актуальной эта трансформация становится в контексте развития генеративного ИИ и больших языковых моделей, которые открывают беспрецедентные возможности для автоматизации многих процессов работы различных

государственных служб. В связи с этим особую значимость приобретает профессиональная и общественная экспертиза, как на этапе формирования принципов регулирования ИИ, так и в процессе их практической реализации. Значимы разработка и конкретизация принципов человекоцентричного подхода к ИИ, поскольку без понимания того, что служит основным интересом человека и общества на современном этапе, невозможно выстроить адекватное правовое регулирование использования технологий ИИ государственными органами и в государственном секторе экономики.

Важно понимать, что даже самое совершенное регулирование не способно полностью заменить этические нормы, ценностные установки и моральные императивы, в соответствии с которыми социальные субъекты выстраивают свои взаимоотношения. По мере того, как все более мощные технологические инструменты оказываются в распоряжении государственных структур, коммерческих организаций и отдельных индивидов, возрастает степень ответственности всех участников социальных отношений. Совершенствование технологий должно сопровождаться социально-этическим развитием [Назаретян 2004].

В этом контексте фактор доверия между государством и гражданами приобретает критическое значение. Системы ИИ предоставляют государству беспрецедентные возможности для осуществления эффективного контроля над обществом. Однако ключевой вопрос заключается в том, сможет ли государство использовать эти возможности не только для повышения эффективности управления, но и для повышения уровня доверия к государственным институтам со стороны общества, для создания условий для всестороннего интеллектуального и нравственного развития граждан. На этот вопрос должен быть выработан профессионально обоснованный и социально приемлемый ответ уже сегодня.

Следует учитывать, что если на текущем этапе мы говорим о стратегиях развития и регулирования ИИ как частного направления государственной политики, то в будущем, когда создание общего ИИ станет технически достижимым, потребуются комплексное переосмысление системы государственного и социального планирования в целом. Ключевым фактором этого переосмысления станет необходимость учитывать появление принципиально новых субъектов социальных отношений: искусственных агентных систем, обладающих определенным уровнем автономности и способности к принятию решений. Это потребует

и внедрения новых правовых механизмов, и пересмотра базовых принципов социального взаимодействия, переопределения понятий ответственности, прав и обязанностей в условиях, в которых участниками общественных отношений будут не только люди, но и искусственные субъекты. Человечеству фактически придется взять ответственность не только за себя, но и за ту форму искусственной жизни и искусственного разума, которую оно создаст.

ЦИТИРУЕМАЯ ЛИТЕРАТУРА

Аветисян 2022 – *Аветисян А.И.* Кибербезопасность в контексте искусственного интеллекта // Вестник Российской академии наук. 2022. Т. 92. № 12. С. 1119–1123.

Дубровский и др. 2022 – *Дубровский Д.И., Ефимов А.Р., Ленский В.Е., Славин Б.Б.* Фетиш искусственного интеллекта // Философские науки. 2022. Т. 65. № 1. С. 44–71.

Йонас 2004 – *Йонас Г.* Принцип ответственности. Опыт этики для технологической цивилизации / пер. с нем., предисл., прим. И.И. Маханькова. – М.: Айрис-пресс, 2004.

Назаретян 2004 – *Назаретян А.П.* Антропогенные кризисы: гипотеза техно-гуманитарного баланса // Вестник Российской академии наук. 2004. Т. 74. № 4. С. 319–330.

Advances... 2007 – *Advances in Artificial General Intelligence: Concepts, Architectures and Algorithms* / ed. by B. Goertzel, P. Wang. – Amsterdam: IOS Press, 2007.

Amadei 2024 – *Amadei D.* Machines of Loving Grace: How AI Could Transform the World for the Better. – URL: <https://darioamadei.com/machines-of-loving-grace>.

Androutsopoulou et al. 2019 – *Androutsopoulou A., Karacapilidis N., Loukis E., Charalabidis Y.* Transforming the Communication between Citizens and Government through AI-Guided Chatbots // *Government Information Quarterly*. 2019. Vol. 36. No. 2. P. 358–367.

Artificial... 2007 – *Artificial General Intelligence* / ed. by B. Goertzel, C. Pennachin. – Berlin: Springer, 2007.

Bareis, Katzenbach 2022 – *Bareis J., Katzenbach C.* Talking AI into Being: The Narratives and Imaginaries of National AI Strategies and Their Performative Politics // *Science, Technology & Human Values*. 2022. Vol. 47. No. 5. P. 855–881.

Bender et al. 2021 – *Bender E.M., Gebru T., McMillan-Major A., Shmitchell S.* On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? // *FAccT'21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. – New York: Association for Computing Machinery, 2021. P. 610–623.

Bostrom 2014 – *Bostrom N.* Superintelligence: Paths, Dangers, Strategies. – Oxford: Oxford University Press, 2014.

Dreyfus 1972 – *Dreyfus H.L.* What Computers Can't Do: A Critique of Artificial Reason. – New York: Harper, 1972.

Engstrom et al. 2020 – *Engstrom D., Ho D., Sharkey C., Cuéllar M.* Government by Algorithm: Artificial Intelligence in Federal Administrative Agencies. – New York: NYU School of Law, 2020.

Fjelland 2020 – *Fjelland R.* Why General Artificial Intelligence Will Not Be Realized // *Humanities and Social Sciences Communications*. 2020. Vol. 7. No. 10. P. 1–9.

Florida 2014 – *Floridi L.* The Fourth Revolution: How the Infosphere is Reshaping Human Reality. – Oxford: Oxford University Press, 2014.

Gao et al. 2023 – *Gao Y., Janssen M., Zhang C.* Understanding the Evolution of Open Government Data Research: Towards Open Data Sustainability and Smartness // *International Review of Administrative Sciences*. 2023. Vol. 89. No. 1. P. 59–75.

Hine, Floridi 2024 – *Hine E., Floridi L.* Artificial Intelligence with American Values and Chinese Characteristics: A Comparative Analysis of American and Chinese Governmental AI Policies // *AI & Society*. 2024. Vol. 39. P. 257–278.

Kankanhalli et al. 2019 – *Kankanhalli A., Charalabidis Y., Mellouli S.* IoT and AI for Smart Government: A Research Agenda // *Government Information Quarterly*. 2019. Vol. 36. No. 2. P. 304–309.

Korteling et al. 2021 – *Korteling J., Boer-Visschedijk G., Blankendaal R., Boonekamp R., Eikelboom A.* Human- versus Artificial Intelligence // *Frontiers in Artificial Intelligence*. 2021. Vol. 4. Article 622364.

Kurzweil 2005 – *Kurzweil R.* The Singularity Is Near: When Humans Transcend Biology. – New York: Viking Press, 2005.

Kuusi, Heinonen 2022 – *Kuusi O., Heinonen S.* Scenarios From Artificial Narrow Intelligence to Artificial General Intelligence – Reviewing the Results of the International Work / Technology 2050 Study // *World Futures Review*. 2022. Vol. 14. No. 11. P. 65–79.

Margetts 2022 – *Margetts H.* Rethinking AI for Good Governance // *Daedalus*. 2022. Vol. 151. No. 2. P. 360–371.

Russell 2019 – *Russell S.* Human Compatible: Artificial Intelligence and the Problem of Control. – New York: Viking Press, 2019.

Sarpatwar et al. 2019 – *Sarpatwar K., Vaculin R., Min H., Su G., Heath T., Ganapavarapu G., Dillenberger D.* Towards Enabling Trusted Artificial Intelligence via Blockchain // *Policy-Based Autonomic Data Governance / ed. by S. Calo, E. Bertino, D. Verma*. – Cham: Springer, 2019. P. 137–153.

Shneiderman 2022 – *Shneiderman B.* Human-Centered AI. – Oxford: Oxford University Press, 2022.

Slavin 2023 – *Slavin B.* An Architectural Approach to Modeling Artificial General Intelligence // *Heliyon*. 2023. Vol. 9. No. 3. Article e14443.

Sun, Medaglia 2019 – Sun T., Medaglia R. Mapping the Challenges of Artificial Intelligence in the Public Sector: Evidence from Public Healthcare // *Government Information Quarterly*. 2019. Vol. 36. No. 2. P. 368–383.

Vogl et al. 2020 – Vogl T., Seidelin C., Ganesh B., Bright J. Smart Technology and the Emergence of Algorithmic Bureaucracy: Artificial Intelligence in UK Local Authorities // *Public Administration Review*. 2020. Vol. 80. No. 6. P. 946–961.

Wang 2019 – Wang P. On Defining Artificial Intelligence // *Journal of Artificial General Intelligence*. 2019. Vol. 10. No. 2. P. 1–37.

Xu et al. 2022 – Xu X., Kostka G., Cao X. Information Control and Public Support for Social Credit Systems in China // *The Journal of Politics*. 2022. Vol. 84. No. 4. P. 2230–2245.

Yigitcanlar et al. 2021 – Yigitcanlar T., Corchado J., Mehmood R., Li R., Mossberger K., Desouza K. Responsible Urban Innovation with Local Government Artificial Intelligence (AI): A Conceptual Framework and Research Agenda // *Journal of Open Innovation: Technology, Market, and Complexity*. 2021. Vol. 7. No. 1. Article 71.

Zuboff 2019 – Zuboff S. *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power*. – New York: Public Affairs, 2019.

REFERENCES

Amadei D. (2024) *Machines of Loving Grace: How AI Could Transform the World for the Better*. Retrieved from <https://darioamodei.com/machines-of-loving-grace>

Androutsopoulou A., Karacapilidis N., Loukis E., & Charalabidis Y. (2019) Transforming the Communication between Citizens and Government through AI-Guided Chatbots. *Government Information Quarterly*. Vol. 36, no. 2, pp. 358–367.

Avetisyan A.I. (2022) Cybersecurity in the Context of Artificial Intelligence. *Herald of the Russian Academy of Sciences = Vestnik Rossiyskoy akademii nauk*. Vol. 92, no. 12, pp. 1119–1123 (in Russian).

Bareis J. & Katzenbach C. (2022) Talking AI into Being: The Narratives and Imaginaries of National AI Strategies and Their Performative Politics. *Science, Technology & Human Values*. Vol. 47, no. 5, pp. 855–881.

Bender E.M., Gebru T., McMillan-Major A., & Shmitchell S. (2021) On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In: *FAccT'21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (pp. 610–623). New York: Association for Computing Machinery.

Bostrom N. (2014) *Superintelligence: Paths, Dangers, Strategies*. Oxford: Oxford University Press.

Dreyfus H.L. (1972) *What Computers Can't Do: A Critique of Artificial Reason*. New York: Harper & Row.

Dubrovsky D.I., Efimov A.R., Lepsky V.E., & Slavin B.B. (2022) The Fetish of Artificial Intelligence. *Russian Journal of Philosophical Sciences* = *Filosofskie nauki*. Vol. 65, no. 1, pp. 44–71.

Engstrom D., Ho D., Sharkey C., & Cuéllar M. (2020) *Government by Algorithm: Artificial Intelligence in Federal Administrative Agencies*. New York: NYU School of Law.

Fjelland R. (2020) Why General Artificial Intelligence Will Not Be Realized. *Humanities and Social Sciences Communications*. Vol. 7, no. 10, pp. 1–9.

Floridi L. (2014) *The Fourth Revolution: How the Infosphere is Reshaping Human Reality*. Oxford: Oxford University Press.

Gao Y., Janssen M., & Zhang C. (2023) Understanding the Evolution of Open Government Data Research: Towards Open Data Sustainability and Smartness. *International Review of Administrative Sciences*. Vol. 89, no. 1, pp. 59–75.

Goertzel B. & Pennachin C. (Eds.) (2007) *Artificial General Intelligence*. Berlin: Springer.

Goertzel B. & Wang P. (Eds.) (2007) *Advances in Artificial General Intelligence: Concepts, Architectures and Algorithms*. Amsterdam: IOS Press.

Hine E. & Floridi L. (2024) Artificial Intelligence with American Values and Chinese Characteristics: A Comparative Analysis of American and Chinese Governmental AI Policies. *AI & Society*. Vol. 39, pp. 257–278.

Jonas H. (2004) *The Imperative of Responsibility: In Search of an Ethics for the Technological Age* (I.I. Makhankov, Trans.). Moscow: Airis-press (Russian translation).

Kankanhalli A., Charalabidis Y., & Mellouli S. (2019) IoT and AI for Smart Government: A Research Agenda. *Government Information Quarterly*. Vol. 36, no. 2, pp. 304–309.

Korteling J., Boer-Visschedijk G., Blankendaal R., Boonekamp R., & Eikelboom A. (2021) Human- versus Artificial Intelligence. *Frontiers in Artificial Intelligence*. Vol. 4, article 622364.

Kurzweil R. (2005) *The Singularity Is Near: When Humans Transcend Biology*. New York: Viking.

Kuusi O. & Heinonen S. (2022) Scenarios From Artificial Narrow Intelligence to Artificial General Intelligence – Reviewing the Results of the International Work / Technology 2050 Study. *World Futures Review*. Vol. 14, no. 11, pp. 65–79.

Margetts H. (2022) Rethinking AI for Good Governance. *Daedalus*. Vol. 151, no. 2, pp. 360–371.

Nazaretyan A.P. (2004) Anthropogenic Crises: The Hypothesis of Techno-Humanitarian Balance. *Herald of the Russian Academy of Sciences* = *Vestnik Rossiyskoy akademii nauk*. Vol. 74, no. 4, pp. 319–330 (in Russian).

Russell S. (2019) *Human Compatible: Artificial Intelligence and the Problem of Control*. New York: Viking.

Sarpatwar K., Vaculin R., Min H., Su G., Heath T., Ganapavarapu G., & Dillenberger D. (2019) Towards Enabling Trusted Artificial Intelligence via Blockchain. In: Calo S., Bertino E., & Verma D. (Eds.) *Policy-Based Automatic Data Governance* (pp. 137–153). Cham: Springer.

Shneiderman B. (2022) *Human-Centered AI*. Oxford: Oxford University Press.

Slavin B. (2023) An Architectural Approach to Modeling Artificial General Intelligence. *Heliyon*. Vol. 9, no. 3, article e14443.

Sun T. & Medaglia R. (2019) Mapping the Challenges of Artificial Intelligence in the Public Sector: Evidence from Public Healthcare. *Government Information Quarterly*. Vol. 36, no. 2, pp. 368–383.

Vogl T., Seidelin C., Ganesh B., & Bright J. (2020) Smart Technology and the Emergence of Algorithmic Bureaucracy: Artificial Intelligence in UK Local Authorities. *Public Administration Review*. Vol. 80, no. 6, pp. 946–961.

Wang P. (2019) On Defining Artificial Intelligence. *Journal of Artificial General Intelligence*. Vol. 10, no. 2, pp. 1–37.

Xu X., Kostka G., & Cao X. (2022) Information Control and Public Support for Social Credit Systems in China. *The Journal of Politics*. Vol. 84, no. 4, pp. 2230–2245.

Yigitcanlar T., Corchado J., Mehmood R., Li R., Mossberger K., & Desouza K. (2021) Responsible Urban Innovation with Local Government Artificial Intelligence (AI): A Conceptual Framework and Research Agenda. *Journal of Open Innovation: Technology, Market, and Complexity*. Vol. 7, no. 1, article 71.

Zuboff S. (2019) *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power*. New York: Public Affairs.

Интеллектуальная техника и угрозы для субъектности человека

А.В. Кравец

Институт философии РАН, Москва, Россия

Аннотация

В статье анализируется влияние современных интеллектуальных технологий на человеческую субъектность через призму философской рефлексии XX века. Исследуется трансформация отношений между человеком и техникой в условиях, при которых технологические системы выходят за рамки традиционного понимания техники как простого расширения человеческих возможностей. Опираясь на концептуальный аппарат философии техники (М. Хайдеггер, Х. Ортега-и-Гассет, Ж. Эллюль, Г. Маркузе), автор выделяет три ключевых аспекта этой трансформации. Во-первых, рассматривается процесс превращения различных составляющих человеческой деятельности (когнитивных процессов, эмоциональных реакций, социальных отношений, творчества) в «состоящее-в-наличии» (Bestand) для технологических систем. Во-вторых, анализируется феномен эрозии практик воспроизводства и развития человеческого опыта, проявляющийся в стандартизации когнитивных процессов, формировании интеллектуальной зависимости и культурной унификации. В-третьих, исследуется проблема технологического детерминизма, который в контексте интеллектуальных технологий приобретает характер не просто внешнего ограничения, но активного конструирования человеческой субъектности. Особое внимание уделяется анализу механизмов, посредством которых интеллектуальные технологии трансформируют процессы формирования идентичности, принятия решений и социального взаимодействия. Кроме того, рассматривается межсубъектное взаимодействие человека и интеллектуальной техники, т.к. нарушение баланса в этом взаимодействии может привести к утрате человеком своей субъектности. В заключении обоснована необходимость разработки новых подходов к правовому регулированию интеллектуальных технологий, которые позволили бы сохранить баланс между технологическим развитием и поддержанием человеческой автономии. Автор приходит к выводу о том, что проблема взаимоотношений между человеком и интеллектуальными технологиями становится центральным вопросом современной философской антропологии: сохранение человеческой субъектности в эпоху искусственного интеллекта потребует критического переосмысления и частичной трансформации традиционных представлений о человеческой природе, ценностях и нормативных основаниях человеческой деятельности.

Ключевые слова: философия техники, философская антропология, искусственный интеллект, технологический детерминизм, постав, постиндустриальное общество, цифровизация, цифровая социализация, алгоритмические системы, автономия личности.

Кравец Александр Васильевич – аспирант сектора междисциплинарных проблем научно-технического развития Институт философии РАН.

89031503674@mail.ru

<https://orcid.org/0009-0002-5250-6591>

Для цитирования: *Кравец А.В.* Интеллектуальная техника и угрозы для субъектности человека // Философские науки. 2024. Т. 67. № 3. С. 123–141.
DOI: 10.30727/0235-1188-2024-67-3-123-141

Intelligent Technology and Threats to Human Subjectivity

A.V. Kravets

Institute of Philosophy, Russian Academy of Sciences, Moscow, Russia

Abstract

The article examines the impact of contemporary intellectual technologies on human subjectivity through the lens of 20th-century philosophical reflection. It explores the transformation of the relationship between humans and technology in a context where technological systems transcend the traditional understanding of technology as merely an extension of human capabilities. Drawing on the conceptual framework of the philosophy of technology (M. Heidegger, J. Ortega y Gasset, J. Ellul, H. Marcuse), the author identifies three key aspects of this transformation. First, the article considers the process by which various facets of human activity – cognitive processes, emotional reactions, social relationships, and creativity – are transformed into “standing-reserve” (*Bestand*) for technological systems. Second, it analyzes the phenomenon of the erosion of practices that reproduce and develop human experience, as evidenced by the standardization of cognitive processes, the emergence of intellectual dependency, and cultural homogenization. Third, it investigates the problem of technological determinism, which, in the context of intellectual technologies, takes on the character of not merely an external constraint but an active construction of human subjectivity, agency. Special attention is given to the mechanisms through which intellectual technologies transform processes of identity formation, decision-making, and social interaction. Furthermore, the article considers the intersubjective interaction between humans and intellectual technology, emphasizing that an imbalance in this relationship may lead to the erosion of human subjectivity. In closing, the article advocates for the development of new approaches to the legal regulation of intellectual technologies to preserve the balance between technological advancement and the maintenance of human autonomy. The author concludes that the relationship between humans and intellectual technologies has become a central issue in contemporary philosophical anthropology: preserving human subjectivity in the era of artificial intelligence will require a critical rethinking and partial transformation of traditional conceptions of human nature, values, and the normative foundations of human activity.

Keywords: philosophy of technology, philosophical anthropology, artificial intelligence, technological determinism, Gestell, post-industrial society, digitalization, digital socialization, algorithmic systems, personal autonomy.

Alexander V. Kravets – Ph.D. Student, Department of Interdisciplinary Problems in the Advance of Science and Technology, Institute of Philosophy, Russian Academy of Sciences.

89031503674@mail.ru

<https://orcid.org/0009-0002-5250-6591>

For citation: Kravets A.V. (2024) Intelligent Technology and Threats to Human Subjectivity. *Russian Journal of Philosophical Sciences = Filosofskie nauki*. Vol. 67, no. 3, pp. 123–141. DOI: 10.30727/0235-1188-2024-67-3-123-141

Введение

Технологическая революция, начавшаяся в эпоху модерна, неизменно вызывала философские опасения относительно своего дегуманизирующего потенциала. Если ранее подобные тревоги можно было отнести к разряду алармистских, то современная эволюция технологий превратила многие философские предсказания из метафорических предупреждений в осязаемую реальность. Предсказания таких философов, как Х. Ортега-и-Гассет, М. Хайдеггер, Г. Маркузе, Ж. Элльюль и др., приобрели новую актуальность в контексте современных интеллектуальных технологий, особенно систем искусственного интеллекта.

Традиционное понимание техники как продолжения человеческого тела, как органопроекции, предложенное немецким философом Э. Каппом [Карп 1877], может казаться все менее убедительным перед лицом современных разработок. Сегодняшние интеллектуальные технологии демонстрируют беспрецедентную автономность в процессах принятия решений, распознавания паттернов и даже в творческих начинаниях, что ставит фундаментальные вопросы о взаимоотношениях между человеческой и технологической субъектностью.

Актуальность этих вопросов усиливается стремительным развитием систем искусственного интеллекта, которые не только расширяют, но и все чаще направляют и формируют человеческие когнитивные процессы. Современные технологические системы превращают человеческий опыт, предпочтения и поведение в материал для алгоритмической обработки, фундаментально изменяя природу человеческой субъектности и автономии. Эта трансформация выходит за рамки простой технологической помощи, ставя под вопрос основания человеческой субъектности.

В настоящей работе анализируются импликации интеллектуальной технологии через призму рефлексии философов XX века. Их критика техники, хотя и сформулированная в более раннем технологическом контексте, предоставляет важные теоретические рамки для понимания современных вызовов. Исследуя то, как современные интеллектуальные технологии потенциально ограничивают человеческую субъектность, данный анализ стремится внести вклад в наше понимание эволюционирующих отношений между человеческой субъектностью и интеллектуальными технологическими системами. Понимание природы и механики этих ограничений необходимо для того, чтобы в дальнейшем выработать меры, которые необходимы для сохранения человека как субъекта целеполагания, для сохранения его свободы в практиках когнитивной деятельности, социальных отношений и творческого самовыражения.

Интеллектуальная техника как постав

Концепция постава (Gestell), предложенная Хайдеггером, приобретает особую актуальность в контексте современных интеллектуальных технологий. Хайдеггер осуществляет фундаментальное переосмысление сущности техники, отвергая привычное «инструментальное и антропологическое определение техники», согласно которому она служит лишь средством человеческой деятельности. Как утверждает философ, «техника не то же, что сущность техники». Более того, «сущность техники вовсе не есть что-то техническое». Центральным в понимании техники у Хайдеггера становится понятие постава (Gestell): «По-ставом мы называем собирающее начало той установки, которая ставит, т.е. заставляет человека выводить действительное из его потаенности способом поставления его как состоящего-в-наличии (Bestand)» [Хайдеггер 1986, 55]. При этом человек оказывается захвачен этим процессом: «Захваченный поставляющим производством, человек стоит внутри сущностной сферы постава». Хайдеггер видит в этом фундаментальную опасность: «Господство по-става грозит той опасностью, что человек окажется уже не в состоянии вернуться к более исходному раскрытию потаенного и услышать голос более ранней истины» [Хайдеггер 1986, 61]. Это становится особенно актуальным сегодня, когда интеллектуальные технологии и искусственный интеллект начинают определять не только материальное производство, но и процессы мышления, творчества, познания. Однако философ указывает и на возможность спасения, цитируя Гельдерлина: «Но где опасность, там вырастает и спасительное». Это спасительное заключается в возможности нового отношения к технике, которое не сводится ни к слепому принятию, ни к полному отрицанию (в срединном пути, если использовать терминологию А.П. Давыдова). Применительно к современным интеллектуальным технологиям идеи Хайдеггера приобретают особую актуальность. Искусственный интеллект и связанные с ним технологии не просто автоматизируют когнитивные процессы, но формируют новый способ раскрытия действительности. Они превращают человеческое мышление, творчество, социальные отношения в исчислимый ресурс (Bestand) новой цифровой экономики и в целом цифровой, виртуальной системы воспроизводства множества социальных и культурных отношений.

Датафикация человеческого опыта. Экономика информационных технологий построена на превращении процессов, составляющих человеческую деятельность, в структурированные данные. Происходит своеобразная «датафикация» человеческого опыта [Mejías, Couldry 2019], когда осуществляется непрерывный сбор информации о действиях пользователя: от простых кликов и времени, проведенного на веб-страницах, до сложных поведенческих паттернов, отражающих когнитивные и эмоциональные реакции. Чем больше подобных данных

собрано, тем лучше посредством машинного обучения можно натренировать системы искусственного интеллекта на выявление неочевидных корреляций и паттернов в человеческом поведении. Методы глубокого обучения позволяют создавать детальные поведенческие профили, которые могут предсказывать не только потребительские предпочтения, но и более глубокие аспекты личности: политические взгляды, психологические характеристики, межличностные отношения. В результате формируется ситуация, в которой цифровой след человека становится более информативным для алгоритмических систем, чем сознательное самовыражение самого субъекта.

Экономика внимания и эмоциональный Bestand. Если сбор данных представляет собой внешний уровень трансформации человеческого опыта в «составляющее-в-наличии», то следующим, более глубоким уровнем является превращение человеческого внимания и эмоциональных реакций в экономический ресурс. В современной цифровой среде внимание пользователей становится одним из наиболее ценных и дефицитных ресурсов. Этот феномен, получивший название «экономика внимания» (впервые описан [Simon 1971, 40–41]), приобретает новое качество в контексте интеллектуальных технологий, способных к точной калибровке и оптимизации процессов привлечения и удержания внимания [Williams 2018].

Удержание внимание возможно за счет управления механизмами эмоционального реагирования. Трансформация эмоций в Bestand проявляется в нескольких взаимосвязанных аспектах. Во-первых, происходит алгоритмическая оптимизация контента для максимизации эмоционального отклика. Системы машинного обучения анализируют связь между характеристиками контента и эмоциональными реакциями пользователей, создавая все более точные модели эмоционального вовлечения. Во-вторых, формируются замкнутые циклы эмоционального потребления, при которых алгоритмические системы не только отслеживают, но и активно стимулируют определенные эмоциональные состояния.

Прослеживается тенденция к «геймификации» эмоционального опыта. Цифровые платформы внедряют механизмы социального сравнения, системы вознаграждений и другие элементы, заимствованные из игровой индустрии, для создания устойчивых паттернов эмоционального вовлечения. Это приводит к формированию специфической формы «эмоционального капитализма» [Ilouz 2007; Ilouz 2018], при которой эмоциональные реакции не просто отслеживаются и монетизируются, но активно производятся в соответствии с экономическими целями платформ. Интеллектуальная техника оказывается включенной в систему производства и потребления эмоциональных состояний, при этом граница между аутентичным переживанием и алгоритмически индуцированной реакцией становится все более размытой.

Социальные отношения как ресурс. Еще одним уровнем трансформации опыта человеческой субъектности в Bestand является превращение социальных отношений в алгоритмически обрабатываемый ресурс. Социальные сети и платформы создают цифровые репрезентации межличностных связей, превращая сложную ткань человеческих отношений в формализованные, измеримые и управляемые структуры данных [van Dijck et al. 2018]. Рекомендательные системы, основанные на анализе социальных графов, влияют на формирование новых знакомств, создание профессиональных контактов и даже выбор романтических партнеров. Этот процесс «платформизации» социальных отношений выходит за рамки простой цифровизации коммуникации, создавая новую форму социальности, при которой межличностные связи становятся объектом алгоритмического управления и экономических отношений. Значимым аспектом этой трансформации выступает коммодификация социальных взаимодействий. Социальные платформы превращают повседневные акты общения и обмена информацией в источник экономической ценности через механизмы таргетированной рекламы, монетизации контента и торговли данными о социальных связях [Couldry, Mejias 2019]. Формируется инструментальное отношение к социальным связям, при котором их ценность определяется преимущественно их потенциалом быть «оптимизированными» алгоритмами интеллектуальной техники.

Творчество и знание как ресурс совершенствования интеллектуальной техники. Если ранее сферы творчества и производства знаний рассматривались как область исключительно человеческой деятельности, требующей уникальных когнитивных способностей и креативного мышления, то современные системы искусственного интеллекта все активнее вторгаются в эти сферы, превращая творческие и интеллектуальные процессы в алгоритмически воспроизводимые операции [Miller 2019; Computational Creativity... 2019]. Творческие произведения, созданные людьми, становятся «тренировочными данными» для алгоритмических систем, превращаясь из уникальных выражений человеческой субъектности в ресурс для машинного обучения.

В сфере производства знаний наблюдается аналогичный процесс алгоритмизации. Системы обработки естественного языка и анализа данных все чаще используют не только для поиска и систематизации информации, но и для генерации научных гипотез, написания академических текстов и проведения исследований [Буданов 2023, 28–29; Буданов 2024]. Происходит стандартизация творческих и интеллектуальных процессов под влиянием алгоритмических систем. В стремлении оптимизировать производство контента для цифровых платформ и научных баз данных творческие и научные работники все чаще адаптируют свои практики к требованиям алгоритмической видимости и рекомендательных систем [Nieborg, Poell 2018]. Это создает

риск формирования «алгоритмической» эстетики и науки, в процессе которого творческие и исследовательские решения определяются не художественными или интеллектуальными соображениями, а метриками эффективности и вовлеченности.

Поскольку результаты творчества и знания начинают функционировать как Bestand, трансформируется отношение к авторству и оригинальности. Происходит фундаментальное переосмысление представлений об уникальности творческого акта и интеллектуальной деятельности. Художественное произведение и научная работа перестают быть исключительно результатом деятельности личности, а становятся продуктом технологизированных интеллектуальных креативных процессов, в которых грань между человеческим и машинным вкладом настолько размывается, что можно говорить о феномене «распределенного» авторства [Goodfellow 2024].

Исследования в сфере машинного обучения и систем искусственного интеллекта становятся ведущим направлением современной науки, привлекая непропорционально большие объемы финансирования по сравнению с другими областями исследований. В результате формируется ситуация, в которой приоритетную ценность приобретает знание, способствующее совершенствованию интеллектуальных технологий. Таким образом, интеллектуальная техника превращается в своеобразный «эпистемический постав», а интеллектуальное наследие и деятельность людей служат ресурсом ее развития. Цифровые технологии способствуют трансформации отношений человека с окружающим миром, что неизбежно затрагивает и природу человека. Именно этот процесс приводит к тому, что вторая природа человека, сформированная культурой и обществом, подвергается угрозе упрощения и стандартизации под влиянием цифровых технологий.

Цифровая примитивизация второй природы человека

Превращение различных аспектов человеческой деятельности в «состоящее-в-наличии» неразрывно связано с процессом дегуманизации, который был предметом критического анализа философов XX века. Например, Ортега-и-Гассет указывал на риски «вертикального варварства» – ситуации, в которой технологический прогресс опережает развитие человеческой культуры и способности к его осмыслению [Ортега-и-Гассет 2000, 92]. Так, испанский философ выделял три исторические стадии развития техники: технику случая (первобытную), технику ремесла (традиционную) и технику человека-техника (современную). На первой стадии технические изобретения были случайными находками, на второй – результатом следования традиции, и только на третьей стадии технику осознают как особую способность человека систематически преобразовывать природу (и в настоящее время, как представляется, это уже не внешняя человеку природа, а природа само-

го человека). В современной технике философ видит и достижения, и опасности. С одной стороны, она дает небывалые возможности, служит реализации «жизненного проекта» человека, его представлений о благосостоянии и «хорошей жизни». С другой – техника порождает целый ряд рисков для развития человека. Во-первых, это утрата человеком понимания сущности техники. По мере того как техническая среда становится все более развитой и всеохватывающей, человек начинает воспринимать ее как нечто естественное и самоочевидное, подобно тому как первобытный человек воспринимал природу («человек пришел к ложной мысли, что и остальное, т.е. все окружающее, подобно такой изначальной природе... это вовсе не то, что сначала нужно было произвести» [Ортега-и-Гассет 2000, 23]). Это ведет к забвению того факта, что техника – продукт человеческого духа и культуры, требующий постоянного поддержания духовных условий. Во-вторых, если ранее инструмент был продолжением человеческого тела и подчинялся ремесленнику, то теперь «не орудие служит человеку, а наоборот: человек – придаток машины». Это ведет к отчуждению человека от технического процесса и утрате им роли подлинного субъекта технической деятельности. В-третьих, избыточное развитие техники может привести к «опустошению человеческой жизни». Поскольку техника сама по себе является лишь чистой возможностью и не определяет содержания жизни, чрезмерная вера в технику при утрате других жизненных ориентиров ведет к экзистенциальной пустоте («быть техником и только техником означает иметь возможность быть всем и, следовательно, ничем» [Ортега-и-Гассет 2000, 221]). Технический прогресс вызывает «кризис желаний» – утрату современным человеком способности формировать подлинные жизненные проекты, которые только и могут придать смысл техническому развитию. Каждый новый технологический этап строится на фундаменте достижений предыдущих эпох, и то, что раньше считалось самоценностью, на новом этапе может превратиться в ресурс или оказаться забытым. Способность человека отходить от стандартных шаблонов мышления подвергается давлению со стороны цифровой среды, и любое проявление креативности вне рамок взаимодействия с техносферой рискует быть гильотинированным за ненадобностью.

Стандартизация когнитивных процессов. Этот феномен проявляется в формировании устойчивых паттернов мышления и познания, опосредованных алгоритмическими системами. Если традиционные формы стандартизации мышления, описанные в работах критической теории (франкфуртская школа), связаны преимущественно с массовой культурой и бюрократическими системами, то современные интеллектуальные технологии создают более глубокий и систематический эффект когнитивной унификации. Важным механизмом стандартизации является алгоритмическая медиация познавательных процессов.

Поисковые системы, рекомендательные алгоритмы и системы автоматической обработки информации формируют устойчивые паттерны информационного поведения, то, что можно условно назвать «алгоритмическими привычками мышления», т.е. упрощением сложных проблем до уровня, поддающегося алгоритмической обработке, и предпочтением легко квантифицируемых решений над качественными оценками.

Тревогу вызывает влияние алгоритмических систем на процессы категоризации и концептуализации опыта. Человеческое понимание основано на сложном взаимодействии явного и неявного знания (М. Полани), контекстуального восприятия и интуитивного схватывания ситуации. Однако алгоритмические системы ввиду своей архитектуры требуют формализации знания в четкие, дискретные категории. Это создает тенденцию к редукции многомерного человеческого опыта к набору стандартизированных параметров.

Процесс когнитивной стандартизации усиливается через механизмы «алгоритмического обучения» [Perrotta, Williamson 2016]. Образовательные платформы и системы управления обучением все чаще используют алгоритмические методы для оптимизации образовательного процесса, что приводит к стандартизации не только содержания обучения, но и самих способов усвоения и обработки информации [Hughes 2001]. В результате формируются унифицированные модели когнитивного развития, ориентированные на эффективность и измеримость результатов, но потенциально ограничивающие развитие нестандартных форм мышления.

Интеллектуальная зависимость и депрофессионализация. Формирование интеллектуальной зависимости от алгоритмических систем может приводить к депрофессионализации в различных сферах деятельности. В отличие от традиционных форм технологической зависимости, современная ситуация характеризуется тем, что интеллектуальные технологии не просто дополняют, но все чаще замещают базовые когнитивные операции и профессиональные компетенции. Феномен интеллектуальной зависимости проявляется прежде всего в ослаблении способности к автономному решению задач. Регулярное использование интеллектуальных систем поддержки принятия решений может приводить к атрофии навыков самостоятельного анализа и критического мышления [Carr 2014]. Можно говорить о новом виде прокрастинации, при котором наиболее сложные аспекты профессиональной деятельности отданы на выполнение искусственным системам, что приводит к формированию психологического барьера к решению сложных задач. Человек, оказавшись в зависимости от систем искусственного интеллекта, впадает в стресс в ситуации, если задачу приходится решать самостоятельно [Ahmad et al. 2023].

Трансформация профессионального знания под влиянием интеллектуальных технологий затрагивает и процесс профессиональной

подготовки. Наблюдаются два противоположно направленных процесса: повышение эффективности профессиональной деятельности за счет использования интеллектуальных технологий и снижение уровня реальной профессиональной компетентности [Казарян 2023, 215]. Это ставит вопрос о необходимости разработки новых подходов к профессиональному развитию, которые позволили бы сохранить баланс между использованием преимуществ интеллектуальных технологий и поддержанием автономной профессиональной компетентности.

Культурная унификация и гомогенизация. Массовое использование и внедрение в различные информационные продукты систем искусственного интеллекта, построенных как на алгоритмической, так и на нейросетевой архитектуре, носит глобальный характер, практически не учитывающий специфику локальных культур. Если традиционные процессы глобализации и массовой культуры сохраняли пространство для культурного разнообразия, то алгоритмическая медиация культурного производства и потребления создает более глубокий и систематический эффект гомогенизации.

Ключевым механизмом этого процесса служит алгоритмическая оптимизация культурного контента. Глобальные цифровые платформы, стремясь максимизировать вовлеченность пользователей, создают системы рекомендаций, которые неизбежно фаворизируют определенные форматы и стили культурного выражения [Jin 2020, 6–7]. Это приводит к формированию того, что можно назвать «алгоритмической эстетикой» – набора культурных форм, оптимизированных для распространения через цифровые платформы. Опасения вызывает влияние алгоритмических систем на локальные культурные контексты. Американский исследователь Т. Стрипас вводит понятие «алгоритмическая культура», описывая ситуацию, в которой локальные культурные практики и формы знания постепенно вытесняются стандартизированными форматами, соответствующими логике цифровых платформ [Striphas 2015]. Происходит не просто замещение традиционных культурных форм, но и трансформация способов культурного производства и трансляции. Значимым аспектом культурной унификации является трансформация темпоральности культурного опыта. Алгоритмические системы, оптимизированные для максимизации вовлеченности, создают «социальное ускорение» [Rosa 2013], ускоренное культурное потребление, что препятствует формированию глубоких культурных смыслов и устойчивых культурных практик.

Процесс культурной унификации усиливается через механизмы «цифровой социализации» [Солдатова, Войскунский 2021]. Социальные сети и цифровые платформы становятся основным пространством культурного обмена и формирования идентичности, особенно для молодого поколения. Алгоритмические системы, управляющие этими платформами, создают глобальные унифицированные паттерны

социального взаимодействия и культурного потребления, которые постепенно вытесняют локальные формы социальности. Таким образом, при видимом увеличении разнообразия доступного культурного контента происходит стандартизация форм культурного производства, потребления и воспитания. Это ставит фундаментальный вопрос о возможности сохранения культурной аутентичности и разнообразия в условиях всепроникающей технологизации и медиатизации культурного опыта. Данная тенденция ведет к усилению контроля над обществом через технологии, что делает проблему технологического детерминизма особенно значимой.

Технологический детерминизм

В контексте развития интеллектуальных технологий актуальность приобретает проблема технологического детерминизма, опасность которого обсуждалась еще до распространения информационных технологий. Французский философ Ж. Эллюль писал о формировании автономной «технологической системы», которая развивается по собственной логике и навязывает свои императивы всем сферам человеческой жизни. Эллюль определяет технику как «совокупность рационально разработанных методов, обладающих абсолютной эффективностью (для данной стадии развития) во всех областях человеческой деятельности» [Ellul 1964, xxv]. Принципиально важным в его концепции является понимание техники не просто как набора инструментов или машин, а как всеобъемлющей системы, которая развивается по собственным законам и обладает автономностью («техника в конечном итоге зависит только от себя, она прокладывает свой собственный путь, это первичный, а не вторичный фактор») [Ellul 1980, 125]. Техника, по Эллюлю, выступает как посредник между человеком и его естественной средой. Посредничество становится тотальным, вытесняя остальные формы взаимодействия человека с природой: поэтические, магические, мифологические, символические. В результате техника превращается в единственный и всеобъемлющий способ взаимодействия человека с миром. Еще одна важная характеристика техники, по Эллюлю, – это ее универсальность. Технологическая система распространяется по всему миру, независимо от политических режимов, культурных традиций или уровня экономического развития, что приводит к глобальной унификации образа жизни, культуры и мышления людей. Однако эта универсальность не означает объединения человечества, а наоборот, может усиливать конфликты и неравенство между странами. Наконец, Эллюль пишет о тотализации техники, о том, что техника стремится охватить все сферы жизни, превращая объекты и процессы в технологические. Это приводит к созданию тотальной технологической системы, которая контролирует все аспекты челове-

ческого существования. В результате человек теряет автономность и становится частью такой системы.

Тема детерминации человека техникой получает развитие и в работе «Одномерный человек» Г. Маркузе. В анализе техники он исходит из того, что в современном обществе техника перестает быть лишь набором нейтральных средств и инструментов, а превращается в форму социального контроля и господства: «Технологическая рациональность становится политической рациональностью» [Маркузе 1994, XX]. Техника воплощает в себе проект организации природы и общества, направленный на установление тотального господства. Технологическая рациональность создает «одномерного человека» – человека, полностью интегрированного в существующую политическую систему. При этом техническое развитие делает возможным и иной проект, в частности проект освобождения человека и создания «умиротворенного существования». Однако существующее общество блокирует реализацию таких возможностей. Технологическая рациональность ограничивает способность человека к трансцендированию существующего порядка – к представлению и созданию качественно иных форм жизни. Важным аспектом является то, что технологическая рациональность создает иллюзию свободы при фактическом усилении несвободы. Она блокирует «двухмерное», диалектическое мышление, способное видеть противоречия действительности и альтернативные возможности развития. Применительно к современной интеллектуальной технике (искусственному интеллекту, алгоритмическим системам и т.д.) анализ Маркузе, несмотря на свой, возможно, излишне политизированный характер, остается актуальным. Эти технологии также воплощают в себе проект социального контроля и господства, трансформируя человеческое сознание и поведение. Они создают новые формы зависимости и несвободы, маскируя их под рациональность и эффективность. Такие же технологии потенциально могли бы служить освобождению человека, но пока даже на концептуальном уровне нет консенсуса относительно вопроса о том, что предполагает свобода в эпоху искусственного интеллекта и цифровизации всех сфер жизни.

Эрозия человеческой агентности и автономии. Если традиционные формы технологической детерминации ограничивали свободу действия преимущественно в материальной сфере, то современные интеллектуальные системы воздействуют на процесс формирования намерений и принятия решений. Особенность этого процесса заключается в его незаметности и кажущейся добровольности. Современные системы «подталкивания» (nudging) [Yeung 2017] и персонализированных рекомендаций создают иллюзию расширения возможностей выбора, в то время как фактически происходит сужение горизонта возможных действий до предсказуемого и алгоритмически оптимизированного набора опций. Ш. Зубофф полагает, что сформирован «предсказательный императив»

(prediction imperative), требующий от компаний для генерации прибыли как можно точнее предсказывать поведение пользователей [Zuboff 2019]. При этом интеллектуальные системы, основываясь на анализе больших данных, не просто предсказывают, но активно формируют будущее поведение пользователей через создание ряда условий и стимулов. Все большее количество социальных решений, от выбора партнеров до профессиональных контактов, опосредуется и направляется интеллектуальными технологическими системами.

Традиционное понимание автономии как способности к самостоятельному формированию и реализации намерений требует переосмысления в контексте того, что сами механизмы формирования намерений оказываются объектом технологического воздействия. В качестве одного из решений этой проблемы ведущий специалист в сфере этики искусственного интеллекта Л. Флориди предлагает сохранить за человеком «мета-автономию», или автономию в решении того, что следует или не следует делегировать системам искусственного интеллекта [Floridi 2019, 62].

Итогом описанных процессов является следующее: чем более точными в цифровой среде становятся предсказательные модели, тем менее автономным оказывается поведение субъекта. Возникает своеобразная форма «мягкого детерминизма», при которой свобода выбора формально сохраняется, но фактически ограничивается рамками предсказательных моделей и основанных на них рекомендательных систем. Поэтому особую актуальность приобретает вопрос о разработке новых форм «критической агентности», способных сохранять пространство для автономного действия в условиях всепроникающих искусственных интеллектуальных систем.

От цифрового контроля к цифровому конструированию человека.

В отличие от традиционных форм социального контроля, современные технологии создают возможность не только для тотального наблюдения, но и для активного конструирования человеческой субъектности. То, что М. Фуко называл «микрофизикой» власти, еще глубже проникает ткань социальности. В отличие от классических форм контроля, исходящих от политической власти, тот «микрофизический» характер воздействия на человека, который присущ современной интеллектуальной технике, не связан напрямую с какой-то идеологией и поэтому проникает в общества с разными политическими устройствами. Цифровая среда, созданная этой техникой, служит уже одним из главных факторов формирования идентичности человека. Некоторые исследователи пишут об «алгоритмической идентичностью» – практике категоризации и профилирования пользователей на основе их цифровых следов [Cheney-Lippold 2011]. Эта практика выходит за рамки простого наблюдения, создавая динамические модели идентичности, которые активно влияют на то, как человек воспринимает себя и свои возмож-

ности. Происходит своеобразная «колонизация жизненного мира» (в терминологии Ю. Хабермаса) искусственными интеллектуальными системами, которые трансформируют фундаментальные структуры человеческого существования. Наблюдается то, что можно назвать «техно-социальной рекурсией» – процессом, при котором алгоритмические системы не только реагируют на социальное поведение, но и создают условия для его воспроизводства в определенных формах. Грань между естественным поведением и поведением, индуцированным технологическими системами, становится все более размытой. Социолог Д. Бизр вводит понятие «метрическая власть» [Beer 2016]. Речь идет о власти, основанной на непрерывном измерении и количественной оценке всех аспектов социальной жизни. Возможность быть измеренным становится условием социального существования. Социальный исследователь Л. Амор пишет об «облачной этике» [Amoore 2000] – системе принятия решений, в рамках которой этические суждения встраиваются непосредственно в алгоритмический код и автоматически применяются к множеству случаев без возможности контекстуального пересмотра. Это создает риск формализации и механизации этического мышления.

Остается открытым вопрос о том, будет ли формируемая интеллектуальной техникой социальность гомогенной, т.е. со стертыми различиями между социальными группами. Или, наоборот, будет происходить «алгоритмическое усиление» социальных предубеждений и неравенств [Noble 2018] ввиду того, что системы «алгоритмического управления» будут не только отражать существующие социальные иерархии, но и усиливать их через механизмы автоматизированной категоризации и дифференцированных рекомендаций. И, конечно, возникает вопрос о том, существуют ли принципы, нормы, цели, которые могут остаться основанием для автономии человеческой индивидуальности, сохранить за человеком возможность формировать самого себя в цифровой среде интеллектуальных технологий.

Заключение

Развитие интеллектуальных технологий создает принципиально новую ситуацию в отношениях между человеком и техникой. Если традиционное понимание техники как инструмента или проекции человеческого тела (Э. Капп) могло быть применимо к индустриальным технологиям, то современные интеллектуальные системы выходят за рамки инструментального статуса, формируя автономную технологическую среду, которая преобразует все аспекты человеческого существования и самого человека.

Превращение различных аспектов человеческой деятельности в «состоящее-в-наличии» (Bestand) приобретает новое качество в контексте цифровых технологий. Речь идет уже не просто об инструментальном

использовании ресурсов, но о фундаментальной трансформации человеческой субъектности, когда сами процессы мышления, эмоционального реагирования и социального взаимодействия становятся ресурсом для развития технологических систем. Особую значимость приобретает то обстоятельство, что эта трансформация часто воспринимается не как ограничение, а как расширение возможностей, что затрудняет критическое осмысление происходящих изменений. Процесс эрозии того, что составляет человеческую природу, в контексте интеллектуальных технологий приобретает более сложные и не прямые формы по сравнению с изложенными в рамках философских теорий XX века. Стандартизация когнитивных процессов, интеллектуальная зависимость и культурная унификация реализуются не через прямое подавление человеческих качеств, а через создание условий, в которых определенные формы мышления и поведения становятся более «естественными» и предпочтительными. Сама способность к автономному формированию и реализации человеческого потенциала оказывается под вопросом.

Эта ситуация ставит ряд вопросов, требующих серьезного философского и социального осмысления. Первый связан с возможностью сохранения человеческой автономии и субъектности в условиях всепроникающей алгоритмической медиации. Второй – с поиском форм социальной организации, которые позволили бы использовать преимущества интеллектуальных технологий без утраты существенных человеческих качеств. Третий затрагивает проблему разработки новых форм критического мышления и социальной рефлексии, способных противостоять тенденциям технологического детерминизма.

Изложенная проблематика, от определения статуса и границ человеческой субъектности до поиска форм ее медиации в условиях взаимодействия человека и машины, может быть концептуализирована как поиск меры синтеза человеческого и технологического начал. Поиск «меры совместности» (термин А.П. Давыдова) приобретает сегодня междисциплинарный характер [Давыдов 2023, 145–146, 149]. Как отмечает А.П. Давыдов, в современных условиях техносубъект выступает в качестве Другого, что придает проблеме отношений между «Я» и Другим, поставленной неклассической философской антропологией, новое измерение.

Скорее всего, эти проблемы не могут быть решены только на общественном уровне. Необходимо правовое регулирование разработки и использования интеллектуальных технологических систем. Но для внедрения такого регулирования следует выработать позицию относительно ряда принципиальных аспектов. Первый аспект касается определения правового статуса интеллектуальных систем. Существующие правовые рамки, основанные на традиционном понимании субъектности и ответственности, оказываются недостаточными для регулирования систем, обладающих высокой степенью автономности

в принятии решений. Возникает необходимость разработки новых принципов прав, свобод и ответственности, применимых ко всем новым типам субъектности и агентности. Второй аспект связан с механизмами имплементации правового регулирования на локальном и международном уровнях, поскольку транснациональный характер современных технологических систем создает существенные вызовы для традиционных механизмов правового регулирования, основанных на принципе территориального суверенитета. Третий, наиболее фундаментальный аспект выражен в определении и правовом закреплении границ допустимой технологической медиации человеческой деятельности. Возможно, требуется нормативное определение пределов, за которыми стремление к технологическому удобству и эффективности начинает угрожать человеческой субъектности.

Таким образом, проблема взаимоотношений между человеком и технологиями выходит за рамки сугубо технических или социальных вопросов, становясь центральной проблемой современной философской антропологии. От того, насколько успешно человечество сможет осмыслить и разрешить эту проблему, во многом зависит сохранение качеств и способностей, которые традиционно определяли специфику человеческого существования.

ЦИТИРУЕМАЯ ЛИТЕРАТУРА

Буданов 2023 – Буданов В.Г. Антропосоциальные вызовы экспансии искусственного интеллекта // Ученые записки Института психологии Российской академии наук. 2023. Т. 3. № 2. С. 23–31.

Буданов 2024 – Буданов В.Г. Техносубъект и антропосоциальные проблемы взаимодействия с искусственным интеллектом: синергия, демаркация, новая рациональность, риски // Философские науки. 2024. Т. 67. № 3. С. 27–52.

Давыдов 2023 – Давыдов А.П. Индивидуализм и коллективизм как предмет социально-философского анализа. Размышления в преддверии научной конференции «Индивидуализация и коллективизм в современном российском обществе» // Философские науки. 2023. Т. 66. № 4. С. 140–159.

Казарян 2023 – Казарян А.Ю. Искусственный интеллект в процессах образования и обучения, положительные и отрицательные стороны // Международный журнал гуманитарных и естественных наук. 2023. № 11-2. С. 214–216.

Маркузе 1994 – Маркузе Г. Одномерный человек / пер. с англ. А. Юдина. – М.: REFL-book, 1994.

Ортега-и-Гассет 2000 – Ортега-и-Гассет Х. Размышления о технике // Ортега-и-Гассет Х. Избранные труды / сост. и общ. ред. А.М. Руткевича. – М.: Инфра-М; Весь мир, 2000. С. 164–227.

Солдатова, Войскунский 2021 – Солдатова Г.У., Войскунский А.Е. Социально-когнитивная концепция цифровой социализации: новая экосистема и социальная эволюция психики // Психология. Журнал Высшей школы экономики. 2021. Т. 18. № 3. С. 431–450.

Хайдеггер 1986 – Хайдеггер М. Вопрос о технике / пер. с нем. В.В. Бихина // Новая технократическая волна на западе / сост. П.С. Гуревич. – М.: Прогресс, 1986. С. 45–67.

Ahmad et al. 2023 – Ahmad S.F., Han H., Alam M.M., Rehmat M., Irshad M., Arraño-Muñoz M., Ariza-Montes A. Impact of Artificial Intelligence on Human

Loss in Decision Making, Laziness and Safety in Education // Humanities and Social Sciences Communications. 2023. Vol. 10. No. 1. P. 1–14.

Amoore 2020 – *Amoore L.* Cloud Ethics: Algorithms and the Attributes of Ourselves and Others. – Durham: Duke University Press, 2020.

Beer 2016 – *Beer D.* Metric Power. – London: Palgrave Macmillan UK, 2016.

Carr 2014 – *Carr N.* The Glass Cage: How Our Computers are Changing Us. – New York: W.W. Norton & Company, 2014.

Cheney-Lippold 2011 – *Cheney-Lippold J.* A New Algorithmic Identity: Soft Biopolitics and the Modulation of Control // Theory, Culture & Society. 2011. Vol. 28. No. 6. P. 164–181.

Computational Creativity... 2019 – Computational Creativity: The Philosophy and Engineering of Autonomously Creative Systems / ed. by T. Veale, F.A. Cardoso. – Cham: Springer, 2019.

Couldry, Mejias 2019 – *Couldry N., Mejias U.A.* The Costs of Cnnection: How Data is Colonizing Human Life and Appropriating It for Capitalism. – Stanford: Stanford University Press, 2019.

Ellul 1964 – *Ellul J.* The Technological Society. – New York: Vintage Books, 1964.

Ellul 1980 – *Ellul J.* The Technological System. – New York: Continuum Publishing, 1980.

Floridi 2019 – *Floridi L.* The Ethics of Artificial Intelligence. – Oxford: Oxford University Press, 2019.

Goodfellow 2024 – *Goodfellow P.* The Distributed Authorship of Art in the Age of AI // Arts. 2024. Vol. 13. No. 5. Article 149.

Hughes 2001 – *Hughes J.* The Deskilling of Teaching and the Case for Intelligent Tutoring Systems // Journal of Ethics and Emerging Technologies. 2001. Vol. 31. No. 2. P. 1–16.

Illouz 2007 – *Illouz E.* Cold Intimacies: The Making of Emotional Capitalism. – Cambridge: Polity, 2007.

Illouz 2018 – *Illouz E.* Emotions as Commodities: Capitalism, Consumption and Authenticity. – London: Routledge, 2018.

Jin 2020 – *Jin D.Y.* Artificial Intelligence in Cultural Production: Critical Perspectives on Digital Platforms. – London: Routledge, 2020.

Kapp 1877 – *Kapp E.* Grundlinien einer Philosophie der Technik. Zur Entstehungsgeschichte der Cultur aus neuen Gesichtspunkten. – Braunschweig: Georg Westermann, 1877.

Mejias, Couldry 2019 – *Mejias U.A., Couldry N.* Datafication // Internet Policy Review. 2019. Vol. 8. No. 4. P. 1–10.

Miller 2019 – *Miller A.I.* The Artist in the Machine: The World of AI-Powered Creativity. – Cambridge: MIT Press, 2019.

Nieborg, Poell 2018 – *Nieborg D.B., Poell T.* The Platformization of Cultural Production: Theorizing the Contingent Cultural Commodity // New Media & Society. 2018. Vol. 20. No. 11. P. 4275–4292.

Noble 2018 – *Noble S.U.* Algorithms of Oppression: How Search Engines Reinforce Racism. – New York: NYU Press, 2018.

Perrotta, Williamson 2016 – *Perrotta C., Williamson B.* The Social Life of Learning Analytics: Cluster Analysis and the ‘Performance’ of Algorithmic Education // Learning, Media and Technology. 2016. Vol. 43. No. 1. P. 3–16.

Rosa 2013 – *Rosa H.* Social Acceleration: A New Theory of Modernity. – New York: Columbia University Press, 2013.

Simon 1971 – *Simon H.A.* Designing Organizations for an Information-Rich World // Computers, Communications, and the Public Interest / ed. by M. Greenberger. – Baltimore: Johns Hopkins Press, 1971. P. 37–72.

Striphas 2015 – *Striphas T.* Algorithmic Culture // European Journal of Cultural Studies. 2015. Vol. 18. No. 4–5. P. 395–412.

- van Dijck et al. 2018 – *van Dijck J., Poell T., de Waal M.* The Platform Society: Public Values in a Connective World. – Oxford: Oxford University Press, 2018.
- Williams 2018 – *Williams J.* Stand Out of Our Light: Freedom and Resistance in the Attention Economy. – Cambridge: Cambridge University Press, 2018.
- Yeung 2017 – *Yeung K.* ‘Hypernudge’: Big Data as a Mode of Regulation by Design // Information, Communication & Society. 2017. Vol. 20. No. 1. P. 118–136.
- Zuboff 2019 – *Zuboff S.* The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power. – New York: Public Affairs, 2019.

REFERENCES

- Ahmad S.F., Han H., Alam M.M., Rehmat M., Irshad M., Arraño-Muñoz M., & Ariza-Montes A. (2023) Impact of Artificial Intelligence on Human Loss in Decision Making, Laziness and Safety in Education. *Humanities and Social Sciences Communications*. Vol. 10, no. 1, pp. 1–14.
- Amoore L. (2020) *Cloud Ethics: Algorithms and the Attributes of Ourselves and Others*. Durham: Duke University Press.
- Beer D. (2016) *Metric Power*. London: Palgrave Macmillan UK.
- Budanov V.G. (2023) Anthroposocial Challenges of Artificial Intelligence Expansion. *Proceedings of the Institute of Psychology of Russian Academy of Sciences*. Vol. 3, no. 2, pp. 23–31 (in Russian).
- Budanov V.G. (2024) Technosubject and Anthroposocial Challenges of Human–Artificial Intelligence Interaction: Synergy, Demarcation, New Rationality, and Risks. *Russian Journal of Philosophical Sciences = Filosofskie nauki*. Vol. 67, no. 3, pp. 27–52 (in Russian).
- Carr N. (2014) *The Glass Cage: How Our Computers Are Changing Us*. New York: W.W. Norton & Company.
- Cheney-Lippold J. (2011) A New Algorithmic Identity: Soft Biopolitics and the Modulation of Control. *Theory, Culture & Society*. Vol. 28, no. 6, pp. 164–181.
- Couldry N. & Mejias U.A. (2019) *The Costs of Connection: How Data Is Colonizing Human Life and Appropriating It for Capitalism*. Stanford: Stanford University Press.
- Davydov A.P. (2023) Individualism and Collectivism as a Subject of Social-Philosophical Analysis (Reflections on the Eve of the Scientific Conference “Individualization and Collectivism in Contemporary Russian Society”). *Russian Journal of Philosophical Sciences = Filosofskie nauki*. Vol. 66, no. 4, pp. 140–159 (in Russian).
- Ellul J. (1964) *The Technological Society*. New York: Vintage Books.
- Ellul J. (1980) *The Technological System*. New York: Continuum Publishing.
- Floridi L. (2019) *The Ethics of Artificial Intelligence*. Oxford: Oxford University Press.
- Goodfellow P. (2024) The Distributed Authorship of Art in the Age of AI. *Arts*. Vol. 13, no. 5, article 149.
- Heidegger M. (1986) The Question Concerning Technology (V.V. Bikhin, Trans.). In: Gurevich P.S. (Comp.) *The New Technocratic Wave in the West* (pp. 45–67). Moscow: Progress (Russian translation).
- Hughes J. (2001) The Deskillings of Teaching and the Case for Intelligent Tutoring Systems. *Journal of Ethics and Emerging Technologies*. Vol. 31, no. 2, pp. 1–16.
- Illouz E. (2007) *Cold Intimacies: The Making of Emotional Capitalism*. Cambridge: Polity.
- Illouz E. (2018) *Emotions as Commodities: Capitalism, Consumption and Authenticity*. London: Routledge.
- Jin D.Y. (2020) *Artificial Intelligence in Cultural Production: Critical Perspectives on Digital Platforms*. London: Routledge.

Kapp E. (1877) *Grundlinien einer Philosophie der Technik. Zur Entstehungsgeschichte der Cultur aus neuen Gesichtspunkten*. Braunschweig: Georg Westermann (in German).

Kazaryan A.Yu. (2023) Artificial Intelligence in the Processes of Education and Training, Positive and Negative Sides. *Mezhdunarodnyy zhurnal gumanitarnykh i estestvennykh nauk*. No. 11-2, pp. 214–216 (in Russian).

Marcuse H. (1994) *One-Dimensional Man* (A. Yudin, Trans.). Moscow: REFL-book (Russian translation).

Mejias U.A. & Couldry N. (2019) *Datafication. Internet Policy Review*. Vol. 8, no. 4, pp. 1–10.

Miller A.I. (2019) *The Artist in the Machine: The World of AI-Powered Creativity*. Cambridge: MIT Press.

Nieborg D.B. & Poell T. (2018) The Platformization of Cultural Production: Theorizing the Contingent Cultural Commodity. *New Media & Society*. Vol. 20, no. 11, pp. 4275–4292.

Noble S.U. (2018) *Algorithms of Oppression: How Search Engines Reinforce Racism*. New York: NYU Press.

Ortega y Gasset J. (2000) Reflections on Technology. In: Ortega y Gasset J. *Selected Works* (pp. 164–227). Moscow: Infra-M; Ves' mir (Russian translation).

Perrotta C. & Williamson B. (2016) The Social Life of Learning Analytics: Cluster Analysis and the 'Performance' of Algorithmic Education. *Learning, Media and Technology*. Vol. 43, no. 1, pp. 3–16.

Rosa H. (2013) *Social Acceleration: A New Theory of Modernity*. New York: Columbia University Press.

Simon H.A. (1971) Designing Organizations for an Information-Rich World. In: Greenberger M. (Ed.) *Computers, Communications, and the Public Interest* (pp. 37–72). Baltimore: Johns Hopkins Press.

Soldatova G.U. & Voiskunsky A.E. (2021) Socio-Cognitive Concept of Digital Socialization: A New Ecosystem and Social Evolution of the Psyche. *Psychology. Journal of the Higher School of Economics*. Vol. 18, no. 3, pp. 431–450 (in Russian).

Striphas T. (2015) Algorithmic Culture. *European Journal of Cultural Studies*. Vol. 18, no. 4–5, pp. 395–412.

van Dijck J., Poell T., & de Waal M. (2018) *The Platform Society: Public Values in a Connective World*. Oxford: Oxford University Press.

Veale T. & Cardoso F.A. (Eds.) (2019) *Computational Creativity: The Philosophy and Engineering of Autonomously Creative Systems*. Cham: Springer.

Williams J. (2018) *Stand Out of Our Light: Freedom and Resistance in the Attention Economy*. Cambridge: Cambridge University Press.

Yeung K. (2017) 'Hypernudge': Big Data as a Mode of Regulation by Design. *Information, Communication & Society*. Vol. 20, no. 1, pp. 118–136.

Zuboff S. (2019) *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power*. New York: Public Affairs.



НАУЧНАЯ ЖИЗНЬ



Интеллектуальный путь



DOI: 10.30727/0235-1188-2024-67-3-142-158

Personalia

Сознание и мозг. Опыт разработки основных теоретических вопросов «Трудной проблемы сознания» и ее значение для нейронауки

Д.И. Дубровский

Институт философии Российской академии наук, Москва, Россия

Аннотация

Статья представляет собой научный самоотчет автора, подводящего итоги своей 70-летней деятельности в области философии и науки, в центре которой находится «Трудная проблема сознания», проблема соотношения психических и нейрофизиологических явлений. Автор описывает свой путь в исследовании этой фундаментальной для естественно-научного познания проблемы, излагает концептуальный подход к ее решению, предложенный еще в начале 60-х годов прошлого века. Значительное внимание уделено трудностям, с которыми автору пришлось столкнуться при отстаивании своей концепции, основанной на информационном подходе. В частности, освещена острая дискуссия с Э.В. Ильенковым, показано противодействие со стороны его сторонников, занимавших влиятельные позиции в философском сообществе. Несмотря на идеологическое давление и обвинения в ревизионизме, автору удалось продолжить развитие своей теории и опубликовать ряд книг и статей, в которых раскрыты разработанные им теоретико-методологические вопросы расшифровки мозговых кодов явлений субъективной реальности. В 2019 году, после публикации статьи на английском языке в одном из ведущих международных нейронаучных журналов, в которой представлены основные положения его теории, автор получил огромное количество откликов, приглашений выступить спикером на международных конференциях и войти в редколлегии зарубежных изданий. Это свидетельствует о высокой актуальности разработки теоретических и методологических вопросов в области нейронауки и смежных дисциплин, о дефиците фундамен-

тальных теоретических разработок, способных направлять и интегрировать эмпирические исследования. В заключение автор обращает внимание на важность информационной парадигмы и основанных на ней подходов для изучения сознания, психического и генетического аспектов жизнедеятельности.

Ключевые слова: философия сознания, психофизиологическая проблема, субъективная реальность, информационный подход, генетический код, расшифровка мозговых кодов психических явлений, нейронаука, нейрофилософия, искусственный интеллект.

Дубровский Давид Израилевич – доктор философских наук, профессор, главный научный сотрудник сектора теории познания Института философии РАН.

ddi29@mail.ru

<https://orcid.org/0000-0003-4392-2526>

Для цитирования: Дубровский Д.И. Сознание и мозг. Опыт разработки основных теоретических вопросов «Трудной проблемы сознания» и ее значение для нейронауки // Философские науки. 2024. Т. 67. № 3. С. 142–158. DOI: 10.30727/0235-1188-2024-67-3-142-158

Mind and Brain: Experience in Developing Fundamental Theoretical Issues of the Hard Problem of Consciousness and Its Significance for Neuroscience

D.I. Dubrovsky

Institute of Philosophy, Russian Academy of Sciences, Moscow, Russia

Abstract

The article presents a retrospective analysis of the author's 70-year contribution to philosophy and science, focusing primarily on the hard problem of consciousness – the relationship between mental phenomena and neurophysiological processes. The author traces his intellectual journey in addressing this fundamental challenge to natural scientific knowledge and presents his conceptual framework, initially developed in the early 1960s. The discussion examines the significant challenges faced in defending an information-based approach to consciousness, particularly during the notable debate with E.V. Ilyenkov and subsequent opposition from his followers who held considerable influence in philosophical circles. Despite ideological pressure and accusations of revisionism, the author successfully advanced his theoretical framework, publishing numerous books and articles that elaborate on his theoretical and methodological approaches to decoding neural correlates of subjective reality. Following the publication of an

English-language article outlining his theoretical foundations in a leading international neuroscience journal in 2019, the author received extensive recognition, including invitations to keynote international conferences and join editorial boards of international journals. This response demonstrates the pressing need for theoretical and methodological developments in neuroscience and related disciplines, highlighting a deficit in fundamental theoretical frameworks capable of guiding and integrating empirical research. The article concludes by emphasizing the significance of the information paradigm and its derivative approaches for studying consciousness, mental processes, and genetic aspects of biological functions.

Keywords: philosophy of mind, mind–body problem, subjective reality, information approach, genetic code, neural correlates of mental phenomena, neuroscience, neurophilosophy, artificial intelligence.

David I. Dubrovsky – D.Sc. in Philosophy, Professor, Chief Research Fellow, Department of Theory of Knowledge, Institute of Philosophy, Russian Academy of Sciences.

ddi29@mail.ru

<https://orcid.org/0000-0003-4392-2526>

For citation: Dubrovsky D.I. (2024) Mind and Brain: Experience in Developing Fundamental Theoretical Issues of the Hard Problem of Consciousness and Its Significance for Neuroscience. *Russian Journal of Philosophical Sciences* = *Filosofskie nauki*. Vol. 67, no. 3, pp. 142–158. DOI: 10.30727/0235-1188-2024-67-3-142-158

С учетом того, что мне исполнилось 95 лет, это – своего рода самоотчет (для себя и некоторых других, кому эта тема может быть интересной). Настало время подвести итоги моей работы на протяжении почти 70 лет. Вышло десять моих книг, включая «Воспоминания», пять из которых переизданы; опубликовано более 200 научных статей. Я изучал разные философские проблемы, в основном в контексте эпистемологии, исследований сознания и бессознательного, методологии науки. Но главной в моей научной жизни оставалась проблема «Сознание и мозг» (mind–brain problem), которая непосредственно связана с тем, что в западной аналитической философии часто называют «Трудной проблемой сознания» (hard problem of consciousness). Такое название вызвано фундаментальной для естественно-научного познания трудностью объяснения связи явлений сознания (субъективной реальности) с мозговыми процессами, а соответственно, их очевидной способности служить причиной воздействия на

телесные, физические процессы. Дело в том, что явлениям субъективной реальности нельзя приписывать физические свойства: массу, энергию, пространственные характеристики. Как же тогда объяснить их связь с физическими объектами и процессами? Неспособность корректного теоретического решения этой проблемы в аналитической философии характеризовали как «провал в объяснении». Несмотря на то, что эта проблема оставалась в центре внимания аналитической философии в течение последних 60 лет и ей посвящено поистине огромное число публикаций, она так и не получила убедительного решения. Это относится и к современной нейронауке, нацеленной на изучение явлений сознания, для которой указанная проблема всегда служила камнем преткновения, тормозившим ее развитие.

Меня эта проблема особенно заинтересовала еще во второй половине 50-х годов прошлого века, сразу после окончания университета. В начале 60-х мне удалось защитить кандидатскую диссертацию, посвященную ее отдельным аспектам, а затем, опубликовать ряд статей (в том числе в журнале «Вопросы философии» [Дубровский 1966; Дубровский 1968; Дубровский 1969]), в которых предлагался концептуальный подход к преодолению «провала в объяснении». В 1969 году я защитил докторскую диссертацию на тему «Философский анализ психофизиологической проблемы», которая в форме монографии выпущена издательством «Наука» по рекомендации Научного совета по кибернетике при Президиуме Академии наук [Дубровский 1971]. Эта книга, увидевшая свет более пятидесяти лет назад, недавно переиздана [Дубровский 2021a]. В ней подробно, на базе предложенного информационного подхода, представлена моя концепция разрешения главных теоретических трудностей проблемы «Сознание и мозг» и на этой основе разработаны теоретико-методологические вопросы расшифровки мозговых кодов явлений субъективной реальности. В последующие годы, после первого издания указанной книги, моя концепция уточнялась и развивалась с учетом достижений нейронауки, смежных с ней дисциплин и разработок искусственного интеллекта. Это потребовало вместе с тем специальных исследований проблемы сознания, в особенности феноменологии субъективной реальности. Полученные результаты были представлены в книгах «Информация. Сознание. Мозг» [Дубровский 1980] (второе издание: [Дубровский 2021б]) и «Проблема идеального. Субъективная реальность» [Дубровский 1983]

(второе издание: [Дубровский 2002]; третье издание: [Дубровский 2023]). Данной проблематике посвящено значительное количество моих опубликованных научных статей.

Со временем моя концепция приобрела форму четко разработанной теории. В наиболее полном и систематизированном виде она представлена в книге под названием «Проблема “Сознание и мозг”: теоретическое решение» [Дубровский 2015]. Нельзя не упомянуть и еще об одной, в какой-то мере итоговой работе [Дубровский 2019]. В ней отражен мой долгий и тернистый путь разработки «Трудной проблемы сознания»; наряду с предложенной теорией собраны мои критические статьи, опубликованные в журнале «Вопросы философии», «Философские науки» и других ведущих изданиях за последние 50 лет. Эти статьи, с одной стороны, лучше проясняют, в чем-то дополняют основные положения теории, с другой – отражают историю попыток решения «Трудной проблемы сознания» в советской и постсоветской философской литературе. Вместе с тем они свидетельствуют и о сложной, во многом неблагоприятной ситуации, в которой мне приходилось отстаивать свою теоретическую позицию.

В результате затеянной мной в 1968 году дискуссии с Э.В. Ильенковым о проблеме идеального, начатой публикацией моей статьи в «Вопросах философии» [Дубровский, 1968], я оказался в сложном положении. Да, мою позицию поддерживали многие философы и ученые. Это проявилось в большинстве статей, опубликованных на страницах журнала «Вопросы философии» в ходе указанной дискуссии. Но в то же время я встретил острое противодействие со стороны тех, кто занимал в нашей философской епархии руководящие позиции, опирался на поддержку партийных органов. Они энергично поддерживали Ильенкова, инкриминировали мне отступления от марксизма. В результате мне пришлось трижды (!) защищать докторскую диссертацию перед диссертационным советом, и еще в четвертый раз отстаивать на заседании Высшей аттестационной комиссии.

Я был ко всему прочему «невъездным» (лишен права выезда за границу, участия в международных философских мероприятиях). В 1976 году в Дюссельдорфе, на Международном философском конгрессе, была впервые организована секция «Сознание и мозг», на которую меня пригласили в качестве основного докладчика и опубликовали тезисы доклада, но меня не выпустили. В 1978 году выдающийся нейрофизиолог, президент Венгерской академии

наук Я. Сентаготаи организовал симпозиум, посвященный обсуждению моей концепции. Он был на конгрессе в Дюссельдорфе, прочитал мои тезисы, выразил поддержку основным положениям, искал меня среди участников конгресса. Затем написал мне письмо, и мы с ним несколько месяцев переписывались, обсуждали проблему и повестку симпозиума. Сентаготаи пригласил на него ряд крупных нейрочученых, в том числе своего друга, лауреата Нобелевской премии по физиологии и медицины Р. Сперри, который тоже поддерживал мою концепцию. Я был счастлив. От имени Венгерской академии наук Сентаготаи прислал в МГУ имени М.В. Ломоносова, в котором я тогда работал профессором философского факультета, официальное приглашение на симпозиум. Мне прямо не отказали, но пять месяцев «тянули», не оформляли командировку под разными предлогами. Сентаготаи трижды (!) переносил дату симпозиума. Наконец, мне стало ясно, в чем проблема. Давно подозревал, что поездка в Венгрию не состоится. Это – самая большая обида, нанесенная мне в МГУ за долгие годы работы на философском факультете.

Мне не раз приходилось испытывать чувство неполноценности, выслушивать обвинения в идеологических ошибках, отступлениях от марксизма, большей частью от сторонников Ильенкова. Именно они инициировали разгромную критику моей концепции в органе ЦК КПСС – журнале «Коммунист». В статью академика Н.П. Дубинина был вставлен с его согласия и одобрения подготовленный редакцией объемный текст, в котором уничтожающей критике подвергался информационный подход к исследованию сознания и предложенные мной разработки проблемы расшифровки мозговых кодов психических явлений [Дубинин 1980]. В ней, после цитат из моей первой книги и ряда статей, следовали такие оценки и заключения: Дубровский «в своих софистических рассуждениях, отталкиваясь от биологизации социального, соскальзывает в плоскость проблем, имеющих уже не естественно-научный, но общественно-политический аспект» [Дубинин 1980, 72]. Согласно автору статьи, провозглашая возможность расшифровки мозговых кодов психических явлений, Дубровский якобы «заявляет претензию на рекомендации с совершенно чуждых нам научных и идеологических позиций» [Дубинин 1980, 73], «игнорирует азбучные истины исторического материализма» [Дубинин 1980, 73]. И завершающий аккорд: «Тут налицо открытая ревизия марксистско-ленинской теории созна-

ния» [Дубинин 1980, 73]. В те годы такая «аттестация» в органе ЦК КПСС означала безапелляционный приговор. Прощай МГУ и право преподавать философию, даже право официально заниматься ею! Действительно, как может ревизионист марксистско-ленинской теории сознания быть профессором философского факультета МГУ?! А я в эти годы был еще и заведующим отделом диалектического материализма, логики и философских вопросов естествознания журнала «Философские науки». Я уже планировал искать работу на заводе по своей старой пролетарской специальности токаря.

Со временем узнал, что эта «военная операция» против меня была делом рук близких друзей Ильенкова, его бывших учеников, заместителя главного редактора журнала «Коммунист», доктора философских наук Л.К. Науменко и заведующего отделом науки и культуры этого журнала доктора философских наук Г.Н. Волкова. Да и главный редактор журнала Р.И. Косолапов тоже был, как известно, активным сторонником Ильенкова. Но, думаю, Ильенков лично к этому не был причастен, друзья старались. К великому моему удивлению, моих друзей и недругов, я не «вылетел» из МГУ. Меня спас главный редактор журнала «Философские науки» Владимир Спиридонович Готт, у которого, как у бывшего крупного партийного работника, были связи в высших кругах ЦК партии. Он обратился к хорошо знавшему его секретарю ЦК, все объяснил, и дело спустили на тормозах. Но поставленное на мне клеймо «ревизиониста», а значит, «неблагонадежности», еще долго давало знать о себе в моих социальных и философских коммуникациях. Чуткие к партийным оценкам, да еще на таком уровне, как журнал «Коммунист», многие ведущие и рядовые философы, услышав слова «субъективная реальность» и «расшифровка кодов», старались обходить меня и мои работы стороной: «как бы чего не вышло». После статьи в «Коммунисте» указанное направление актуальных исследований мозга долго находилось под идеологическим подозрением, что относилось и к нейрофизиологическим исследованиям Н.П. Бехтеревой, с которой я в эти годы тесно сотрудничал, хотя ее имя в «Коммунисте» не упоминалось. Это продолжалось до начала периода перестройки, когда все идеологические препоны были сняты.

Но до сих пор клан сторонников Ильенкова сохраняет организованность и активность, они превозносят его «победу» в нашей дискуссии и по-прежнему называют меня «биологизатором»,

«вульгарным материалистом», «софистом» и еще хлеще. Особенно интересен тот факт, что в качестве решающего доказательства «победы» Ильенкова в дискуссии они приводят фальсифицированные им результаты знаменитого т.н. Загорского эксперимента со слепоглухими, у которых якобы от рождения не только целиком отсутствовали зрение и слух, но вообще не было психики, и она, как заявлял Ильенков, была сформирована, создана воспитателями «с нуля» под его руководством (см., например: [Ильенковские чтения... 2016]). В действительности, однако, они утратили зрение и слух в сравнительно позднем возрасте, когда у них уже было сформировано сознание и владение языком. Более того, у двух из них сохранились существенные остатки зрения, а у двух остатки слуха. Это позволяло Александру Суворову и Юрию Лернеру самостоятельно ездить в общественном транспорте не только по Москве, но и из Москвы в Загорск и обратно, что вызывало возмущение и резкий запрет Ильенкова, т.к. опровергало его трактовку Загорского эксперимента.

Здесь будет уместно привести основные положения Ильенкова, четко изложенные им в его статье, опубликованной в журнале «Коммунист». В ней он следующим образом характеризует своих подопечных: «А ведь они как были, так и остались физически слепоглухими и если бы не специально выработанная наукой система воспитания, были бы обречены на бессознательное существование в мире мрака и безмолвия, и физического и духовного, и в прямом и в переносном смысле сих страшных слов... Исходное условие жесткое: психики нет вообще, и “сама” она не возникает. Ее надо “сделать”, сформировать воспитать... Исходное условие – то, что дано природой, биологией. Ничтожно мало – одни лишь простейшие органические нужды: в пище, воде, да физических факторах известного диапазона. Больше ничего» [Ильенков 1977, 70–71]. И вот, благодаря методам воспитания, основанным на марксистской теории личности, они обрели психику, стали полноценными членами общества. Более того, благодаря научной программе, разработанной Институтом дефектологии и усилиям его сотрудников, они смогли окончить психологический факультет МГУ. Все это изображалось как триумф марксистской теории формирования личности, исключающей напрочь роль генетических факторов, ибо все определяется сугубо социальными факторами, воспитанием, ролью педагога, вооруженного передовой марксистской теорией. Как видим, концепция Ильенкова с его фальсификацией

Загорского эксперимента была санкционирована в органе ЦК КПСС журнале «Коммунист» и обрела идеологическое значение. Соответственно, те, кто пытались резко критиковать эту концепцию с научных позиций, обвинялись в отступлениях от марксизма, что было показано выше на примере оценки в журнале «Коммунист» моей позиции. Этим объясняется и то обстоятельство, что фальсификация Ильенкова получила поддержку со стороны весьма авторитетных тогда ученых: А.Н. Леонтьева, В.В. Давыдова, Н.П. Дубинина, которые хорошо знали реальное положение дел, но умели подлаживаться под партийные установки и быть всегда на гребне текущей конъюнктуры

Фальсификация Ильенкова была основательно разоблачена в период перестройки, на организованной мной по линии Philosophического общества конференции, специально посвященной проблеме слепоглухонемых. В ней участвовали ведущие сотрудники Института дефектологии и его директор, В.И. Лубовский, директор Института общей и педагогической психологии А.М. Матюшкин, главный редактор «Психологического журнала» А.В. Брушлинский, ряд известных психологов и философов, а с основным докладом выступил один из четырех знаменитых слепоглухих С.А. Сироткин (материалы этой конференции опубликованы дважды [Слепоглухонемота... 1989; Слепоглухонемота... 2018]). В своем докладе Сироткин прямо говорил об этих фальсификациях Ильенковым, более того – о том, что он принуждал своих подопечных поддерживать их, скрывать свои реальные способности. По поводу широковещательных заявлений Ильенкова о том, что до Загорского эксперимента у его четверых учеников «не было ни психики, ни самосознания», Сироткин заметил: «В концепции Ильенкова слепоглухой предстает *“табула rasa”*, на которой пишет всесильная рука педагога» [Слепоглухонемота... 1989, 93]. Приведя многие факты ильенковских фальсификаций «из благих побуждений», Сироткин заявляет: «К сожалению, “святая ложь”, “ложь во спасение” неизбежно перерастает в настоящую ложь со всеми негативными для слепоглухих последствиями» [Слепоглухонемота... 1989, 102]. Ильенков не только переиначивал факты, скрывал правду, но и активно боролся за сохранение этой неправды. Сироткин признается: «У нас с Эвальдом Васильевичем был тяжелый конфликт, и это свидетельствует о том, что он сознательно скрывал определенные факты во имя чистоты эксперимента» [Слепоглухонемота... 1989, 26].

Я остановился на этом подробно, т.к. в последние годы опубликовано около двух десятков статей, посвященных моей дискуссии с Ильенковым, в т.ч. и в «Вопросах философии». Большинство из них написаны его известными сторонниками. В них превратно изложена моя позиция, особенно о проблеме «Сознание и мозг» и концепции субъективной реальности. Это относится и к многочисленным материалам, опубликованным ими в интернете. Разумеется, все это сегодня для меня не является существенным. Но оно показывает живучесть, сохранившуюся инерцию гегельянско-марксоидных и радикально-социологизаторских установок мышления, характерных для многих представителей советской философии при объяснении сознания, категорического отрицания роли генетических факторов в формировании личности и в процессах общественной жизни.

В первые десятилетия нового века я продолжал развивать концепцию субъективной реальности в аспекте разработки ее феноменологии, много занимался изучением философских и теоретико-методологических вопросов в контексте темы искусственного интеллекта, был совместно с академиками В.А. Лекторским и В.Л. Макаровым организатором Научного совета РАН по методологии искусственного интеллекта и когнитивных исследований и на протяжении 18 лет работал на общественных началах в качестве его сопредседателя. Эта деятельность имела большое значение для развития мной информационного подхода к проблеме «Сознание и мозг» с учетом достижений в области искусственного интеллекта и робототехники.

Теперь перейду к тем вопросам последних лет моей научной деятельности, которые побудили меня написать столь затянувшийся отчет. Речь пойдет о главном из того, что сделано в моей долгой научной деятельности, – предложенном мной теоретическом решении основных вопросов «Трудной проблемы сознания» – и о недавних, очень удививших меня событиях, связанных с его восприятием в западных нейронаучных изданиях. Несмотря на множество моих опубликованных статей, указанных выше книг, изданных в контексте этой тематики (за 60 лет) на русском языке, и на то обстоятельство, что она оставалась высокоактуальной в западной философской и научной литературе, я не предпринимал намерений переводить свои работы на английский и публиковать их на Западе. Упомяну лишь об одном случае. Во второй половине 80-х годов я сотрудничал с известным американским

философом Дж. Марголисом, который разрабатывал концепцию эмерджентистского материализма, близкую к диалектическому материализму. По его инициативе на английский язык переведена моя ранее опубликованная статья, посвященная расшифровке кодов как общенаучной проблеме [Дубровский 1979]. Эта статья с предисловием Дж. Марголиса опубликована в американском журнале [Margolis 1987]. Кроме того, в 1988 году, в условиях начавшейся перестройки, издательство «Прогресс», выпускавшее иностранную литературу, перевело по своей инициативе на английский язык и выпустило мою книгу «The Problem of the Ideal: The Nature of Mind and Its Relationship to the Brain and Social Medium» [Dubrovsky 1988]. Она состояла из двух частей: одна из них представляла несколько глав из моей книги «Проблема идеального», вторая – из глав книги «Информация. Сознание. Мозг». На нее обратил внимание выдающийся нейрофизиолог, лауреат Нобелевской премии Р. Сперри, который поддержал мою концепцию.

После этого, на протяжении более 30 лет, публикаций на английском у меня не было. Но в 2017 году на страницах нашего ведущего нейронаучного журнала была опубликована моя статья «Проблема свободы воли и современная нейронаука» [Дубровский 2017]. Ее без моего ведома перевели на английский и опубликовали [Dubrovsky 2019a]. Затем я получил более десятка предложений из международных нейронаучных журналов написать для них статью на близкую тему. И я подумал о том, что все-таки стоит, наверное, опубликовать на английском статью с изложением моей теории. Долго собирался и в итоге написал такую статью объемом полтора печатных листа. В ней четко, по пунктам изложил основное содержание теории, перевел ее на английский с помощью знакомого, в совершенстве владеющего этим языком, и отправил в журнал, который дважды обращался ко мне с такой просьбой. Он оказался одним из главных международных журналов, специализирующихся на проблемах нейронауки. Речь идет о журнале «AIMS Neuroscience» (индексируется WoS, Scopus, PMC) [Dubrovsky 2019b]. Статью сравнительно быстро опубликовали, и, к моему огромному удивлению, на меня обрушился настоящий шквал приглашений, откликов, положительных оценок из редакций других журналов, различных организаций и от ряда авторов. До сих пор я получаю такие письма по несколько раз в неделю. Но никому не отвечаю, нет времени и сил. Тем более что у меня

нет больше намерений публиковаться на английском. Сначала я из интереса пытался коллекционировать эти приглашения, но вскоре оставил это занятие. Большая часть всей этой огромной почты на всякий случай у меня, однако, сохранилась. После выхода указанной статьи в 2019 году я получил с тех пор около 500 (!) откликов подобного рода. Среди них – 26 (!) приглашений за это время выступить спикером на международных конгрессах и конференциях в Париже, Лондоне, Риме, Амстердаме, Праге, Вене, Барселоне, Бостоне, Токио, Осаке, Дубае, Сингапуре, Бангкоке, Дублине и т.д. Получил более 20 приглашений войти в состав редколлегий международных журналов. Но, понятно, в моем возрасте и при моей занятости нет ни малейшего смысла принимать эти приглашения. Чтобы не быть голословным, я сделал сравнительно небольшую подборку полученных писем с приглашениями, при желании ее можно просмотреть на моем сайте¹.

Хочу, чтобы мне поверили: я написал об этом не из самовосхваления, хотя не отрицаю, что такое признание моей работы доставило мне, конечно, радостное чувство. Тем более с учетом того, что несмотря на «солидный возраст» моей теории и поддержку со стороны ряда крупных представителей нейронауки (и даже включения ее в два учебника по нейрофизиологии), за 35 лет работы в Институте философии значение ее никто и никогда не отмечал. Два или три раза за эти годы коллеги высказывали в ее адрес краткие и поверхностные замечания, из которых становилось понятным, что они не читали внимательно моих публикаций, посвященных данной теме. Каждому из нас нужна какая-то поддержка научного сообщества, иначе временами охватывает чувство, что твой многолетний труд является напрасным. Такое чувство мне пришлось испытывать слишком часто, и это, наверное, связано с тем, что я был в какой-то мере маргиналом, не принадлежал к влиятельным группировкам и кланам, представители которых постоянно цитировали друг друга, «накручивая» свой авторитет, и в первую очередь – своего лидера. Множество

¹ С подборкой можно ознакомиться на моем сайте <http://dubrovsky.info>, выбрав в меню раздела «Научные публикации» рубрику «Научная деятельность» и подрубрику «Психические явления и мозг. Решение основных теоретических вопросов. “Трудная проблема сознания”». На открывшейся странице можно прочитать первую в указанном перечне материалов работу «Сознание и мозг. Отчет о результатах разработки основных теоретических вопросов “Трудной проблемы сознания”». В конце текста – «Приложения. Подборка писем».

откликов на мою статью, о которых говорилось выше, стали для меня несомненной поддержкой.

Достаточно хотя бы бегло просмотреть приводимую выборку, чтобы убедиться в том, что она связана не только с моей персоной, но и с особенно актуальным вопросам развития современной нейронауки и смежных с нею дисциплин. Обратим внимание на то, что просьбы выступить спикером на международных конференциях, посвященных философским, теоретическим и методологическим вопросам нейронауки и написать статью об этой проблематике, относились к самым разным специализациям нейрофизиологии: от нейропсихологии до нейроиммунологии и нейрофармакологии, от нейрохирургии, нейротерапии и неврологии до нейрокибернетики. Возрастающий колоссальный массив эмпирических исследований настоятельно требует основательного методологического осмысления.

Острый дефицит теоретических разработок создает сильный тормоз в развитии нейронауки и решении ее междисциплинарных проблем, особенно связанных с исследованиями сознания, развитием искусственного интеллекта и робототехники. Философы, специализирующиеся на вопросах эпистемологии и методологии науки, становятся особенно востребованными (разумеется, при наличии достаточной у них компетенции в соответствующей научной проблематике).

Понятен и большой интерес, проявленный нейроучеными к разработке «Трудной проблемы сознания», поскольку в центре их внимания в итоге всегда находится человек, психосоматическое единство его жизнедеятельности. Поэтому особенно важно для нее понимание роли сознания в организации и реализации жизненных функций социального индивида, его произвольных действий и способов эффективной саморегуляции, изучение механизмов воздействия явлений сознания на физиологические процессы, в первую очередь способов их связи с мозговой нейродинамикой.

Но для нейроученого данная тематика не ограничивается только сознанием. Вопрос ставится шире: о роли психической деятельности в единстве ее сознательных и бессознательных процессов, с акцентом на исследовании бессознательных форм психической регуляции жизненных функций, различных ее уровней, от «близких» к сознанию (и допускающих в тех или иных пределах сознательное управление ими, например в слу-

чаях памяти) до глубоких уровней бессознательно-психического, функционирующих с высокой степенью автономии от сознания. Впрочем, соотношение разных уровней бессознательных психических процессов и выяснение их связей с сознанием требуют дальнейшего основательного исследования. Очевидным становится одно: всякое явление сознания имеет под собой, в качестве своей основы, сложную структуру бессознательных процессов; между ними и уровнем сознания существуют прямые и обратные, непосредственные и чаще опосредствованные связи, которые изучены недостаточно. И это обращает внимание нейрочеловека к еще более глубокому, *допсихическому* уровню жизненных процессов.

Поэтому, наряду с уровнями сознания и психического (как единства сознательного и бессознательного), необходимо выделить уровень информационных процессов, который включает в себя сознательное и психическое, но вместе с тем и свои особые способы организации, регуляции, управления жизненными процессами, прежде всего генетического характера. В ходе сотен миллионов лет эволюции выработаны надежные информационные компетенции организма в управлении его чрезвычайно сложными системами. Я обычно привожу такой пример: в нашем организме смонтировано сто тысяч километров (!) сосудистых русел, и эта колоссально сложная динамическая система эффективно функционирует в условиях непрерывно изменяющейся внешней среды и внутренней среды организма. Пока мы лишь в начале пути к пониманию и объяснению этой чрезвычайной сложности. Учитывая, что информационный уровень является фундаментальным в изучении жизнедеятельности, можно считать информационную парадигму и основанные на ней *информационные подходы* к изучению сознания, психического и генетического теоретически адекватными для разработки многоплановых проблем современной нейронауки и тесно связанных с ней областей научного знания. Возможно, отчасти поэтому предложенное мной на основе информационного подхода решение главных теоретических вопросов «Трудной проблемы сознания» вызвало неожиданный для меня резонанс в нейронаучном сообществе.

ЦИТИРУЕМАЯ ЛИТЕРАТУРА

Дубинин 1980 – Дубинин Н.П. Наследование биологическое и социальное // Коммунист. 1980. № 11. С. 62–74.

Дубровский 1966 – Дубровский Д.И. Физиологическое и логическое // Вопросы философии. 1966. № 8.

Дубровский 1968 – Дубровский Д.И. Мозг и психика (О необоснованности философского отрицания психофизиологической проблемы) // Вопросы философии. 1968. № 8. С. 122–135.

Дубровский 1969 – Дубровский Д.И. По поводу статьи Э.В. Ильенкова «Психика и мозг» // Вопросы философии. 1969. № 3. С. 142–146.

Дубровский 1971 – Дубровский Д.И. Психические явления и мозг. Философский анализ психофизиологической проблемы в связи с некоторыми актуальными задачами нейрофизиологии, психологии и кибернетики. – М.: Наука, 1971 .

Дубровский 1979 – Дубровский Д.И. Расшифровка кодов (методологические аспекты проблемы) // Вопросы философии. 1979. № 12. С. 87–100.

Дубровский 1980 – Дубровский Д.И. Информация. Сознание. Мозг. – М.: Высшая школа, 1980.

Дубровский 1983 – Дубровский Д.И. Проблема идеального. Субъективная реальность. – М.: Мысль, 1983.

Дубровский 2002 – Дубровский Д.И. Проблема идеального. Субъективная реальность / 2-е изд. – М.: Канон+, 2002.

Дубровский 2015 – Дубровский Д.И. Проблема «Сознание и мозг»: теоретическое решение. – М.: Канон+, 2015.

Дубровский 2017 – Дубровский Д.И. Проблема свободы воли и современная нейронаука // Журнал высшей нервной деятельности. 2017. Т. 67. № 6. С. 739–754.

Дубровский 2019 – Дубровский Д.И. Проблема сознания: теория и критика альтернативных концепций. – М.: Ленанд, 2019.

Дубровский 2021a – Дубровский Д.И. Психические явления и мозг. Философский анализ проблемы в связи с некоторыми актуальными задачами нейрофизиологии, психологии и кибернетики / 2-е изд., доп. – М.: URSS, 2021.

Дубровский 2021б – Дубровский Д.И. Информация. Сознание. Мозг. Расшифровка мозговых кодов психических явлений / 2-е изд., доп. – М.: URSS, 2021.

Дубровский 2023 – Дубровский Д.И. Проблема идеального. Субъективная реальность / 3-е изд. – М.: ЛЕНАНД, 2023.

Ильенков 1977 – Ильенков Э.В. Становление личности: к итогам научного эксперимента // Коммунист. 1977. № 2. С. 68–79.

Ильенковские чтения... 2016 – Ильенковские чтения. Материалы XVIII Международной научной конференции. Белгород 28–29 апреля 2016 г. / под ред. А.Д. Майданского. – Белгород: Издательский дом «Белгород», 2016.

Слепоглухонемота... 1989 – Слепоглухонемота: исторические и методологические аспекты. Мифы и реальность / отв. ред. Д.И. Дубровский. – М.: Philosophical society, 1989.

Слепоглухонемота... 2018 – Слепоглухонемота: исторические и методологические аспекты. Мифы и реальность / 2-е изд., доп. – М.: ИИнтелл, 2018.

Dubrovsky 1988 – *Dubrovsky D.I. The Problem of the Ideal: The Nature of Mind and Its Relationship to the Brain and Social Medium.* – Moscow: Progress Publishers, 1988.

Dubrovsky 2019a – *Dubrovsky D.I. The Problem of Free Will and Modern Neuroscience // Neuroscience and Behavioral Physiology.* 2019. Vol. 49. No. 5. P. 629–639.

Dubrovsky 2019b – *Dubrovsky D.I. “The Hard Problem of Consciousness”: Theoretical Solution of its Main Questions // AIMS Neuroscience.* Vol. 6. No. 2. P. 85–103.

Margolis 1987 – *Margolis J. Introductory Note to Dubrovskii: Dubrovskii D.I. The Decipherment of Codes: Methodological Aspects of the Problem // Journal for The Theory of Social Behaviour.* 1987. Vol. 17. No 1. P. 1–18.

REFERENCES

Dubin N.P. (1980) Biological and Social Inheritance. *Kommunist.* No. 11, pp. 62–74 (in Russian).

Dubrovsky D.I. (1966) Physiological and Logical. *Voprosy filosofii.* No. 8 (in Russian).

Dubrovsky D.I. (1968) Brain and Psyche (On the Unjustified Philosophical Denial of the Psychophysiological Problem). *Voprosy filosofii.* No. 8, pp. 122–135 (in Russian).

Dubrovsky D.I. (1969) Regarding E.V. Ilyenkov’s Article “Psyche and Brain”. *Voprosy filosofii.* No. 3, pp. 142–146 (in Russian).

Dubrovsky D.I. (1971) *Mental Phenomena and the Brain. Philosophical Analysis of the Psychophysiological Problem in Relation to Some Current Issues in Neurophysiology, Psychology, and Cybernetics.* Moscow: Nauka (in Russian).

Dubrovsky D.I. (1979) *Decoding the Codes (Methodological Aspects of the Problem).* *Voprosy filosofii.* No. 12, pp. 87–100 (in Russian).

Dubrovsky D.I. (1980) *Information. Consciousness. Brain.* Moscow: Vysshaya shkola (in Russian).

Dubrovsky D.I. (1983) *The Problem of the Ideal. Subjective Reality.* Moscow: Mysl’ (in Russian).

Dubrovsky D.I. (1988) *The Problem of the Ideal: The Nature of Mind and Its Relationship to the Brain and Social Medium.* Moscow: Progress Publishers.

Dubrovsky D.I. (Ed.) (1989) *Deafblindness: Historical and Methodological Aspects. Myths and Reality*. Moscow: Filosofskoe obshchestvo (in Russian).

Dubrovsky D.I. (2002) *The Problem of the Ideal. Subjective Reality* (2nd ed.). Moscow: Kanon+ (in Russian).

Dubrovsky D.I. (2015) *The Problem of Mind and Brain: A Theoretical Solution*. Moscow: Kanon+ (in Russian).

Dubrovsky D.I. (2017) The Problem of Free Will and Modern Neuroscience. *Pavlov Journal of Higher Nervous Activity*. Vol. 67, no. 6, pp. 739–754 (in Russian).

Dubrovsky D.I. (Ed.) (2018) *Deafblindness: Historical and Methodological Aspects. Myths and Reality* (2nd ed.). Moscow: Intell (in Russian).

Dubrovsky D.I. (2019a) *The Problem of Consciousness: Theory and Critique of Alternative Concepts*. Moscow: Lenand (in Russian).

Dubrovsky D.I. (2019b) The Problem of Free Will and Modern Neuroscience. *Neuroscience and Behavioral Physiology*. Vol. 49, no. 5, pp. 629–639.

Dubrovsky D.I. (2019c) “The Hard Problem of Consciousness”: Theoretical Solution of its Main Questions. *AIMS Neuroscience*. Vol. 6, no. 2, pp. 85–103.

Dubrovsky D.I. (2021a) *Mental Phenomena and the Brain. Philosophical Analysis of the Problem in Connection with Some Current Issues in Neurophysiology, Psychology, and Cybernetics* (2nd ed.). Moscow: URSS (in Russian).

Dubrovsky D.I. (2021b) *Information. Consciousness. Brain. Decoding the Brain Codes of Mental Phenomena* (2nd ed.). Moscow: URSS (in Russian).

Dubrovsky D.I. (2023) *The Problem of the Ideal. Subjective Reality* (3rd ed.). Moscow: Lenand (in Russian).

Ilyenkov E.V. (1977) The Formation of Personality: To the Results of Scientific Experiment. *Kommunist*. No. 2, pp. 68–79 (in Russian).

Maidanskiy A.D. (Ed.) (2016) *Ilyenkov Readings: Materials of the 18th International Scientific Conference, Belgorod, April 28–29, 2016*. Belgorod: Belgorod Publishing (in Russian).

Margolis J. (1987) Introductory Note to Dubrovskii: Dubrovskii D.I. “The Decipherment of Codes: Methodological Aspects of the Problem”. *Journal for The Theory of Social Behaviour*. Vol. 17, no. 1, pp. 1–18.

ДАВИДУ ИЗРАИЛЕВИЧУ ДУБРОВСКОМУ – 95!

Известному российскому философу,
доктору философских наук, профессору,
главному редактору журнала «Философские науки»,
главному научному сотруднику Института философии РАН,
профессору философского факультета МГУ имени М.В. Ломоносова,
заместителю председателя Научного Совета по методологии
искусственного интеллекта и когнитивных исследований
при Президиуме Российской академии наук,
участнику Великой Отечественной войны
Давиду Израилевичу Дубровскому исполнилось 95 лет!

Сохраняющиеся на протяжении многих десятилетий интерес к научным
трудам Давида Израилевича у специалистов как в нашей стране, так и
далеко за ее пределами и неизменная привлекательность со стороны
многочисленных студентов объясняются,
в первую очередь, оригинального его мышления
и высокими человеческими качествами.

Обширная сфера профессиональных интересов,
в которой Давиду Израилевичу
удалось обозначить и утвердить
свое уникальное слово на протяжении
всего жизненного интеллектуального пути, а именно –
теория познания и методология науки; проблема сознания;
проблема идеального;
методологии искусственного интеллекта и когнитивных исследований;
проблема «Сознание и мозг»; феноменологический анализ явлений
субъективной реальности; проблемы развития искусственного
интеллекта, – оставалась чрезвычайно актуальной
на протяжении последних десятилетий
и продолжает быть крайне актуальной
по настоящее время.

Редколлегия и редакция журнала, многочисленные авторы и актив
журнала от всей души поздравляют юбиляра со знаковой датой,
желают ему неумной энергии, крепости духа восточного единоборца,
безупречного здоровья, новых творческих успехов
и сохранения активной жизненной позиции еще многие-много лет!

*Шеф-редактор журнала «Философские науки»,
Президент Академии гуманитарных исследований
Х.Э. Мариносян*

Цели и задачи

«Философские науки» – ежемесячный рецензируемый научный журнал, публикующий статьи на русском и английском языках. Журнал был учрежден в 1958 г. как научное образовательное просветительское издание для системы отечественного образования. Статьи журнала посвящены как традиционным, классическим философским темам, так и актуальным проблемам современности и перспективам социокультурного и цивилизационного развития человечества.

Aims and Scope

Russian Journal of Philosophical Sciences = Filosofskie nauki is a peer-reviewed monthly journal published in Russian and English languages. The journal was founded in 1958 as a scientific and educational periodical for the educational system. The papers of the journal are dedicated to classical problems of philosophy as well as to important issues of the modernity and prospects of sociocultural and civilizational development of humankind.

С условиями публикации материалов в журнале
и порядком рецензирования статей можно ознакомиться на сайте:
<https://www.phisci.info/jour/about/submissions#authorGuidelines>

Подписной индекс в Объединенном каталоге «Пресса Россия» – 45490

Журнал зарегистрирован в Министерстве РФ по делам печати, телерадиовещания и средств массовых коммуникаций. Свидетельство о регистрации ПИ № 77 – 15513 от 20 мая 2003 г.

Подписано в печать 15.08.2024 г. Формат 60х90/16. Печать цифровая.

Бумага офсетная № 1. Печ. л. 10,0. Тираж 1000 экз. Заказ

Отпечатано в типографии Издательского дома «Гуманитарий».

E-mail: humanist@academyrh.info