ISSN 0235-1188 (Print)
ISSN 2618-8961 (Online)

65(1)` 2022

ФИЛОСОФСКИЕ НАУКИ

ФН

RUSSIAN JOURNAL OF PHILOSOPHICAL SCIENCES

- *ЧТЕНИЕ МОЗГА!*
- *ФЕТИШ XXI ВЕКА*
- *ПСЕВДОНАУЧНОЕ КЛИШЕ*
- *НЕКОМПЛЕМЕНТАРНЫЕ СТРУКТУРЫ*
- *АУТОПОЙЕТИЧЕСКИЕ РАЗРЫВЫ*
- *БЛОКИРОВКА ЭМПАТИИ*
- *СОЦИОГУМАНИТАРНАЯ ЗНАЧИМОСТЬ*

МОСКВА
ГУМАНИТАРИЙ

MOSCOW
HUMANIST PUBLISHING HOUSE

МИНИСТЕРСТВО НАУКИ
И ВЫСШЕГО ОБРАЗОВАНИЯ
РОССИЙСКОЙ ФЕДЕРАЦИИ

АКАДЕМИЯ
ГУМАНИТАРНЫХ
ИССЛЕДОВАНИЙ

ФН

ФИЛОСОФСКИЕ НАУКИ

Том 65. № 1/2022

Научный образовательный просветительский журнал
Издается с 1958 года. Выходит ежемесячно

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

E-mail: academyRH@list.ru

http://www.phisci.info

Шеф-редактор Мариносян Х.Э.

ISSN 0235-1188 (Print)

ISSN 2618-8961 (Online)

THE MINISTRY OF SCIENCE
AND HIGHER EDUCATION OF
THE RUSSIAN FEDERATION

ACADEMY FOR
RESEARCH INTO
THE HUMANITIES

ΦН

RUSSIAN JOURNAL OF PHILOSOPHICAL SCIENCES (FILOSOFSKIE NAUKI)

Vol. 65. No. 1/2022

Scientific and Educational Journal
Published since the Year 1958. Issued Monthly

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

ISSN 0235-1188 (Print)
ISSN 2618-8961 (Online)

ОГЛАВЛЕНИЕ

ПЕРСПЕКТИВЫ ЧЕЛОВЕЧЕСТВА. ФИЛОСОФИЯ ГУМАНИТАРНО-ТЕХНОЛОГИЧЕСКОГО РАЗВИТИЯ

CONTENTS

PROSPECTS FOR MANKIND. PHILOSOPHY OF HUMANITARIAN AND TECHNOLOGICAL DEVELOPMENT

ПЕРСПЕКТИВЫ ЧЕЛОВЕЧЕСТВА. ФИЛОСОФИЯ ГУМАНИТАРНО-ТЕХНОЛОГИЧЕСКОГО РАЗВИТИЯ

Введение в рубрику
Section introduction

Уважаемые читатели!

Специальный выпуск журнала «Философские науки» посвящен одной из наиболее актуальных проблем современности – **определению и анализу социогуманитарных оснований критериев оценки инноваций, использующих цифровые технологии и искусственный интеллект.**

В настоящее время нарастает бум цифровых трансформаций, внедрения искусственного интеллекта (ИИ) во все сферы жизнедеятельности и развития социальных систем. Как правило, уделяется недостаточно внимания оценке социальных последствий от такого рода инноваций. На это все чаще указывают не только ученые, но и представители международных организаций, государственных структур. Общество занимает все более активную позицию в связи с неконтролируемыми последствиями от внедрения цифровых технологий и прежде всего в сферу образования.

Базовые причины связаны с доминированием западной модели техногенной цивилизации, воплощением которой является технократический подход к цифровой трансформации социальных систем. Отечественные философы предупреждали о кризисе модели техногенной цивилизации и негативных последствиях ее доминирования (В.С. Степин) и предлагали философско-методологические основания для разработки модели посттехногенной цивилизации на основе постнеклассической научной рациональности. Актуальность такого рода исследований и разработок резко повышается с учетом нарастающих процессов цифровой трансформации.

Выделим наиболее актуальные философско-методологические проблемы совершенствования процессов разработки, использования цифровых технологий и ИИ.

Во-первых, исследовательский интерес представляет проблема «онтологического парадокса» в разработках и внедрении цифровых технологий и ИИ. Суть этого парадокса состоит в разрыве парадигм ИИ и парадигм социальных систем. С одной стороны, категориальный аппарат и базовые понятия из парадигм и онтологий социальных систем пытаются использовать в парадигмах и онтологиях систем ИИ («этика ИИ», «доверие ИИ» и др.), фактически некорректно «очеловечивая» ИИ, наделяя его субъектностью. С другой стороны, на основе представлений парадигм ИИ внедряют в социальные системы наработки ИИ без учета специфики парадигм социальных систем. Для преодоления онтологического парадокса предлагается использовать постнеклассическую модель саморазвивающихся полисубъектных сред – кибернетику третьего порядка (В.Е. Лепский).

Во-вторых, важной видится проблема учета «цивилизационной специфики» заказчиков и разработчиков цифровых технологий и ИИ в определении базовых понятий, категориального аппарата и критериев оценки такого рода инноваций. Например, принципиально отличаются подходы к ИИ специалистов из США и Европы по сравнению с подходами специалистов из Китая. У первых доминирует либеральный подход с ярко выраженным превалированием интересов индивидов над общественными интересами, у вторых – ориентация на коллективизм с «жестким» контролем над индивидами. Оба подхода вступают в определенное противоречие со спецификой подходов, характерных для российской цивилизации. Как следствие, уместна постановка проблемы поиска целесообразного подхода с учетом опыта развития российской цивилизации. Актуальность этой проблемы резко повышается в условиях сложившейся международной обстановки (Д.И. Дубровский).

В-третьих, значима проблема определения категориального аппарата, особенно критериев оценки инноваций, использующих цифровые технологии и ИИ в конкретных сферах жизнедеятельности и развития социальных систем. Данная проблема зависит от теоретико-рефлексивных позиций, которые актуализируются для ее решения. Среди них актуальны

позиции субъектов стратегических проектов, государств и их объединений, общих научных подходов, а также подходов, существующих в прикладных областях знания и практических сферах деятельности. Эти позиции влияют друг на друга в зависимости от их сочетания и сложившихся ситуаций. Как следствие, важным шагом для решения этой проблемы является разработка методологической платформы оценки инноваций в форме специализированного универсального конфигуратора взаимосвязанных позиций, социогуманитарного анализа инноваций и выработки критериев их оценки. Это позволит создать инструментарий, инвариантный к различным сферам применения цифровых технологий и ИИ. Принципиально важно отметить, что создаваемая методологическая платформа может рассматриваться как научная парадигма. Следовательно, она должна удовлетворять сложившимся в науке критериям становления новых научных парадигм, в частности сформировавшемуся в естественных науках принципу соответствия Н. Бора и общенаучным идеям становления новых парадигм Т. Куна.

В специальном выпуске журнала представлены результаты исследований участников научного коллектива по теме гранта Российского научного фонда, проект № 21-18-00184 «Социогуманитарные основания критериев оценки инноваций, использующих цифровые технологии и искусственный интеллект». Проведен предварительный анализ и представлены рекомендации для корректной постановки и решения выделенных трех актуальных философско-методологических проблем совершенствования процессов разработки, использования цифровых технологий и ИИ.

В.Е. ЛЕПСКИЙ
Научный руководитель проекта

Д.И. ДУБРОВСКИЙ
Ответственный координатор проекта

Философия искусственного интеллекта

DOI: 10.30727/0235-1188-2022-65-1-10-26
Оригинальная исследовательская статья
Original research article

Эпистемологический анализ социогуманитарной значимости новаций искусственного интеллекта в контексте общего искусственного интеллекта*

Д.И. Дубровский
Институт философии РАН Москва, Россия

Аннотация

В последние годы возникли новые направления развития искусственного интеллекта (ИИ), поставлена задача создания общего искусственного интеллекта (ОИИ), который способен выйти за пределы «узкого» ИИ, обрести высокую степень автономности, самостоятельного решения задач в разных условиях внешней среды и таким образом иметь возможность выполнять функции естественного интеллекта. В связи с этим возникают важные философские и теоретико-методологические вопросы, касающиеся определения и оценки социальной значимости новых достижений ИИ, особенно под углом соотношения их социогуманитарных и технологических аспектов. Необходимо преодолеть сугубо технократический подход, который обычно игнорирует отрицательные последствия цифровизации, возможные риски и угрозы развития ИИ. Этому способствует парадигмальный разрыв между системой понятий, используемых для описания технологий, и системами понятий, специфичных для социогуманитарных описаний и объяснений. Возникает так называемый онтологический парадокс при цифровой трансформации и внедрении ИИ в социальные системы. Онтологический парадокс является также и эпистемологическим парадоксом, поскольку всякое утверждение онтологического типа предполагает его эпистемологическое обоснование. Для преодоления указанного разрыва успешно используется предложенная В.Е. Лепским концепция саморазвивающихся полисубъектных сред – кибернетики третьего порядка. Она позволяет создать концептуальный «мост» между двумя системами понятий, не имеющих между собой прямых

* Работа поддержана Российским научным фондом (РНФ), грант № 21-18-00184 «Социогуманитарные основания критериев оценки инноваций, использующих цифровые технологии и искусственный интеллект».

логических связей. Продуктивным инструментом для этого может служить информационный подход, широко используемый для решения подобных междисциплинарных проблем. Он способен теоретически корректно связать технологические и социогуманитарные описания в единой концептуальной структуре и использоваться для анализа выделенных В.Е. Лепским основных социогуманитарных критериев оценки цифровых новаций. В статье подробно рассматриваются основные положения информационного подхода и его применения для разработки систем ОИИ, социогуманитарной оценки автономного развития ИИ и решения вопросов безопасности.

Ключевые слова: «узкий» искусственный интеллект, социогуманитарные критерии, онтологический парадокс, эпистемологический парадокс, саморазвивающиеся полисубъектные среды, информационный подход, безопасность.

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.* Эпистемологический анализ социогуманитарной значимости новаций искусственного интеллекта в контексте общего искусственного интеллекта // Философские науки. 2022. Т. 65. № 1. С. 10–26. DOI: 10.30727/0235-1188-2022-65-1-10-26

An Epistemological Analysis of the Social and Humanitarian Significance of Artificial Intelligence Innovations in Context of Artificial General Intelligence*

D.I. Dubrovsky
Institute of Philosophy, Russian Academy of Sciences, Moscow, Russia

Abstract

Nowadays, new directions for the development of artificial intelligence (AI) have emerged, the task has been set to develop artificial general intelligence (AGI), which is able to go beyond the narrow AI, gain a high degree of autonomy, independently solve problems in different environmental conditions and thus have the ability to perform the functions of natural

* The work was supported by the Russian Science Foundation, grant no. 21-18-00184 “Social and humanitarian foundations for evaluation criteria for innovations based on digital technologies and artificial intelligence.”

intelligence. In this regard, important philosophical, theoretical, and methodological questions arise concerning the definition and evaluation of the social significance of new AI achievements, especially regarding the correlation of their socio-humanitarian and technological aspects. It is necessary to overcome a purely technocratic approach, which usually ignores the negative consequences of digitalization, the possible risks and threats of AI development. These difficulties are conditioned by the paradigm gap between the system of concepts used to describe technologies and the systems of concepts specific to socio-humanitarian descriptions and explanations. The so-called ontological paradox arises during digital transformation and the introduction of AI into social systems. An ontological paradox is also an epistemological paradox, since any statement of an ontological type presupposes its epistemological justification. To overcome this gap, V.E. Lepskiy proposes the conception of self-developing polysubjective environments – cybernetics of the third order. This conception allows to create a conceptual "bridge" between two systems of definitions that do not have direct logical connections between them. A productive tool for this can be the information approach, which is widely used to solve such interdisciplinary problems. This makes possible to theoretically correctly connect technological and socio-humanitarian descriptions in a single conceptual structure and to analyze main socio-humanitarian criteria for evaluating digital innovations. The article discusses in detail the main provisions of the information approach and its application for the development of AGI systems, the socio-humanitarian assessment of AI autonomous development, and the resolution of security issues.

Keywords: narrow artificial intelligence, socio-humanitarian criteria, ontological paradox, epistemological paradox, self-developing polysubjective environments, informational approach, security.

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. (2022) An Epistemological Analysis of the Social and Humanitarian Significance of Artificial Intelligence Innovations in Context of Artificial General Intelligence. *Russian Journal of Philosophical Sciences = Filosofskie nauki*. Vol. 65, no. 1, pp. 10–26.
DOI: 10.30727/0235-1188-2022-65-1-10-26

Введение

Нынешний этап развития ИИ, постановка и реализация задачи создания общего ИИ остро ставит новые философские

и теоретико-методологические вопросы, особенно касающиеся необходимости основательного эпистемологического осмысления в единой концептуальной структуре соотношения технологических и социогуманитарных аспектов этого развития. Здесь наблюдается всё еще недопустимый разрыв.

Техногенная цивилизация является по своей сути потребительской, замкнутой в параноидном круге «еще больше производить, чтобы еще больше потреблять, чтобы еще больше производить», и она подчиняет такому курсу развитие самого ИИ, а вместе с этим навязывает и сугубо технократический подход к его объяснению и осмыслению его социальной значимости. Сейчас крайне важно основательно анализировать и оценивать не только действительные социально значимые результаты цифровизации и развития ИИ, но и тех процессов, которые создают явно негативные последствия социогуманитарного порядка. Крупнейшие IT-корпорации в постоянно возрастающем масштабе продуцируют такие новации, которые направлены на разжигание аппетитов массового потребителя, на создание изощренных средств манипуляции массовым сознанием в своих экономических и политических интересах.

Онтологический и эпистемологический парадоксы

Критический анализ технократического подхода показал его полную концептуальную несостоятельность и позволил вскрыть слишком частую его ангажированность идеологами либерализма, глобализма, западной потребительской системы ценностей. Но в тоже время такой анализ позволил уточнить реальные теоретические трудности использования социогуманитарных критериев при моделировании и реализации функций ИИ и создать условия для разработки новых подходов.

Среди них наиболее значительной и продуктивной является, на наш взгляд, предложенная В.Е. Лепским концепция саморазвивающихся полисубъектных сред – кибернетики третьего порядка [Лепский 2010; Лепский 2017]. Она служит основой для преодоления указанного парадигмального разрыва, того, что именуется ***онтологическим парадоксом*** при цифровой трансформации и внедрении новаций ИИ в социальные системы.

Но необходимо учитывать, что онтологический парадокс является одновременно и ***эпистемологическим парадоксом***, поскольку всякое обоснованное утверждение онтологического типа предполагает его эпистемологическую рефлексию, т.е. рассмотрение достаточности тех познавательных, понятийных средств, которые при этом используются.

Эпистемологический парадокс выражает концептуальный разрыв между такими двумя системами понятий, которые не имеют между собой прямых логических связей. Одна из них служит описанию и объяснению социогуманитарных явлений (это такие понятия, как смысл, ценность, интенциональность, целеполагание, воля, справедливость, доверие и т.п.), вторая – описанию и объяснению физических свойств, устройства и функционирования технологических систем (это понятия физикалистского типа, такие, как масса, энергия, механические взаимодействия, пространственно-временные отношения и структуры и т.п.). Как же возможно приписывать технологическому агенту этические свойства или способность понимания смысла причинных связей? Подобная ситуация, именуемая в аналитической философии «провалом в объяснении», широко обсуждалась многими ее представителями в связи с попытками решения того, что принято называть трудной проблемой сознания, т.е. проблемой объяснения связи между явлениями субъективной реальности, которым нельзя приписывать физические свойства (массу, энергию, пространственные характеристики), и мозговыми процессами. Как известно, предлагавшиеся редукционистские способы решения этой проблемы оказывались мнимыми, так как в результате все те специфические особенности сознания, такие как субъективная реальность, которые требовали объяснения, фактически элиминировались и «провал в объяснении» оставался в силе. В работах [Дубровский 2020; Дубровский 2015; Dubrovsky 2019] предложен способ нередукционистского решения основных теоретических вопросов трудной проблемы сознания на основе информационного подхода.

Сейчас во многих областях науки и технологий настоятельно требуется теоретически обоснованное преодоление

такого провала, что особенно ярко проявляется в развитии ИИ, в том числе при моделировании и реализации описанных В.Е. Лепским основных социогуманитарных критериев оценки цифровых новаций (продуктивность, безопасность, развитие, удовлетворенность) [Лепский 2018], [Лепский 2020]. Для этого требуется определенный категориальный «мост», позволяющий теоретически корректно объединить в единой концептуальной структуре указанные два типа описания и объяснения, которые логически разобщены.

Информационный подход

Как показывает опыт последних десятилетий, категориальный «мост» создается посредством информационного подхода, который широко используется в различных областях науки для решения сложных междисциплинарных проблем. Он может служить эффективным инструментом для многоплановой реализации концепции В.Е. Лепского о саморазвивающихся полисубъектных средах. Несмотря на отсутствии общепринятой теории информации, понятие информации широко используется практически во всех научных дисциплинах. Это обусловлено тем, что понятие «информация» обладает рядом общепринятых определений. Среди них фундаментальное значение имеют следующие: 1) информация необходимо воплощена в своем физическом носителе, не существует вне и помимо него; 2) информация инвариантна по отношению к физическим свойствам своего носителя, т.е. одна и та же для данной системы информация может иметь разные по своим физическим свойствам носители, т.е. кодироваться по разному; 3) информация в биологических, социальных и социотехнических системах обладает не только синтаксическим, но также *семантическим* и *прагматическим* аспектами, которые релевантны для описания и объяснения таких свойств социогуманитарных явлений как смысл, ценность, цель, волевая активность, вера, управление. Поскольку эти определения общеприняты (по крайней мере, в биологических, социальных и социотехнических дисциплинах, к которым относится и область ИИ), они могут служить основанием для построения в них требуемых теоретических объяснений.

Здесь принципиальное значение приобретает вопрос о *кодировании и декодировании информации,* а также вопрос о способности системы к *самоорганизации*. Всякий носитель информации представляет собой определенную физическую структуру, которая является ***кодовым воплощением*** данной информации. Она сложилась в ходе биологической эволюции, а затем и социального развития (включая процессы развития и использования технологий). Наряду с кодовыми структурами, типическими для всего класса биологических и социальных систем, отдельные, единичные представители этих классов формируют свои специфичные кодовые структуры, обусловленные индивидуальным опытом их существования, приспособления, обучения. Во всех случаях информация, как таковая, выражает по сути функциональное ***значение*** для данной самоорганизующейся системы определенного комплекса физических свойств и физических воздействий, сложившегося в виде соответствующей кодовой структуры. Одни физические комплексы имеют для системы временное значение (например, в виде условного рефлекса, отдельных навыков), другие – для всего периода существования данного индивида (скажем, родной язык), а некоторые – для всей истории существования живых существ (генетический код). Информация выполняет функции программирования действий и управления ими. Но лишь в том случае, когда она актуализована, т.е. ***декодирована.*** Можно выделить два вида кодов: «естественные» и «чуждые». В случае «естественного» кода информация выступает для системы «прозрачной», так сказать, в «чистом» виде, т.е. декодируется автоматически; сразу «понятна», и готова для реализации функции управления; это имеет место в организме на многих уровнях регуляции жизненных процессов, например при мгновенном понимании человеком значения слов родного языка. В «чуждом» коде информация закрыта, «не понятна» для системы. Чтобы она смогла информацию «понять» и использовать, нужна специальная операция декодирования, расшифровки кода, требующая нередко значительных усилий. Это хорошо видно на примерах решения многих познавательных задач, особенно при расшифровке тайных сообщений или зна-

чения слов забытого языка. Опыт решения такого рода задач представляет значительный эпистемологический интерес для разработки и использования информационного подхода в различных научных дисциплинах (см.: [Сингх 2007]).

Нейронаука и «чтение мозга»

Нейронаука широко использует информационный подход для объяснения связи явлений сознания с мозговыми процессами, их декодирования и для объяснения феномена психической причинности. Сравнительно новое направление нейронауки, именуемое «чтением мозга» (brain-reading), используя методы картирования и визуализации мозговых процессов, исследует и выясняет разнообразные нейродинамические корреляты психических явлений. Оно ставит своей задачей расшифровку их мозговых кодов и достигло уже существенных результатов в области изучения явлений субъективной реальности. Еще более десяти лет тому назад японские исследователи Ё. Мияваки, Ю. Камитани и их сотрудники расшифровали нейродинамические эквиваленты зрительного восприятия статичных черно-белых объектов (отводя сигналы от мозга испытуемых на компьютер, воспроизводили на его экране переживаемые ими в данном интервале зрительные образы) [Miyawaki et al. 2008; Fujiwara, Miyawaki, Kamitani 2009]. При этом они смогли расшифровывать не только образы непосредственно воспринимаемого объекта, но и воспоминание о нем. Через некоторое время они научились расшифровывать цветные динамические образы (из кинофильма).

За последние годы на основе технологии глубоких нервных сетей расшифровка мозговых кодов психических явлений значительно усовершенствовалась. Нейроинформатики из Киотского университета в Японии предложили новый способ непосредственно по данным МРТ визуализировать на экране компьютера изображение, которое видит человек в момент сканирования мозга [Shen et al. 2019]. Впечатляющий результат получен в области прямого перевода из мозга в компьютер (т.е. декодирования) текста из 30–50 предложений на английском языке с помощью отведения сигналов примерно от 250 пунктов

коры мозга испытуемого. При этом средняя частота ошибок в словах не превышала 3% [Makin et al. 2018].

К настоящему времени установлено большое число достаточно четких корреляций между определенными психическими явлениями и соответствующими нервными процессами. Они широко используются для создания новых нейротехнологий в медицине, в различных интерфейсах «мозг – компьютер – машина». Они применяются в робототехнике, где их перспективы могут быть особенно значительными в области создания гибридных человеко-робототехнических систем с прямыми интерфейсами «мозг – интеллектуальный робот». Как известно, давно созданы и успешно развиваются средства, позволяющие парализованному человеку мысленно управлять курсором компьютера, инвалидной коляской и даже экзоскелетом. Сюда же относятся мысленно управляемые протезы конечностей. Установленные нейрокорреляты касаются не только действий, но и довольно сложных явлений субъективной реальности, имеющих прямое отношение к процессам мышления и его операциям, к феноменам внимания, намерений человека, выяснения ложных ответов на задаваемые ему вопросы. Все это позволяет считать теоретически возможным и весьма перспективным создание систем мысленного управления роботом.

Психическая и информационная причинность

Требуется основательное теоретическое осмысление вопроса о *психической причинности*. Очевидно, что моя мысль способна вызывать желаемое движение моей руки. Это простейший пример психической причинности. Но ведь по своим мысленным планам мы совершаем сложные системы действий и достигаем желаемых результатов. Как это объяснить, если мысли нельзя приписывать физические свойства? Достаточное теоретическое объяснение может быть дано на основе информационного подхода, согласно которому информация (в данном интервале) и ее физический носитель суть явления однопричинные, одновременные и постольку находятся в отношении взаимно однозначного соответствия. Моя мысль (содержанием которой является «желание и решение взять на столе стакан с водой»)

есть информация в виде сложного ментального образования, включающего наряду с моим желанием и решением мои образы собственной руки, стола, стакана с водой, ряда деталей наличной обстановки. Этот паттерн информации имеет своим физическим носителем определенную кодовую мозговую нейродинамическую структуру, способную запускать соответствующие моторные функции, реализующие мое желание. Здесь описано то, что именуют ***информационной причинностью.*** Как видим, она необходимо включает физический компонент. Но она отличается от обычной ***физической причинности*** тем, что вызываемое следствие тут определяется не сугубо физическими свойствами носителя информации, а именно содержанием информации, так как то же самое следствие может быть вызвано носителем этой же информации, имеющей другие физические свойства – в силу *принципа инвариантности информации по отношению к физическим свойствам ее носителя.* ***Психическая причинность есть вид информационной причинности.*** Помимо нее существуют другие виды информационной причинности (генетические, соционормативные, технические).

Вряд ли надо доказывать фундаментальное значение информационной причинности в процессах цифровизации, в развитии новых направлений ИИ и робототехники, которые отвечают требованиям социогуманитарной значимости, но в то же время и задачам своевременной диагностики и блокирования тех новаций ИИ, которые несут угрозу нашим жизненно важным интересам. Эти вопросы приобретают особенно высокую актуальность в связи с разработкой ОИИ, ставшей сейчас главным мировым трендом в развитии ИИ (в принятой международной номенклатуре ОИИ обозначается сокращенно *AGI*, от *Artificial General Intelligence*).

Разработка ОИИ происходит в условиях нарастающей конкуренции между крупнейшими научными центрами и специализированными корпорациями, – а в более широком масштабе, – между Россией и такими лидирующими в области ИИ государствами, как США, Япония, страны Западной Европы. Наше отставание, однако, не должно вызывать пессимистического настроя. Надо учитывать, что наши конкуренты – лидеры

в области классических направлений ИИ, главным образом, в области «узкого» ИИ. Для успехов же в области ОИИ необходимы принципиально новые теоретические и методологические подходы, новые технологические решения. В этом отношении у нас в стране ведется успешная работа.

Важнейшим условием успешной разработки ОИИ является преодоление ограниченности классической методологии А. Тьюринга, на которой до сих пор основывалось развитие ИИ и выдающиеся успехи информационных технологий, изменивших мир. Но сейчас мы вступили в стратегически новый этап развития ИИ. Методология А. Тьюринга носит сугубо операциональный, бихевиоральный характер и тем самым устраняет использование понятия сознания для описания и моделирования интеллекта. Этот недостаток не раз отмечался рядом специалистов в области ИИ; в последние годы предпринимаются попытки его устранения. Одним из таких наиболее показательных примеров может служить концепция посттьюринговой методологии, предложенная А.Р. Ефимовым, которая существенно расширяет диапазон функций ИИ в интеллектуальной робототехнике [Ефимов 2020].

Требуемые ресурсы

Для успешной разработки ОИИ требуется тщательный анализ пригодных для этих целей ресурсов, имеющихся в различных научных и философских дисциплинах, особенно в нейронауке, психологии, лингвистике и, конечно, в эпистемологии и феноменологических исследованиях структур субъективной реальности, в парадигме постнеклассической рациональности. Значительные ресурсы содержатся в результатах эпистемологического анализа специфических для естественного интеллекта когнитивных структур, касающихся таких феноменов как внимание, категоризация стимулов и действий, дискретизация непрерывных процессов, обобщение и оценка эмпирических данных, соотношение высказываний от первого и от третьего лица, обучение новому навыку, понимание причинных связей и др., т.е. того, что необходимо для ОИИ. Следует подчеркнуть, что большое значение для указанных целей имеют результаты

феноменологических исследований субъективной реальности, позволяющие выделить и описать динамические структуры и операции целереализующего мыслительного процесса, которые могут служить основанием для моделирования специфических когнитивных архитектур ОИИ. Некоторые вопросы этого плана рассмотрены в моей статье [Дубровский 2021].

Вопрос о ресурсах является широким и многоплановым, требует специального анализа. Но здесь хотелось бы еще раз отметить первостепенное значение нейронаучных исследований сознания для проблематики ОИИ. Поскольку отличительным свойством системы ОИИ выступает ее *автономность*, способность самостоятельного решения задач в разных средах, то для моделирования такой способности самообучения может служить описание двуединой функции зеркальных нейронов, которые одновременно кодируют образ предмета и способ действия с ним, что, по мнению исследователей, лежит в основе быстрого обучения новым навыкам. Для понимания и моделирования механизмов самообучения имеют важное значение многие другие достижения нейронауки, раскрывающие механизмы самоорганизации психических процессов [Анохин 2021; Дубровский 2022].

Продвижение в разработке ОИИ будет означать появление все новых функций ИИ, в том числе таких, о которых мы ранее не подозревали, что связано с развитием у технологического агента способности самообучения и самоорганизации. Это ставит сложные вопросы социогуманитарной оценки достижений ОИИ. Одно дело, когда мы проектируем функции робота, число которых ограничено, и заведомо опасные для нас функции могут блокироваться, другое дело, когда робот сам производит новые непредвиденные функции. Такая ситуация должна быть предметом внимания и специального изучения.

Когда речь идет о социогуманитарных оценках новаций ИИ и определении их критериев, то последние имеют два противоположных плана: в одном описывается и оценивается позитивное значение, в другом – негативное. Трудности возникают в связи с тем, что невозможно выстроить строго однозначную иерархическую структуру социогуманитарных ценностей для

всех жизненных ситуаций, т.е. такую, которая бы позволяла во всех случаях производить строго альтернативный выбор. Это хорошо видно при попытках моделирования этически приемлемых функций интеллектуального агента. Например, такая высокая этическая ценность как следование правде и истине в ряде ситуаций вступает в резкое противоречие с другой столь же высокой этической ценностью как, например, сохранение жизни человеку. В свою очередь, сохранение жизни как высшей ценности ставится под вопрос, когда речь идет о таких высоконравственных поступках, как исполнение воинского долга и в других случаях героического самопожертвования в экстремальных ситуациях.

Однако надо подчеркнуть, что во многих видах ситуаций одна определенная этическая ценность, несмотря на ее относительность, приобретает абсолютное значение, не имеет альтернативы и должна быть неукоснительно реализована. Такие этические решения и такие виды ситуаций допускают четкую классификацию для *определенных видов деятельности,* например, при создании интеллектуальных роботов. Сфера видов их деятельности постоянно расширяется, но она все же ограничена. Поэтому при проектировании робота, предназначенного для данного вида деятельности, можно жестко программировать исполнение одного определенного действия и категорический запрет другого.

Тем не менее проблема моделирования этически санкционированных функций у интеллектуального агента, взятая в ее общем виде, остается нерешенной. Судя по существующим исследованиям в той области, которая именуется «этикой искусственного интеллекта», позитивные результаты могут быть достигнуты с помощью построения ***частных моделей*** на основе классификатора видов деятельности, что позволит осуществить алгоритмизацию определенных этических функций. При этом применяются различные методы (например, метод обучения с подкреплением, используемый для определения оптимального выбора в ситуации неопределенности). В большинстве концептуальных подходов, характерных для проблем «этики искусственного интеллекта», в центре внимания находятся задачи успешного взаимодействия людей с искусственными

автономными системами, в особенности с продвинутыми интеллектуальными роботами, которых учат или которые уже «научились» выполнять ряд санкционированных оператором этически приемлемых действий.

Однако все это большей частью касается лишь функциональных аспектов поведения искусственного интеллектуального агента, остается вне контекста рассмотрения социогуманитарной значимости развития ИИ и, как правило, вне рассмотрения проблематики ОИИ. Между тем именно вопросы разработки ОИИ, уже достигнутые и ожидаемые результаты, приобретают в современных условиях общественной жизни исключительно высокую актуальность. Все более острыми и неотложными становятся проблемы социогуманитарной безопасности, взятые во всем их масштабе и разнообразии. Это вопросы политической, экономической, военной, информационной безопасности и, соответственно, укрепления и дальнейшего развития соответствующих сфер жизнедеятельности.

Заключение

В настоящее время идет ломка системы информационных связей внутри страны и за рубежом, слом мировой финансовой системы и экономики, разжигание межнациональных и иных конфликтов. В этом контексте надо всемерно укреплять безопасность нашей страны во всех областях, включая информационную безопасность, понимаемой как безопасность на всех уровнях и во всех системах и видах коммуникаций, органически связанных с ИИ и ОИИ, в том числе для целей двойного назначения. Это – задача первостепенной политической и социогуманитарной значимости. В ее решении должны принимать самое активное участие не только профессионалы, работающие в области ИИ и ОИИ, масс-медиа, инженерно-технологических специальностей, но и представители социальных и гуманитарных дисциплин.

ЦИТИРУЕМАЯ ЛИТЕРАТУРА

Анохин 2021 – *Анохин К.В.* В поисках фундаментальной нейронаучной теории // Журнал высшей нервной деятельности. 2021. Т. 71. № 1. С. 39–71.

Дубровский 2020 – *Дубровский Д.И.* Психические явления и мозг. Философский анализ проблемы в связи актуальными задачами нейрофизиологии, психологии и кибернетики / 2-е изд., доп. – М.: ЛЕНАНД, 2020.

Дубровский 2015 – *Дубровский Д.И.* Проблема «Сознание и мозг»: Теоретическое решение. – М.: Канон+, 2015.

Дубровский 2021 – *Дубровский Д.И.* Задача создания Общего искусственного интеллекта и проблема сознания // Философские науки. 2021. № 64. № 1. С. 13–44.

Дубровский 2022 – *Дубровский Д.И.* Значение нейронаучных исследований сознания для разработки общего искусственного интеллекта (методологические вопросы) // Вопросы философии. 2022. № 2. С. 83–93.

Ефимов 2020 – *Ефимов А.Р.* Посттьюринговая методология: разрушение стены на пути к общему искусственному интеллекту // Интеллект. Инновации. Инвестиции. 2020. № 2. С. 74–80.

Лепский 2010 – *Лепский В.Е.* Рефлексивно-активные среды инновационного развития. – М.: Когито-Центр, 2010.

Лепский 2017 – *Лепский В.Е.* Седьмой социогуманитарный технологический уклад – контуры будущего человечества // Глобальный мир: системные сдвиги, вызовы и контуры будущего: XVII Международные Лихачевские научные чтения, 18–20 мая 2017 г. – СПб.: СПбГУП, 2017. С. 357–360.

Лепский 2018 – *Лепский В.Е.* Социогуманитарные критерии оценки новаций цифровой реальности // Социальное время. 2018. № 4 (16). С. 16–26.

Лепский 2020 – *Лепский В.Е.* Философско-методологические основания оценки социально-психологических последствий внедрения новых технологий // Психологический журнал. 2020. Т. 41. № 4. С. 105–108.

Сингх 2007 – *Сингх С.* Книга кодов. Тайная история кодов и их «взлома». – М.: АСТ, Астрель, 2007.

Dubrovsky 2019 – *Dubrovsky D.I.* “The Hard Problem of Consciousness”. Theoretical Solution of its Main Questions // AIMS Neuroscience. Vol. 6. No. 2. P. 85–103.

Fujiwara, Miyawaki, Kamitani 2009 – *Fujiwara Y., Miyawaki Y., Kamitani Y.* Estimating Image Bases for Visual Image Reconstruction from Human Brain Activity // Advances in Neural Information Processing Systems. Vol. 22. P. 576–584.

Makin, Moses, Chang 2020 – *Makin J.G., Moses D.A., Chang E.F.* Machine Translation of Cortical Activity to Text with an Encoder-Decoder Framework // Nature-Neuroscience Technical Report. 2020. Vol. 23. No. 4. P. 575–582.

Miyawaki et al. 2008 – *Miyawaki Y., Uchida H., Yamashita O., Sato M.A., Morito Y., Tanabe H.C., Sadato N., Kamitani, Y.* Visual Image Reconstruction from Human Brain Activity Using a Combination of Multiscale Local Image Decoders // Neuron. Vol. 60. No. 5. P. 915–929.

Shen et al. 2019 – *Shen G., Horikawa T., Majima K., Kamitani Y.* Deep Image Reconstruction from Human Brain Activity // PLoS Computational Biology. Vol. 15, no. 1, e1006633.

REFERENCES

Anokhin K.V. (2021) In Search of a Fundamental Neuroscientific Theory. *Journal of Higher Nervous Activity.* Vol. 71, no. 1, pp. 39–71 (in Russian).

Dubrovsky D.I. (2015) *"Consciousness and Brain" Problem: A Theoretical Solution.* Moscow: Kanon+ (in Russian).

Dubrovsky D.I. (2019) "The Hard Problem of Consciousness." Theoretical Solution of Its Main Questions. *AIMS Neuroscience.* Vol. 6, no. 2, pp. 85–103.

Dubrovsky D.I. (2020) *Mental Phenomena and the Brain. Philosophical Analysis of the Problem in Connection with Actual Problems of Neurophysiology, Psychology and Cybernetics* (2nd ed.). Moscow: LENAND (in Russian).

Dubrovsky D.I. (2021) The Task of Creating General Artificial Intelligence and the Problem of Consciousness. *Russian Journal of Philosophical Sciences = Filosofskie nauki.* Vol. 64, no. 1, pp. 13–44 (in Russian).

Dubrovsky D.I. (2022) The Value of Neuroscience Research of Consciousness for the Development of Artificial General Intelligence (Methodological Issues). *Voprosy filosofii.* No. 2, pp. 83−93 (in Russian).

Efimov A.R. (2020) Post-Turing Methodology: Breaking the Wall toward Artificial General Intelligence. *Intellekt. Innovatsii. Invesititsii.* No. 2, pp. 74–80 (in Russian).

Fujiwara Y., Miyawaki Y., & Kamitani Y. Estimating Image Bases for Visual Image Reconstruction from Human Brain Activity. *Advances in Neural Information Processing Systems.* Vol. 22, pp. 576–584.

Lepskiy V.E. (2010) *Reflexively Active Environments of Innovative Development.* Moscow: Kogito-tsentr (in Russian).

Lepskiy V.E. (2017) The Seventh Socio-Humanitarian Technological Order − the Outlines of the Future of Mankind. In: Zapesotsky A.S., Markov A.P., Paseshnikova L.A. (Eds.) *Global World: Systemic Shifts, Challenges and Outlines of the Future: 17th International Likhachev Scientific Readings, May 18–20, 2017* (pp. 357–360). Saint Petersburg: SPbGUP (in Russian).

Lepskiy V.E. (2018) Socio-Humanitarian Criteria for Evaluating the Innovations of Digital Reality. *Sotsial'noye vremya.* No. 4, pp. 16−26 (in Russian).

Lepskiy V.E. (2020) Philosophical and Methodological Foundations for Assessing the Socio-Psychological Consequences of the Introduction of New Technologies. *Psychological Journal.* Vol. 41, no. 4, pp. 105–108 (in Russian).

Makin J.G., Moses D.A., & Chang E.F. (2020) Machine Translation of Cortical Activity to Text with an Encoder-Decoder Framework. *Nature-Neuroscience Technical Report.* Vol. 23, no. 4, pp. 575–582.

Miyawaki Y., Uchida H., Yamashita O., Sato M.A., Morito Y., Tanabe H.C., Sadato N., & Kamitani Y. (2008) Visual Image Reconstruction from Human Brain Activity Using a Combination of Multiscale Local Image Decoders. *Neuron.* Vol. 60, no. 5, pp. 915–929.

Shen G., Horikawa T., Majima K., & Kamitani Y. (2019) Deep Image Reconstruction from Human Brain Activity. *PLoS Computational Biology.* Vol. 15, no. 1, e1006633.

Singh S. (1999) *The Code Book: The Science of Secrecy from Ancient Egypt to Quantum Cryptography.* New York: Doubleday (Russian translation: Moscow: AST, Astrel, 2007).

DOI: 10.30727/0235-1188-2022-65-1-27-43
Оригинальная исследовательская статья
Original research article

Искусственный интеллект: туманность определений в неопределенности реалий*

А.И. Агеев
Институт экономических стратегий РАН, Москва, Россия

Аннотация

Развитие систем искусственного интеллекта (СИИ) и цифровая трансформация в целом ведут к образованию коллективов автономных агентов искусственной и смешанной генеалогии, а также сложных конструкций информационной и нормативной среды со множеством возможностей и патологий и растущим уровнем неопределенности при принятии управленческих решений. Ситуацию осложняет сохраняющаяся множественность понимания сущности СИИ. Современное расширенное понимание ИИ восходит к представлениям, сформулированным более 100 лет назад. В официальных страновых программных документах о развитии СИИ предпочтение отдается рабочим определениям ИИ. Текущий статус жизненного цикла СИИ можно оценить как завершение стартового этапа развития систем, связанных со «слабым» ИИ. Способность искусственных систем к осознанию себя в качестве отдельной личности становится одним из серьезных научно-практических вызовов. Внимание к этике СИИ свидетельствует о начале их работы в пространстве целеполагания и расширении классов и контуров используемых данных. Новые морально-этические проблемы возникают и в связи с созданием в обозримой перспективе подлинно осознающих субъектов. Наблюдается усиливающийся феномен деградации естественного интеллекта. Требуется учитывать разнородность данных, генерируемых человеком, электронными сенсорами и сетевыми устройствами в динамических проблемных средах цифровой экономики, сложности процесса коэволюции СИИ, коллективного и индивидуального естественного сознания. Особая сфера возможностей и рисков – развитие нейротехнологий. Объектом управления становятся цифровые двойники, через которые может осуществляться манипуляция реальными установками, оценками и поведением личности. Как следствие, развиваются технологические возможности

* Работа поддержана Российским научным фондом (РНФ), грант № 21-18-00184 «Социогуманитарные основания критериев оценки инноваций, использующих цифровые технологии и искусственный интеллект».

для провоцирования деструктивных явлений и формируется новый класс массовых зависимостей.

Ключевые слова: философия искусственного интеллекта, естественный интеллект, киберфизические системы, цифровая трансформация, жизненный цикл, подлинно осознающие субъекты.

Агеев Александр Иванович – доктор экономических наук, профессор, директор Института экономических стратегий РАН.
ageev@inesnet.ru
https://orcid.org/0000-0002-2826-2702

Для цитирования: *Агеев А.И.* Искусственный интеллект: туманность определений в неопределенности реалий // Философские науки. 2022. Т. 65. № 1. С. 27–43. DOI: 10.30727/0235-1188-2022-65-1-27-43

Artificial Intelligence: The Opacity of Concepts in the Uncertainty of Realities*

A.I. Ageev
Institute for Economic Strategies, Russian Academy of Science, Moscow, Russia

Abstract

The development of the systems of artificial intelligence (AI) and digital transformation in general lead to the formation of multitude of autonomous agents of artificial and mixed genealogy, as well as to complex structures in the information and regulatory environment with many opportunities and pathologies and a growing level of uncertainty in making managerial decisions. The situation is complicated by the continuing plurality of understanding of the essence of AI systems. The modern expanded understanding of AI goes back to ideas formulated more than 100 years ago. In official national policy documents on the development of AI, working definitions of AI are preferred. The current stage of AI systems life cycle can be assessed as the completion of the initial period in the development of systems associated with weak AI. The ability of artificial systems to realize themselves as a separate person becomes one of the serious scientific and practical challenges. Attention to the issues of the ethics of AIS indicates the expansion of the diversity of its forms and the beginning of the work in the field of goal-setting. New moral and ethical problems also arise in connection with

* The work was supported by the Russian Science Foundation, grant no. 21-18-00184 “Social and humanitarian foundations for evaluation criteria for innovations based on digital technologies and artificial intelligence.”

the possibility of the creation of genuine conscious subjects in the foreseeable future. There is an increasing phenomenon of degradation of natural intelligence. It is required to take into account the issue of the heterogeneity of data generated by humans, electronic sensors and network devices in the dynamic complex environments of the digital economy, the issue of the complexity of the process of co-evolution of AI systems, collective and individual natural consciousness. A special area of opportunities and risks is the development of neurotechnologies. The object of control is digital twins, through which there can be manipulation of real attitudes, preferences, and behavior of individuals. As a result, there are the development of technological capabilities that provoke destructive phenomena as well as the formation of a new class of mass addictions.

Keywords: philosophy of artificial intelligence, natural intelligence, cyber-physical systems, digital transformation, life cycle, genuine conscious subjects.

Alexander I. Ageev – D.Sc. in Economics, Professor, Director of the Institute for Economic Strategies, Russian Academy of Science.
ageev@inesnet.ru
https://orcid.org/0000-0002-2826-2702

For citation: Ageev A.I. (2022) Artificial Intelligence: The Opacity of Concepts in the Uncertainty of Realities. *Russian Journal of Philosophical Sciences = Filosofskie nauki*. Vol. 65, no. 1, pp. 27–43.
DOI: 10.30727/0235-1188-2022-65-1-27-43

Постановка проблемы

Системы искусственного интеллекта (СИИ) справедливо признаются ансамблем наиболее всеобъемлющих технологий современности, которые окажут и уже оказывают существенное влияние на развитие систем управления на всех уровнях. Достаточно указать на внедрение системы социального рейтинга в Китае, опыт Сингапура или эксперимент с внедрением ИИ в практику управления социально-экономическими процессами в Москве, а главное – на колоссальные инвестиции в разработку и внедрение СИИ прежде всего в Китае и США.

СИИ являются ключевыми среди всех сквозных цифровых технологий. Цифровая трансформация ведет к экспоненциальному образованию автономных агентов искусственной и смешанной генеалогии, их сообществ с различной устойчивостью, сложных, противоречивых и динамичных конструкций информационной

и нормативной среды. Возрастает риск деструктивных действий формально естественных агентов социальных процессов, в том числе в случаях воздействия на их сознание искусственных агентов. Это порождает множество вызовов для принятия управленческих решений на всех уровнях и для общественной динамики в целом. На фоне множественных угроз и усиливающегося кризиса возможно быстрое манифестирование как предвидимых угроз, так и событий, которые принято называть «черными лебедями».

Следует учитывать масштаб уже наступившего «информационного потопа», который ежесекундно усиливается за счет новых массивов данных в силу пользовательской активности человека и электронных устройств и систем. Динамичность виртуального пространства, появление принципиально новых феноменов создают проблемы идентификации активных и пассивных субъектов социальной деятельности, предотвращения целенаправленных и случайных деструктивных действий. Необходимо создавать новые модели анализа и прогнозирования обстановки, учитывающие непрерывное изменение состояния объектов и субъектов, их кластеризацию или распад на базе информационных взаимосвязей и их рефлексивных эффектов.

Философские аспекты данной проблематики охватывают вопросы определения сущности ИИ, коэволюции искусственных и естественных систем, их идентичности и коллективности, перспектив и последствий внедрения СИИ и особенно нейротехнологий в социальную реальность.

Запутанность понимания

Пока не существует общепринятого определения ИИ. Международный и национальный стандарты ИИ, задающие его базовые определения и области применения, находятся в фазе интенсивной разработки, хотя уже введено в действие множество стандартов, охватывающих различные аспекты тематики.

В структуре Технического комитета (ТК) ИСО и МЭК 42 «Искусственный интеллект» в 2019–2022 годах существовали 5 рабочих групп, первая – под названием «Основополагающие принципы ИИ» была нацелена на решение базовых вопросов

ИИ (включая определения и этику ИИ). Сам российский ТК 164 «Искусственный интеллект» в РФ образован в июле 2019 года, по структуре подобен ТК 42 «ИИ» ИСО и МЭК[1]. Всего по проекту базового стандарта ИИ уполномоченные представители более чем 70 стран высказали более 1000 комментариев. Бóльшая часть из них была так или иначе учтена. В настоящее время разрабатывается более 30 профильных стандартов. Все это означает, что еще некоторое время будут существовать риски неточного, несогласованного (недоопределенного) толкования СИИ и, как следствие, недобросовестного или некомпетентного отнесения к ним продукции, которая не соответствует критериям ИИ.

Применяемые для конкретных целей определения делаются либо дедуктивно, либо индуктивно, либо функционально, либо через родовидовое отличие, либо через отрицание. При определении сущности ИИ представляется полезным взять за точку отсчета понимание естественного интеллекта. Латинский термин *intellectus* означает «ум, рассудок, разум; мыслительные способности человека». В словаре Брокгауза и Ефрона статья «интеллект» отсылает к весьма обширной статье «умъ». В статье выделены три вида умственной деятельности: 1) восприятие явлений и их интеллектуальная переработка; 2) изменение эмоционального равновесия; 3) волевые импульсы. Только содержание первого пункта относилось к собственно «уму», хотя делались попытки либо расширить, либо сузить даже это определение. Более ста лет назад в понимании «ума» подчеркивалось выдающееся влияние памяти, внимания и «утомляемости личности» на интеллектуальную жизнь. Важнейшая функция умственной деятельности, как полагалось, состояла в сочетании вновь воспринимаемых явлений с накопленными воспоминаниями и опытом и выработке «целесообразной и планомерной реакции». Считалось, и не без оснований, что сложные последовательности реакций на разные впечатления рутинизируются, становясь инстинктами, и протекают рефлекторно, «не проникая в сознание» [Энциклопедический словарь… 1902, 731–734].

[1] См. приказ Росстандарта от 18.03.2020 № 579 «О внесении изменений в Программу национальной стандартизации на 2020 год, утвержденную приказом Федерального агентства по техническому регулированию и метрологии от 1 ноября 2019 г. № 2612».

История ИИ как нового направления в науке начинается в середине XX века. К этому времени сложилось множество предпосылок в математике, кибернетике, гносеологии, нейрофизиологии, психологии. В науке сформировались многочисленные вычислительные традиции (теория алгоритмов, первые компьютеры). Алан Тьюринг в статье «Вычислительная техника и интеллект» (1950) обсуждал критерии, позволяющие считать машину интеллектуальной [Turing 1950]. Важно отметить, что в английском языке словосочетание *artificial intelligence* не имеет такой антропоморфной окраски, которую оно приобрело в русском переводе. В то время одним из ключевых вопросов при обсуждении ИИ стала способность компьютеров мыслить и осознавать себя как отдельную личность. Вопрос об этике и самоидентификации СИИ вновь стал актуальным в наши дни, поскольку стала очевидна техническая возможность СИИ оперировать в пространстве целеполагания и расширять классы и контуры используемых данных, в частности в Интернете вещей.

В настоящее время интеллект понимается как «общая познавательная способность, которая проявляется в том, как человек воспринимает, понимает, объясняет и прогнозирует происходящее, какие решения он принимает и насколько эффективно он действует (прежде всего в новых, сложных и необычных ситуациях)» [БРЭ 2008, 429–430].

Палитра пониманий интеллекта включает: 1) рассмотрение специфики организации его как «базы знаний» (объем, разнообразие, актуальность, компетентность), которая рассматривается в качестве критерия развитости интеллекта; 2) интеллект как система мыслительных операций: анализ, синтез и обобщение, при этом скорость переработки данных выступает критерием развития интеллекта; 3) механизмы ментального самоуправления, форма организации ментального опыта и т.п. [БРЭ 2008, 429–430].

Интеллект также определяют как «способность приходить к решению при помощи вычислений» [McCarthy 2007, 2]. Интеллект разного вида и уровня присутствует у людей, многих животных и некоторых машин. Говард Гарднер в 1980-е годы выделил семь сторон интеллекта, которые выражены у людей в разной степени и в разных пропорциях: лингвистическую интеллектуальность,

логико-математические составляющие, оцениваемые тестом IQ, музыкальные способности, способность к пространственному видению, кинестетические способности и др. [Гарднер 2007].

Согласно самому общему современному определению, интеллект трактуется как способность к процессу познания и эффективному решению проблем, в частности при овладении новым кругом жизненных задач. Нельзя не заметить, что данное понимание недалеко ушло от определения «ума» в словаре Брокгауза и Ефрона более века назад. Тем не менее определение интеллекта варьирует в зависимости от сферы применения. Релевантной может быть трактовка интеллекта как «социально полезной адаптации». Кроме того, представляется уместным использовать понятие «домен»: принципиальное отличие и преимущество человека до сих пор заключаются именно в способности оперировать знаниями и опытом из разных доменов, развивая способности, создавая новые знания, навыки и даже домены.

В центре современной дискуссии об определениях и границах СИИ находится вопрос об алгоритмах, который играет роль своего рода понятийного водораздела. В математике и кибернетике класс задач определенного типа считается решенным, когда для их решения установлен алгоритм. Изучение и нахождение алгоритмов является естественной целью человека при решении разнообразных проблем. В определенном смысле прогресс культуры можно описать как накопление запаса алгоритмов (стереотипов, рутин, навыков) решения тех или иных проблем. Задачи, для решения которых алгоритм еще не найден и требуется усилие, изобретательность и проницательность человеческого ума, относят к категории интеллектуальных [Каляев 2019]. Соответственно, в строгом смысле этот класс задач и следует относить к объему проблем СИИ. Стремление к компромиссу в определении по критерию «интеллектуальности задач» в зависимости от нахождения алгоритма привело к разделению СИИ на «слабые» и «сильные»[2].

[2] Сильный ИИ – система, способная моделировать человеческие чувства, намерения и мышление путем обработки символов, физических полей и других видов материи и энергии для решения сложных междисциплинарных задач с глубоким пониманием того, что она делает. Слабый ИИ – система, способная решать интеллектуальные задачи, обрабатывая

В российском программном документе по СИИ – «Национальной стратегии развития искусственного интеллекта на период до 2030 года» – дано прагматичное определение ИИ как «комплекса технологических решений, позволяющих имитировать когнитивные функции человека (включая самообучение и поиск решений без заранее заданного алгоритма) и получать при выполнении конкретных задач результаты, сопоставимые, как минимум, с результатами интеллектуальной деятельности человека»[3].

В стандарте *ISO CD* 22989 ИИ понимается как «дисциплина, которая изучает инженерные системы, способные приобретать, обрабатывать и применять знания и опыт». В рамках этой дисциплины разработано несколько подходов и методов: от машинного обучения и машинного мышления, включающего прогнозирование, планирование, представление знаний и рассуждение, поиск и оптимизацию, от робототехники до интеграции различных методов в киберфизические системы.

Анализ определений в национальных и международных документах по ИИ показывает, что предпочтение отдается рабочим определениям. Однако есть основания рассчитывать, что в ближайшее время на уровне ИСО/МЭК будут сформулированы относительно общепринятые представления об ИИ и утверждены в качестве стандартов.

В национальном стандарте «Системы искусственного интеллекта. Классификация систем искусственного интеллекта»[4], подготовленном с учетом предложений упомянутой выше рабочей группы (РГ 01 ТК 164 ИИ), указано, что он разрабатывался в целях установления принципов классификации систем ИИ и повышения

только символы, не понимая, что она делает, но быстрее, чем это сделали бы люди [Raikov 2021]. Правда, в формирующихся международных стандартах (проект *ISO-IEC* 22989) отмечается, что обозначения «слабый ИИ» и «сильный ИИ» в основном важны для философов и неактуальны для практиков ИИ.

[3] См. Указ Президента Российской Федерации от 10.10.2019 № 490 «О развитии искусственного интеллекта в Российской Федерации». – URL: https://www.kremlin.ru/acts/bank/44731

[4] См. национальный стандарт РФ «Системы искусственного интеллекта. Классификация систем искусственного интеллекта». – URL: https://protect.gost.ru/default.aspx/v.aspx?control=7&id=239563

эффективности их использования для решения прикладных задач как в автономном режиме, так и во взаимодействии с человеком-оператором.

Классификация позволяет сравнивать различные решения в СИИ по таким параметрам, как вид деятельности, структура знаний, функции контура управления, безопасность, конфиденциальность, степень автоматизации, методы обработки информации, интеграция/интероперабельность, комплексность системы, архитектура, специализация [Кукшев 2020].

Таким образом, для прагматических целей смыслы и значения понятия ИИ (точнее – СИИ) можно считать определенными. Неопределенности являются предметом дальнейших научных изысканий, требуют учета при осмыслении практических проблем, возникающих в ходе цифровой трансформации, чреватой серьезными социальными последствиями.

Статус жизненного цикла СИИ

В истории развития СИИ имели место периоды как бурного развития, сопровождавшегося значительными инвестициями, так и утраты интереса к феномену искусственного интеллекта, разочарования в ожиданиях. В 1960-х годах ставились и решались достаточно сложные задачи в науке и технике, управлении, военном деле. Далеко не все удалось решить на удовлетворительном уровне из-за отсутствия вычислительных мощностей, низкой производительности компьютеров и отчасти из-за социальных причин. В настоящее время наблюдается новый всплеск интереса к СИИ, при этом отличающиеся от прежних формулировки создают иллюзию новизны. В любом случае можно констатировать фазу завершения этапа стартового развития систем, связанных со слабым ИИ. Созданы социальные сети, электронные торговые площадки, системы видеомониторинга, военные и транспортные устройства, центры сбора и обработки больших данных, где используются СИИ.

Имелись обоснованные ожидания, что к 2025 году массовое внедрение технологий ИИ обеспечит удвоение темпов роста ВВП ведущих государств мира и увеличение мирового ВВП на 15 трлн долл. [Каляев 2019]. Внедрение технологий ИИ

даст значительный экономический и социальный эффект в промышленности, энергетике, сельском хозяйстве, финансовом секторе, образовании, городской инфраструктуре, в области безопасности и противодействия терроризму, на транспорте, в обороне и во многих других сферах.

В 2017 году первые пять стран приняли национальные программы в сфере ИИ, в 2022 году их уже больше сорока. Пока только пять стран (среди которых и Россия) создали такие важные собственные элементы цифровой экосистемы, как поисковик, социальные сети, развитые школы криптографии, инфраструктура кибербезопасности, система подготовки кадров в области математики, логики, программирования, ИТ.

Выделяются следующие приоритетные направления исследований в области ИИ: нейронные сети, глубокое обучение, теория управления, умный контроль, компьютерное зрение, технологии поиска и оптимизации, анализ речи, генерация естественного языка и речи, компьютерная логика и рассуждение, когнитивные вычисления, вероятностные методы выбора в условиях неопределенности, классификаторы и методы статистического обучения, технологии взаимодействия систем с искусственным интеллектом, нейроинтерфейсы, чтение сигналов мозга, нейроинформатика, электроэнцефалография [Осоченко, Макушкин 2019]. Работа по этим направлениям ведется практически во всех научно-исследовательских организациях РФ. Однако на предприятиях промышленного сектора вопросами применения ИИ интересуются в меньшей степени. Как показало наше исследование [К цифре… 2018], одной из двух наиболее существенных угроз, сопровождающих цифровизацию, является деградация естественного интеллекта. Готовность же ее парировать является низкой прежде всего потому, что возникновение такой угрозы является результатом сочетания множества процессов эволюции социума. В связи с этим в рамках исследований СИИ приоритетной должна быть задача понимания особенностей эволюции «естественного» сознания, коэволюции и гибридизации искусственных и естественных систем.

СИИ и нейротехнологии

В настоящее время замещение человека на СИИ происходит во многих сферах и в нарастающих масштабах. При создании подлинно осознающих субъектов (ПОС), способных выбирать цели своей деятельности и произвольно работать с разными базами знаний, неизбежно возникают морально-этические проблемы. Под ПОС понимаются живые или подобные живым существа, которые обладают самосознанием, субъективными переживаниями, схожими с самосознанием и субъективными переживаниями человека или другого высокоразвитого существа. Степень развитости этих свойств ПОС простирается между тем, что понимается под слабым и сильным ИИ [Социально-экономические аспекты… 2020].

При разработке СИИ неизбежно допускается некоторая неопределенность деятельности такой системы, которая сходна со свободой воли человека. Действия ПОС должны учитывать этические самоограничения, знания о нормах поведения людей, механизмы саморегуляции поведения, способность к эмпатии, механизм прогнозирования рисков и последствий собственных действий, возможность исправления собственной ошибки, в том числе исключение действий, связанных с особым риском для человечества.

Среди наиболее острых вопросов развития ИИ – нейротехнологии, которые используют или помогают понять работу мозга, мыслительные процессы, высшую нервную деятельность, в том числе технологии по усилению, улучшению работы мозга и психической деятельности. В сфере нейротехнологий наблюдается стремительный прогресс НИР, ОКР, практического применения. Следует обратить внимание на введенный в действие еще в 2009 году стандарт[5], описывающий практические вопросы работы операторов сложных технических систем, а также их взаимодействия с социумом [Агеев, Логинов 2019].

Формирование целостности личности происходит в процессе динамического взаимодействия людей и надличностных систем

[5] См. национальный стандарт РФ «Информационное обеспечение техники и операторской деятельности. Ноон-технология в технической деятельности». – URL: https://protect.gost.ru/document.aspx?control=7&id=176044

в современном социуме. На личность влияют когнитивные проекции социума, индивидуализированные биологически (тело), информационно (первичная и вторичная социализация с освоением видов грамотности, коммуникативные связи и базы данных), когнитивно (знания, чувствования и понимания), социально (принадлежность к агрегированным группам, которых может быть множество по разным критериям). Отсюда – неизбежность комплексирования различных моделей поддержки, лояльных к правовым нормам, управленческим ключевым установкам (в том числе цифровым, анонимным) и глубинным регуляторам жизнедеятельности («матрицам») социальной среды. При этом сбои в работе киберфизических систем могут быть вызваны дистанционным способом. Современный уровень технологий позволяет осуществлять зондаж и воздействие с учетом широкого спектра психосемантических качеств личности (официальной и реальной политической ориентации, качества профессиональной подготовки, культурного уровня, интересов, волевых качеств, внутренней мотивации и т.п.) [Дьяков 2015; Севостьянов 2014; Ясницкий, Сичинава 2011].

Новейшие поколения массовых технических устройств дают возможность заинтересованному пользователю не только идентифицировать и определять геолокацию обладателя устройства, его эмоциональное отношение к содержанию электронных сообщений, но и выявлять характеристики окружающего оборудования, опираясь на автоматизированные системы сбора, накопления, обработки и использования данных. Получение информации о состоянии личности через «гаджеты здоровья» дают возможность при необходимости прогнозировать изменение состояния человека в нормальных и чрезвычайных условиях [Ляхов, Тришин 2013; Самарцев, Латенкова 2016]. Выявление особенностей пользователя по характеристикам просматриваемых ею информационных программ, активности в социальных сетях, выбору компьютерных игр и т.п. (портфель данных индивидуального электронного контента) позволяет сформировать когнитивно-рефлексивную модель личности. На основе такой модели возможно нейропрограммирование мировоззренческих и ситуативных ориентиров и актов поведения личности и групп людей [Волынский-Басманов 2010].

Сведение данных, полученных из различных форм электронного контента в пакет информации о психосемантической субъектности позволяет с высокой степенью достоверности обнаружить поведенческие доминанты личности и скрываемые качества, а также принадлежность к социопатической группе [Пономарева, Устюжанин 2016]. Объектом управления становится цифровой двойник, через воздействие на параметры которого можно корректировать поведение, мышление реального человека, интерпретации событий и процессов. Сам двойник непрерывно актуализируется по мере онлайн-активности самого человека. Применение СИИ позволяет осуществлять всю необходимую предиктивную аналитику двойника и его прототипа.

По сути, в настоящее время традиционные методы управления социумом превращаются в единую систему с новыми коммуникационными интерфейсами, нейро- и биоинтерфейсами. Для разработки методов прогнозирования интеллектуальной динамики поведенческой активности накоплен значительный исследовательский опыт [Агеев, Логинов 2017, 2022; Лефевр 2003; Смирнов, Безносюк 1995; Райков 2015; Холодов 1982]. Имеется множество перспективных концепций создания многофункциональной информационной мониторинговой системы как платформы прогнозирования (с обратной связью) явных и неявных глубинных процессов и тенденций в социуме, техносфере и природной среде.

Заключение

В окружающей нас социальной реальности растет явное и латентное присутствие цифровых технологий, прежде всего СИИ. Это создает нарастающий поток новых явлений в массовом сознании, среди них наибольшую опасность представляет манипулятивное использование новых зависимостей человеческого мышления и поведения от информационных и других виртуализированных, киберфизических систем. Практически по экспоненте растут риски дестабилизации социума вследствие как техногенных катастроф, вызванных человеческим фактором, так и деструктивного воздействия с применением цифровых технологий. Новейший опыт показывает феноменальные возможности

современных СМИ и Интернета по влиянию на массовое сознание, политический выбор, повседневность.

ЦИТИРУЕМАЯ ЛИТЕРАТУРА

Агеев, Логинов 2017 – *Агеев А.И., Логинов Е.Л.* Битва за будущее: кто первым в мире освоит ноомониторинг и когнитивное программирование субъективной реальности? // Экономические стратегии. 2017. Т. 19. № 2(144). С. 124–139.

Агеев, Логинов 2019 – *Агеев А.И.*, *Логинов Е.Л.*, *Шкута А.А.*, *Деркач А.К.* Сетевое нейрокогнитивное управление сложноорганизованными структурами с политической компонентой в нечетких информационных средах // Микроэкономика. 2019. Т. 15. № 5(88). С. 5–13.

Агеев, Логинов 2022 – *Агеев А.И., Логинов Е.Л.* Нейроменеджмент личности / 2-е изд. – М.: ИНЭС, 2022.

БРЭ 2008 – Большая Российская энциклопедия. Т. 11. – М.: Большая Российская энциклопедия, 2008.

Волынский-Басманов 2010 – *Волынский-Басманов Ю.М.* Применение методов нейролингвистического программирования для выявления потенциально опасных лиц // Проблемы безопасности и чрезвычайных ситуаций. 2010. № 5. С. 124–128.

Гарднер 2007 – *Гарднер Г.* Структура разума. Теория множественного интеллекта. – М.: Вильямс, 2007.

Дьяков 2015 – *Дьяков С.И.* Психосемантическая модель и техника анализа и оценки субъектности личности // Научная конференция «Ломоносовские чтения – 2015». – М.: МАКС Пресс, 2015. С. 121–122.

К цифре… 2018 – К цифре готов? Оценка адаптивности высокотехнологичного комплекса России к реалиям цифровой экономики / под ред. А.И. Агеева. – М.: ИНЭС, 2018.

Каляев 2019 – *Каляев И.А.* Искусственный интеллект: Камо грядеши? // Экономические стратегии. 2019. Т. 21. № 5(163). С. 6–15.

Кукшев 2020 – *Кукшев В.И.* Классификация систем искусственного интеллекта // Экономические стратегии. 2020. Т. 22. № 6(172). С. 58–67.

Лефевр 2003 – *Лефевр В.А.* Рефлексия. – М.: Когито-Центр, 2003.

Ляхов, Тришин 2013 – *Ляхов А.Ф., Тришин И.М.* Компьютерное моделирование поведения игрока в интеллектуальной карточной игре с помощью нейронной сети // Компьютерные инструменты в образовании. 2013. № 5. С. 54–64.

Осоченко, Макушкин 2019 – *Осоченко Е.А.*, *Макушкин А.Г.* Атлас сквозных технологий цифровой экономики России. – М.: Гринатом, 2019.

Пономарева, Устюжанин 2016 – *Пономарева О.С., Устюжанин В.Н.* О состоянии и перспективах использования психосемантических методов познания личности подозреваемого в деятельности следственного работника // Вестник Санкт-Петербургского университета МВД России. 2016. № 2(70). С. 190–194.

Райков 2015 – *Райков А.Н.* Моделирование коллективного бессознательного при принятии решений // Труды Международной научной конференции СРТ-2014 Московского физико-технического института (государственного университета), Института физико-технической информатики. – М.; Протвино: Институт физико-технической информатики, 2015. С. 146–156.

Райков 2021 – *Райков А.Н.* Гибридный сильный искусственный интеллект // Экономические стратегии. 2021. Т. 23. № 1(175). С. 62–63.

Самарцев, Латенкова 2016 – *Самарцев О.Р., Латенкова В.М.* Психосемантические аспекты восприятия интерактивного дискурса в Интернет-СМИ // Вестник Череповецкого государственного университета. 2016. № 2(71). С. 87–91.

Севостьянов 2014 – Севостьянов Ю.О. Изменение психосемантической структуры готовности работать в команде у студентов // Научный вестник Южного института менеджмента. 2014. № 2. С. 94–97.

Смирнов, Безносюк 1995 – *Смирнов И., Безносюк Е., Журавлев А.* Психотехнологии. Компьютерный психосемантический анализ и психокоррекция на неосознаваемом уровне. – М.: Издательская группа «Прогресс» – «Культура», 1995.

Социально-экономические аспекты… 2020 – Социально-экономические аспекты внедрения искусственного интеллекта / под науч. ред. А.И. Агеева. – М.: Айти Сервис, 2020.

Холодов 1982 – *Холодов Ю.А.* Мозг в электромагнитных полях. – М.: Наука, 1982.

Энциклопедический словарь… 1902 – Энциклопедический словарь Брокгауза и Ефрона. Т. XXXIV-А. – СПб.: Ф.А. Брокгауз, И.А. Ефрон, 1902.

Ясницкий, Сичинава 2011 – *Ясницкий Л.Н., Сичинава З.И.* Нейросетевые алгоритмы анализа поведения респондентов // Нейрокомпьютеры: разработка, применение. 2011. № 10. С. 59–64.

McCarthy 2007 – *McCarthy J.* What Is Artificial Intelligence? // Professor John McCarthy. 2007, Nov. 2. – URL: http://jmc.stanford.edu/articles/whatisai/whatisai.pdf

Raikov 2021 – *Raikov A.* Cognitive Semantics of Artificial Intelligence: A New Perspective. – Singapore: Springer, 2021.

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

REFERENCES

Ageev A.I. (Ed.) (2018) *Ready for the Digital? Assessment of the Adaptability of the High-Tech Complex of Russia to the Realities of the Digital Economy.* Moscow: INES (in Russian).

Ageev A.I. (Ed.) (2020) *Socio-Economic Aspects of the Implementation of Artificial Intelligence.* Moscow: Aiti Servis (in Russian).

Ageev A.I. & Loginov E.L. (2017) Battle for the Future: Who Will Be the First in the World to Master the Noomonitoring and Cognitive Programming of Subjective Reality? *Ekonomicheskie strategii.* Vol. 19, no. 2, pp. 124–139 (in Russian).

Ageev A.I. & Loginov E.L. (2022) *Neuromanagement of Personality.* Moscow: INES (in Russian).

Ageev A.I., Loginov E.L., Shkuta A.A., & Derkach A.K. (2019) Network neurocognitive management of complex organizations with a political component in fuzzy information environments. *Mikroekonomika.* Vol. 15, no. 5, pp. 5–13 (in Russian).

Arseniev K.K. & Petrushevsky F.F. (Eds.) (1902) *Brockhaus and Efron Encyclopedic Dictionary* (Vol. XXXIV-A). Saint Petersburg: F.A. Brochkaus, I.A. Efron (in Russian).

Dyakov S.I. (2015) Psychosemantic Model and Technique of Analysis and Evaluation of the Subjectivity of the Personality. In: *"Lomonosov Readings – 2015" Scientific Conference* (pp. 121–122). Moscow: MAKS Press (in Russian).

Gardner H. (1983) *Frames of Mind: The Theory of Multiple Intelligences.* New York: Basic Books (Russian translation: Moscow: Vil'yams, 2007).

Kalyaev I.A. (2019) Artificial Intelligence: Whither Goest Thou? *Ekonomicheskie strategii.* Vol. 21, no. 5, pp. 6–15 (in Russian).

Kholodov Yu.A. (1982) *The Brain in Electromagnetic Fields.* Moscow: Nauka (in Russian).

Kukshev V.I. (2020) Classification of Artificial Intelligence Systems. *Ekonomicheskie strategii.* Vol. 22, no. 6, pp. 58–67 (in Russian).

Lefebvre V.A. (2003) *Reflection*. Moscow: Kogito-Tsentr (in Russian).

Lyakhov A.F. & Trishin I.M. (2013) Computer Simulation of Player Behavior in an Intellectual Card Game Using a Neural Network. *Komp'yuternye instrumenty v obrazovanii.* No. 5, pp. 54–64 (in Russian).

McCarthy J. (2007) *What Is Artificial Intelligence?* Retrieved from http://jmc.stanford.edu/articles/whatisai/whatisai.pdf

Osipov Y.S. (Ed.) (2008) *Great Russian Encyclopedia* (Vol. 11). Moscow: Bol'shaya Rossiyskaya entsiklopediya (in Russian).

Osochenko E.A. & Makushkin A.G. (2019) *Atlas of End-to-End Technologies of the Digital Economy of Russia.* Moscow: Grinatom (in Russian).

Ponomareva O.S. & Ustyuzhanin V.N. (2016) On the State and Prospects for the Use of Psychosemantic Methods of Cognition of the Personality of a Suspect in the Activities of an Investigative Officer. *The Bulletin of the St.*

Petersburg University of the Ministry of Internal Affairs of Russia. No. 2, pp. 190–194 (in Russian).

Raikov A.N. (2015) Modeling the Collective Unconsciousness in Decision Making. In: *Proceedings of the International Scientific Conference CPT-2014 of the Moscow Institute of Physics and Technology (State University), Institute of Physical and Technical Informatics* (pp. 146–156). Protvino: Institute of Physical and Technical Informatics. (in Russian).

Raikov A.N. (2021a) Hybrid Strong Artificial Intelligence. *Ekonomicheskie strategii*. Vol. 23, no. 1, pp. 62–63 (in Russian).

Raikov A. (2021b) *Cognitive Semantics of Artificial Intelligence: A New Perspective.* Singapore: Springer.

Samartsev O.R. & Latenkova V.M. (2016) Psychosemantic Aspects of the Perception of Interactive Discourse in the Internet Media. *Cherepovets State University Bulletin.* No. 2, pp. 87–91 (in Russian).

Sevostyanov Yu.O. (2014) Change of Psychosemantic Structure of Willingness to Work in a Team of Students. *Scientific Bulletin of the Southern Institute of Management.* No. 2, pp. 94–97 (in Russian).

Smirnov I., Beznosyuk E., & Zhuravlev A. (1995) *Psychotechnologies: Computer Psychosemantic Analysis and Psychocorrection at the Unconscious Level*. Moscow: Progress – Kul'tura (in Russian).

Turing A.M. (1950) Computing Machinery and Intelligence. *Mind.* Vol. 59, no. 236, pp. 433–460.

Volynsky-Basmanov Yu.M. (2010) Application of Neuro-Linguistic Programming Methods to Identify Potentially Dangerous Persons. *Problemy bezopasnosti i chrezvchaynykh situatsiy.* No. 5, pp. 124–128 (in Russian).

Yasnitsky L.N. & Sichinava Z.I. (2011) Neural Network Algorithms for Analyzing the Behavior of Respondents. *Neirokomp'yutery: razrabotka, primenenie.* No. 10, pp. 59–64 (in Russian).

DOI: 10.30727/0235-1188-2022-65-1-44-71
Оригинальная исследовательская статья
Original research article

Фетиш искусственного интеллекта*

Д.И. Дубровский
Институт философии РАН, Москва, Россия

А.Р. Ефимов
ПАО «Сбербанк», Москва, Россия,
Национальный исследовательский технологический университет «МИСиС», Москва, Россия

В.Е. Лепский
Институт философии РАН, Москва, Россия

Б.Б. Славин
Финансовый университет при Правительстве РФ, Москва, Россия

Аннотация

В статье представлены основания, позволяющие констатировать фетиш искусственного интеллекта (ИИ). Выделяются принципиальные отличия ИИ от всех предшествующих технологических инноваций, связанные прежде всего с внедрением в когнитивную сферу человека и принципиально новыми неконтролируемыми последствиями для общества. Представлены убедительные аргументы того, что лидеры глобалистского проекта являются главными интересантами и заказчиками фетиша ИИ. Это отчетливо проявляется в работах философов, приближенных к гигантским ИТ-корпорациям, и в мегапроектах этих корпораций. Предлагается к рассмотрению проблема использования возможности ИИ для преодоления нарастающих международных конфликтов и в целом мирового кризиса. В центре внимания оказывается вопрос субъектности, решение которого с позиций антропоморфного подхода к ИИ чревато серьезными негативными последствиями. При наделении субъектностью ИИ неявно снимается ответственность с человека, который применяет эту технологию, а также разрушается сложившаяся законодательная практика. Предлагается представление ИИ как агента, наделенного набором инвари-

* Работа поддержана Российским научным фондом (РНФ), грант № 21-18-00184 «Социогуманитарные основания критериев оценки инноваций, использующих цифровые технологии и искусственный интеллект».

антных упрощенных качеств, которыми обладают естественные субъекты. Среди этих качеств – способность к целеустремленности, своего рода рефлексивность, коммуникативность и упрощенные элементы социальности. Такое представление ИИ как агента (псевдосубъекта) согласуется с принципом распределенного управления в биологии и психологии, который был назван принципом двойного субъекта. В сочетании с системами принципов и онтологий, задаваемых в концепции постнеклассической кибернетики саморазвивающихся сред, это позволит использовать ИИ как средство социальных инноваций при сохранении контроля над технологиями ИИ, а также ставить и решать проблему интеграции образований искусственного и естественного интеллекта при сохранении базовых качеств носителей естественного интеллекта.

Ключевые слова: философия искусственного интеллекта, естественный интеллект, глобалистский проект, антропоморфный подход, субъект, псевдосубъект, постнеклассическая кибернетика.

Дубровский Давид Израилевич – доктор философских наук, профессор, главный научный сотрудник Института философии Российской академии наук.
ddi29@mail.ru
https://orcid.org/ 0000-0003-4392-2526

Ефимов Альберт Рувимович – кандидат философских наук, директор Управления исследований и инноваций ПАО «Сбербанк», заведующий кафедрой инженерной кибернетики Национального исследовательского технологического университета (НИТУ) «МИСиС».
makkawity@gmail.com
https://orcid.org/0000-0001-6857-8659

Лепский Владимир Евгеньевич – доктор психологических наук, главный научный сотрудник сектора междисциплинарных проблем научно-технического развития Института философии РАН.
VELepskiy@mail.ru
https://orcid.org/0000-0002-6893-0234

Славин Борис Борисович – доктор экономических наук, профессор департамента бизнес-информатики Финансового университета при Правительстве РФ.
bbslavin@gmail.com
https://orcid.org/0000-0003-3465-0311

Для цитирования: *Дубровский Д.И., Ефимов А.Р., Лепский В.Е., Славин Б.Б.* Фетиш искусственного интеллекта // Философские науки. 2022. Т. 65. № 1. С. 44–71. DOI: 10.30727/0235-1188-2022-65-1-44-71

The Fetish of Artificial Intelligence*

D I. Dubrovsky
Institute of Philosophy, Russian Academy of Sciences, Moscow, Russia

A.R. Efimov
PJSC Sberbank, Moscow, Russia,
National University of Science and Technology MISiS, Moscow, Russia.

V.E. Lepskiy
Institute of Philosophy, Russian Academy of Sciences, Moscow, Russia

B.B. Slavin
Financial University under the Government of the Russian Federation, Moscow, Russia

Abstract

The article presents grounds for defining the fetish of artificial intelligence (AI). We highlight the fundamental differences of AI from all earlier technological advances, as they are primarily related to its introduction into the human cognitive sphere and generating fundamentally new uncontrollable consequences for society. We provide solid evidence that the leaders of the globalist project are the main beneficiaries of the AI fetish. This is clearly manifested in the works of philosophers who are close to major technology corporations and their mega-projects. We suggest considering the problem of how to use the capabilities of AI to overcome the growing international conflicts and the global crisis. The focus is on the problem of agency, which solution from the standpoint of an anthropomorphic approach to AI is fraught with serious negative consequences. Endowing AI with agency, responsibility is implicitly removed from the person who uses the technology, and the established legislative practice is also destroyed. We present AI as an agent endowed with a set of invariant generalized qualities that is similar to natural subjects. These qualities include: the ability to deliberation, reflexivity, communication and elements of sociability. Such a representation of AI as an

* The work was supported by the Russian Science Foundation, grant no. 21-18-00184 "Social and humanitarian foundations for evaluation criteria for innovations based on digital technologies and artificial intelligence."

agent (pseudo-subject) is consistent with the principle of distributed control in biology and psychology, which was called the principle of a dual subject. In combination with the systems of principles and ontologies specified in the concept of post-nonclassical cybernetics of self-developing environments, this will allow the use of AI as a means of social innovation, while maintaining control over AI technologies. This will also help to pose and solve the problem of integrating formations of artificial and natural intelligence while maintaining the basic qualities of carriers of natural intelligence.

Keywords: philosophy of artificial intelligence, globalist project, anthropomorphic approach, subject, pseudo-subject, post-nonclassical cybernetics.

David I. Dubrovsky – D.Sc. in Philosophy, Professor, Chief Research Fellow, Department of Theory of Knowledge, Institute of Philosophy, Russian Academy of Science.
ddi29@mail.ru
https://orcid.org/ 0000-0003-4392-2526

Albert R. Efimov – Ph.D. in Philosophy, Vice-President of Innovation and Research, PJSC Sberbank; Head of the Department of Engineering Cybernetics, National University of Science and Technology MISiS.
makkawity@gmail.com
https://orcid.org/0000-0001-6857-8659

Vladimir E. Lepskiy – D.Sc. in Psychology, Chief Research Fellow, Department of Interdisciplinary Problems in the Advance of Science and Technology, Institute of Philosophy, Russian Academy of Science.
VELepskiy@mail.ru
https://orcid.org/0000-0002-0590-4020

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: Dubrovsky D.I., Efimov A.R., Lepskiy 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.
DOI: 10.30727/0235-1188-2022-65-1-44-71

Введение

По мере того, как экономика и социальные коммуникации оцифровываются, возрастает и роль цифровых технологий. Тех-

нологические инновации всегда привлекали к себе внимание: так было и во времена появления паровых машин, и в период электрификации, и в эпоху расцвета электроники. Появление новых технологий приводило к трансформации экономических и общественных отношений, и поэтому часто технологиям приписывали субъектные возможности по переустройству мира. Однако время все расставляет на свои места, и предсказания наподобие «Со временем, телевидение перевернет жизнь всего человечества. Ничего не будет: ни кино, ни театра, ни книг, ни газет – одно сплошное телевидение», как правило, не оправдываются. Технологии находят свою нишу в общей экосистеме научного прогресса и не ломают привычных устоев. Поэтому всегда необходим элемент скепсиса при появлении новых революционных технологий.

Изложенный подход можно было бы применить и к современным цифровым технологиям, долго не обсуждая вопрос о возможных трансформациях, к которым они приводят. Однако цифровые технологии имеют особенность, которая их существенно отличает от остальных технологий. Все новшества, появившиеся в доцифровую эпоху, были призваны повысить производительность труда человека либо облегчить и сделать более комфортной его жизнь. Они преобразовывали окружающую среду, делая ее более удобной для человека, но при этом не затрагивали его сущность и когнитивные возможности. Конечно, совершенствование орудий труда вело к тому, что человечество накапливало новые знания, но эти знания по-прежнему находились в головах людей. Книги были лишь инструментом передачи знаний от человека к человеку, но без человека они не имеют смысла.

Цифровая эпоха дала возможность накапливать знания в цифровом виде. В начале это, как и любое нововведение, лишь облегчало использование знаний человеком, поскольку стало возможным читать книги и статьи с электронных носителей, искать в них информацию. Но по мере развития технологий обработки данных оказалось, что аналитические программы способны настолько глубоко анализировать данные (т.н. *data mining*), что можно найти в них то, чего не обнаружишь непосредственно в тексте или в данных. Впервые технологии покусились на самое святое – на когнитивные возможности человека. Если технологии смогут хотя бы частично замещать мыслительные способности

человека, не означает ли это, что технологии могут получить ту самую субъектность по трансформации общества, которой они ранее никогда не обладали? Этот вопрос сегодня в той или иной мере будоражит умы многих ученых.

Нарастающая волна интереса к ИИ и предвестники шторма

Благодаря увеличению производительности вычислительной техники, успехам в области разработки алгоритмов использования искусственных нейронных сетей (в первую очередь машинного обучения, в особенности глубокого) и появлению инструментов для работы с большими данными технологии искусственного интеллекта (ИИ) перешли из разряда перспективных в разряд самых востребованных технологий. Стоит обратить внимание на то, что существующие технологии ИИ в основном направлены на выполнение задач распознавания, предсказаний, имитации человеческой деятельности и не претендуют на создание реального аналога человеческого интеллекта.

Текущие успехи ограничены решением проблем определенных классов: ИИ все еще нельзя доверить самостоятельное принятие сложных решений, от которых зависит жизнь человека. Однако даже в таком виде ИИ стал широко использоваться в предиктивной аналитике, скоринге, распознавании лиц и игровых приложениях. В 2022 году по прогнозу *IDC* крупнейшие мировые компании вложат более 430 млрд долларов США в исследования и разработки в области ИИ[1]. При этом мировой рынок ИИ технологий составит 554,3 млрд долл. США к 2024 году при среднегодовом темпе роста 17,5 % [Прорывные инновации… 2022, 44].

Одновременно с ростом успеха ИИ началась череда принятия стратегий его развития на национальных уровнях. В 2016 году в США Национальным советом по науке и технологиям подготовлены доклады «Подготовка к будущему искусственного интеллекта», «Национальный стратегический план исследований и разработок в области искусственного интеллекта» и «Искусственный интеллект, автоматизация и экономика», определившие стратегию развития ИИ в стране. В июле 2017 года Государствен-

[1] IDC Forecasts Companies to Increase Spend on AI Solutions by 19.6% in 2022 // IDC. 15 February 2022. – URL: https://www.idc.com/getdoc.jsp?containerId=prUS48881422

ный совет Китая опубликовал «План развития искусственного интеллекта нового поколения», рассчитанный на создание индустрии ИИ и превращение Китая в ведущую державу в области ИИ к 2030 году. В 2018 году, на ежегодной встрече в рамках Всемирного экономического форума, премьер-министр Великобритании Тереза Мэй объявила, что собирается сделать «Великобританию мировым лидером в области искусственного интеллекта». В 2019 году в России Президентом В.В. Путиным была утверждена «Национальная стратегия развития искусственного интеллекта на период до 2030 года», а позднее, в рамках программы «Цифровая экономика», выделен проект по развитию ИИ. В последние годы набирает оборот новый мировой тренд – разработка общего искусственного интеллекта, который по своим функциям должен приблизиться к решению задач, специфичных для естественного интеллекта, о чем пойдет речь далее в статье.

Блицкриг ИИ в бизнесе и на государственной ниве совпал с протестами и требованиями об ограничении использования инструментов распознавания, которые нарушали права личности и несли риски принятия ошибочных решений. Так, европейские страны уже при подписании в апреле 2018 года Декларации о сотрудничестве в области искусственного интеллекта делали акцент на необходимости учета социальных, экономических, этических и юридических вопросов. В частности, Европейская комиссия учредила Группу экспертов, которые опубликовали руководящие принципы этики ИИ в апреле 2019 года. В сентябре 2021 года Совет ОНН по правам человека выпустил доклад с рекомендациями «для государств и предприятий относительно разработки и внедрения гарантий для предотвращения и сведения к минимуму вредных последствий и содействия полному использованию преимуществ, которые может предоставить искусственный интеллект» [The Right to Privacy… 2021]. Руководитель Совета (верховный комиссар ООН по правам человека) Мишель Бачелет, комментируя доклад, призвала ввести мораторий на системы ИИ, которые угрожают правам человека, до тех пор, пока правительства не смогут установить гарантии[2].

[2] Мишель Бачелет призвала ввести мораторий на использование систем искусственного интеллекта // Новости ООН. 15 сентября 2021. – URL: https://news.un.org/ru/tags/iskusstvennyy-intellekt/date/2021-09

В докладе Совета ООН по правам человека, в частности, говорится о том, что «процессы принятия решений во многих системах искусственного интеллекта непрозрачны. Сложность информационной среды, алгоритмов и моделей, лежащих в основе разработки и функционирования систем искусственного интеллекта, а также преднамеренная секретность государственных и частных субъектов являются факторами, которые подрывают значимые способы понимания общественностью вопросов влияния систем искусственного интеллекта на права человека» [The Right to Privacy… 2021]. Интересен тот факт, что именно машинное обучение вызывает у защитников прав человека наибольшие опасения, поскольку результат таких вычислений непредсказуем: «Системы машинного обучения добавляют важный элемент непрозрачности; они могут быть способны выявлять закономерности и разрабатывать рецепты, которые трудно или невозможно объяснить. Это часто называют проблемой “черного ящика”. Непрозрачность затрудняет тщательное изучение системы искусственного интеллекта и может стать препятствием для эффективной подотчетности в тех случаях, когда системы искусственного интеллекта наносят ущерб» [The Right to Privacy… 2021]. Кроме того, вызывает большую озабоченность то обстоятельство, что любое машинное обучение очень сильно зависит от тех данных, на которых происходит само обучение. Если эти данные не проверять на соблюдение элементарных этических норм или, что еще хуже, специально подобрать их, чтобы они им противоречили, то системы машинного обучения неизбежно будут выдавать заведомо неэтичные рекомендации.

Главные интересанты и заказчики фетиша ИИ

В этом году в одном из зарубежных изданий опубликована статья «К теории справедливости для искусственного интеллекта» Ясона Габриэля [Gabriel 2022]. Автор работает штатным философом в компании *DeepMind*, принадлежащей *Google*. В статье речь идет о разработке гуманистических принципов использования технологий ИИ. Проанализируем философско-методологические основания позиции автора, их влияние на представления о потенциальных последствиях развития ИИ, а также, что особенно важно, о главных интересантах предложенного подхода.

В первую очередь обратим внимание на идеологическую установку автора статьи, опирающегося на труды Джона Ролза, посвященные его теории справедливости [Rawls 1999]. Суть данной установки отражена в следующем: люди заинтересованы в увеличении своей и уменьшении общей доли выгоды. Это – ярко выраженная позиция идеологии социального либерализма. Для того, чтобы сформировать представление о роли и месте ИИ, автор предлагает понимать основную структуру общества как совокупность социотехнических систем, функционирование которых складывается под все возрастающим влиянием ИИ. Такое представление об ИИ может быть охарактеризовано как яркое проявление технократического редукционизма в организации социальных процессов. В качестве следствия неявно предлагается нарастающее под влиянием ИИ снижение роли государства и общества. Интересантами таких результатов являются лидеры глобалистского проекта. Но кто именно? Это следует конкретизировать.

Лидерами глобалистского подхода и яркими выразителями, защитниками идеологии западного либерализма выступают в первых рядах именно владельцы гигантских ИТ-корпораций, которые, подобно спруту, охватывают мировое коммуникативное пространство. Владельцы таких компаний, как *Facebook*, *Twitter*, *Amazon*, *Google* и др., их наемные теоретики, приписывая системам ИИ *качество субъектности*, пытаются создать впечатление о том, что они наделены некими «божественными» функциями. Они же не просто разрабатывают новые программные продукты, а создают инструменты, которые способны якобы устанавливать справедливость и защищать права человека. Такие инструменты – утопия. Но прокламирование веры в них, ее имитация служит благодатной почвой для защиты глобалистских идей, а в то же время и сокровенных интересов крупнейших мировых производителей ИТ-технологий. Манипулируя массовым сознанием, используя все формы социальных коммуникаций, все средства ИИ, они заявляют о том, что будут для всех нас «сеять разумное, доброе, вечное». Неслучайно основатель *Facebook* Марк Цукерберг решил на базе своих ИТ-продуктов создавать т.н. метавселенную, в которой будет построена новая, более правильная и благополучная жизнь. Однако крайне трудно допустить мысль о том, что Марк

Цукерберг действительно озадачен настолько высокими гуманистическими устремлениями. Судя по его бизнес-деятельности и его опыту манипуляции массовым сознанием в целях достижения максимальной прибыли, сказки о «метавселенной» служат лишь маскировкой для осуществления этих же целей.

Глобалистский подход, опирающийся на идеологию западного либерализма с его псевдогуманистическими клише, убедительно показал свою несостоятельность в условиях пандемии COVID-19 [Лепский 2020] и тем самым продемонстрировал свою негативную роль в деле оценки и разработки способов преодоления нарастающих угроз для человечества, в том числе угроз, связанных с цифровыми трансформациями и развитием ИИ.

Перспективы ИИ в контексте глобального кризиса мировой цивилизации

Именно этот контекст рассмотрения процесса развития ИИ и его социальной значимости, отодвигаемый часто на дальний план (особенно в публикациях, подобных упомянутой выше), приобретает сегодня первостепенное, судьбоносное значение для человечества. Неуклонное нарастание глобального кризиса нашей потребительской цивилизации ведет к разжиганию все более масштабных экономических, политических, всевозможных социальных конфликтов, бескомпромиссной борьбы за ресурсы, за сферы влияния и в конечном итоге – за передел мировой структуры социально-экономической и политической самоорганизации в целом. В этой связи ИИ становится инструментом борьбы в противостоянии различных сил.

Особое внимание привлекает задача создания общего искусственного интеллекта (ОИИ). Ее решение стало в последние годы предметом конкуренции между крупнейшими бигтехами, а в более широком понимании – между государствами-лидерами в области ИИ, в числе которых находятся наши стратегические противники. Это обязывает нас максимально сконцентрировать усилия в данном направлении и добиваться опережения конкурентов. В прошлом году в России вышла первая книга, посвященная специфической проблематике ОИИ [Сильный искусственный интеллект… 2021]. В ней подробно проанализированы главные теоретические и методологические вопросы в области разработки

ОИИ и необходимые для этого научные подходы, обозначены его специфические функции, которые должны быть созданы и, что особенно интересно, предполагаемые масштабные перемены, которые он способен произвести в нарождающемся мироустройстве.

Авторы выделяют две главные способности ОИИ, которые характерны для естественного интеллекта. В отличие от т.н. узкого ИИ, решающего одну определенную задачу, он должен быть интегральным, т.е. способным решать много разных задач. И он должен обрести качество автономности, т.е. способность самостоятельно и эффективно действовать в широком диапазоне сред. Все это качественно повышает деятельные возможности систем ИИ, их использование для военной техники и военных действий, для решения задач производства, управления, планирования, организации экономических процессов, оптимизации самых разнообразных сфер общественной жизни и научных исследований, что чрезвычайно важно для нашей страны в нынешних условиях. Это не менее значимо для осмысления и осуществления грядущих исторических изменений в развитии земной цивилизации, связанных с крушением принципа и практики монополярного мира.

Вместе с тем развитие ОИИ ставит новые сложные теоретические вопросы о его взаимодействии с естественным интеллектом, создании и перспективах гибридного интеллекта, возможном состязании с естественным интеллектом, возможных рисках и угрозах для человека и общества. Достижение высокой степени автономности общим интеллектом создает вероятность появления таких видов и способов его «самодеятельности», которые могут представлять опасность для человека и общества, потребуют разработки новых методов обеспечения безопасности. Перед нами окажется новый аспект все той же проблемы *субъектности* систем ИИ, сохраняющей свою высокую актуальность.

ИИ – это субъект или агент, контролируемый субъектами естественного интеллекта?

Обратимся еще раз к истолкованиям теоретических вопросов о субъектности систем ИИ по отношению к реальным человеческим субъектам. Подход Ясона Габриэля, автора упомянутой выше статьи [Gabriel 2022], заключается в том, чтобы перенести на

ИИ такие же принципы, которые установлены для людей. Автор пишет о том, что «ИИ все больше формирует элементы базовой структуры общества», и, «следовательно, его проектирование, разработка и развертывание потенциально взаимодействуют с принципами правосудия». По мнению Габриэля, «ИИ взаимодействует с поведением людей, принимающих решения, и формирует характер этих практик, включая распределение выгод и бремени среди населения» [Gabriel 2022].

Таким образом, у автора технологии будто «оживают». Он пишет: «Для нашей цели важно учесть, что в современных обществах фоновая справедливость все больше осуществляется алгоритмически» [Gabriel 2022]. Вводя понятие фоновой (видимо, массовой) справедливости, неявно предполагается, что эту справедливость осуществляет алгоритм. Автор продолжает: «Делая оценки или прогнозы на основе прошлого выбора человека и предоставляя решения или рекомендации, которые затем формируют набор возможностей, доступных этому человеку в будущем, эти системы сильно влияют на разворачивающуюся взаимосвязь между индивидуальным выбором и коллективными результатами» [Gabriel 2022]. Снова неявно предполагается, как системы что-то «разумно» делают и что-то предоставляют. Автор придает субъектность технологиям ИИ и требует распространить законы социальных отношений на использование ИИ, которое должно «поддерживать основные свободы граждан, способствовать справедливому равенству возможностей и приносить наибольшую пользу тем, кто находится в наихудшем положении» [Gabriel 2022].

Такого рода антропоморфный подход к ИИ чреват существенными последствиями. Перенося субъектность на ИИ, неявно снимается ответственность с человека, который применяет эту технологию, что нивелирует законодательную практику. Любая технология несовершенна, и она не может сама по себе принимать решение. Ошибка в работе детектора лжи никогда не будет равна нулю, как и при использовании ИИ. Задача людей состоит в том, чтобы учесть ограничения технологий, а не пытаться просто их поставить в какие-то заданные рамки. Неслучайно в судебной практике при всех возможностях криминалистики окончательное решение принимают люди. Системы ИИ ничем не отличаются

от других технологий. Поэтому технологии не должны снимать ответственность с человека, а следовательно, не должны обладать человеческими характеристиками, т.е. быть справедливыми, гуманными и т.д.

Успехи ИИ оказались настолько значительными, что многие решили, что эта технология может заменить человека, только нужно ее поставить в определенные рамки. Фактически ИИ стал фетишем XXI века, который одни стали превозносить как наше будущее, а другие начали вести с ним борьбу. В действительности ИИ, хотя и может распознать то, что не удается человеческому взгляду, или выявить корреляцию, которую не может найти человек, вместе с тем остается далеким от реальных когнитивных возможностей человека. Не исключено, что в будущем удастся создать ИИ, интегрированный в социальную среду, т.н. сильный ИИ, но пока мы далеки от этого, и придавать ему субъектность не только неправильно, но опасно.

Этические вопросы развития ИИ

Вопросы этики использования ИИ широко обсуждаются сегодня как общественными деятелями, политиками, так и учеными. Большое внимание привлек скандал, возникший в связи со статьей об этике ИИ, подготовленной к публикации сотрудником компании *Google* Тимнит Гебру, одним из ведущих мировых экспертов по проблемам необъективности алгоритмов и извлечения данных (*data mining*) [Hao 2020]. Научные и коммерческие позиции принципиально разошлись, и, как следствие, исследовательница покинула компанию *Google*.

Дискуссии ведутся и среди российских ученых. Наряду с конструктивными предложениями отдельные авторитетные ученые предлагают сомнительные идеи, призывают к политизации науки и технологий, наделяя технологии свойствами «патриотизма». Так, 23 ноября 2021 года состоялось заседание Президиума Российской академии наук, на которой академики обсуждали в том числе и этическую сторону ИИ, и возможности для ИИ быть доверенным. Один из выступающих заявил: «С моей точки зрения, ИИ должен быть не только доверенным, о чем сегодня говорилось, но и патриотичным, т.е. он должен в первую очередь работать на интересы страны, а не против нее» [Славин 2021, 34].

Сегодня многие протестуют против технологий распознавания лиц (в некоторых городах США такое распознавание запрещено, Европарламент тоже предложил запретить технологии распознавания лиц). Однако вредно не распознавание как таковое, а использование его в противоправных целях. Человек не скрывает свои болезни перед врачом потому, что доверяет ему и рассчитывает на помощь. Необходимо регулировать законы применения технологий, в том числе и ИИ, серьезно наказывать, если они были использованы во вред человеку или незаконно. Россия сегодня оказалась во многом слабо защищенной перед мошенниками, которые используют средства коммуникаций для обмана доверчивых граждан. Все, что сегодня власть может сделать, – это предупреждать население о новых способах мошенничества. При этом власть вполне эффективно, в том числе и с использованием ИИ, борется с политическими противниками. Этические проблемы ИИ должны решать не путем ограничения технологий, а в первую очередь путем ограничения действий людей, которые их неправомерно используют.

Решение этических вопросов применения ИИ – очень сложное, комплексное дело, которое невозможно выполнить лишь посредством принятия декларативных кодексов этики ИИ. Необходимо учитывать, что даже наши этические принципы – от библейских заповедей до кодекса строителя коммунизма – есть лишь точки в бесконечном пространстве морально-этических решений, в котором мы движемся ежедневно. Дискуссии об этике ИИ только начинают разворачиваться, и сообщества философов, инженеров должны тесно сотрудничать для выработки ответов на вызовы времени.

Сегодня крайне важно рассматривать этические проблемы применительно к развитию ИИ в более широком концептуальном плане: под углом реальных особенностей функционирования этических норм в социальной жизнедеятельности, реального состояния нравственности массового сознания, индивидуальных, групповых, институциональных субъектов. Сплошь и рядом, всегда, на всех этапах истории человечества ясно наблюдался разрыв между знанием этических норм и их исполнением. Вспомним древнеримскую поговорку: «Вижу лучшее и одобряю, но следую худшему». Слишком часто интерес оказывался выше нравственных установлений, а обман подавлял правду и вил себе уютные

гнезда в самых высоких этических наставлениях. Говорить о нравственном прогрессе в развитии человечества нет достаточных оснований (обстоятельные материалы, посвященные данной теме, представлены во множестве философских исследований [Дубровский 2007]). Все эти обстоятельства следует учитывать, если мы рассуждаем на тему «Этика ИИ», причем как в отношении создателей систем ИИ, так и в отношении пользователей.

При попытках моделирования принципов этики и воплощения их в работе ИИ ситуация осложняется тем, что совокупность этических норм не может быть упорядочена в виде четкой иерархической структуры, допускающей альтернативный выбор. Выбор практически всегда может быть сделан лишь при рассмотрении и оценке конкретных условий. Поэтому указанное моделирование представляется возможным только в специально определенных частных случаях.

Вместе с тем проблема субъектности в области разработок ИИ по-прежнему заслуживает пристального внимания. Так или иначе способность системы ИИ решать сложные задачи мы связываем с описаниями некоторых функций естественного интеллекта. Если ИИ нецелесообразно представлять в качестве субъекта, аналогичного человеку, то как понимать и определять ИИ, которому передаются возможности принятия решений в определенных ситуациях и который побеждает чемпиона мира по шахматам или игры в го? Наиболее адекватным подходом, на наш взгляд, может быть представление ИИ как агента, наделенного набором инвариантных упрощенных качеств, которыми обладают естественные субъекты. К таким качествам можно отнести подобие целеустремленности, своего рода рефлексивность, коммуникативность и упрощенные элементы социальности. Представление ИИ как агента (псевдосубъекта) согласуется с принципом распределенного управления в биологии и психологии, который был назван принципом двойного субъекта [Лепский 1998]. Это представление ИИ в сочетании с системами принципов и онтологий, задаваемых в концепции постнеклассической кибернетики саморазвивающихся сред, позволяет использовать ИИ как средство социальных инноваций, при сохранении контроля над технологиями ИИ, а также ставить и решать проблему интеграции образований искусственного и естественного интеллекта при

сохранении базовых качеств носителей естественного интеллекта [Lepskiy 2018; Лепский 2021].

Заключение

ИИ переживает фазу бурного роста. Масштабные цифры эффектов от внедрения не должны вводить нас в заблуждение: нынешний период – это лишь начало тотального проникновения ИИ в нашу жизнь. Именно поэтому нам следует очень внимательно относиться к возможным когнитивным искажениям при исследованиях возникающих феноменов. К примеру, антропоморфизм в применении к ИИ может заставить нас легко поверить в ложную субъектность машины. Настоящая работа авторов, специалистов в философии и методологии, служит призывом к более широкому диалогу и переходу от создания кодексов поведения ИИ к созданию следующего поколения ИИ, действующего вместе с человеком и для человека.

The Fetish of Artificial Intelligence

Introduction

As currently economy and social communications are undergoing digitalization, the role of digital technologies also increases. Technological innovations have always attracted attention. This was the case during the advent of steam engines, later during electrification, and in the heyday of electronics. The emergence of new technologies led to a transformation of economic and social relations. Due to this, technology was considered to possess capabilities of restructuring the world. However, time puts everything in its place, and the prediction from the classical Soviet movie script has not come true: “Over time, television will change the life of all mankind. There will be nothing else: no cinema, no theater, no books, no newspapers. Only television.” Technologies fill their niches in the general ecosystem of scientific progress but do not destroy the traditional bases. Therefore, we should use some skepticism to new revolutionary technologies.

Such an approach could be applied to modern digital technologies, thus postponing any lengthy discussions they may lead to. However,

digital technologies have one feature that significantly distinguishes them from all others. All the innovations that appeared in the pre-digital era were designed to increase human labor productivity or to make our life easier and more comfortable. They transformed the social environment, making it more convenient for humans, but at the same time they did not affect the very essence of human beings or their cognitive capabilities. Of course, the improvement of tools led to accumulation of new knowledge, but this knowledge still remained in the people's minds. Books were just tools for transferring knowledge from person to person.

Our epoch has made it possible to accumulate knowledge in digital form. At first this only facilitated personal use of knowledge: it became possible to read books and articles in electronic formats and to search for information. But with the development of data processing technologies, it turned out that analytical software could analyze data in such a way (the so-called data mining) that made it possible to find what was not directly presented in the text or in the data. For the first time in the history of mankind, technology encroached on the most sacred thing: man's cognitive capabilities. If technology can replace (even partially) the ability of a human to think, does this not mean that technology can get the very agency for the transformation of society, which they have never possessed before? This question today stirs the minds of many scientists in various degrees.

Increasing interest in AI and harbingers of a storm

Thanks to the increasing performance rates in modern computing, advances in the development of algorithms for artificial neural networks (primarily machine learning and especially deep learning) and the emergence of tools for working with big data, artificial intelligence (AI) technologies have outgrown the category of promising technologies, to enter the category of the most popular ones. The existing AI technologies mainly focus on tasks of recognition, prediction and imitation of human activity, and do not claim to createa real analogue of human intelligence.

The current successes are limited to solving problems of certain classes, and AI still cannot be trusted to independently make complex decisions on which life depends. However, even in this form, AI is now widely used in predictive analytics, scoring, face recognition

and game applications. IDC predicts that in 2022, the world's largest companies will invest more than 430 billion US dollars in AI research and development[3]. Also, the global AI technology market will amount to 554.3 billion US dollars by 2024 with an average annual growth rate of 17.5% [Gokhberg, Efimov, & Milshina 2022, 44].

Simultaneously with the growing success of AI, adoption of strategies for its development at national levels began. In 2016, the U.S. National Science and Technology Council prepared the reports "Preparing for the future of Artificial Intelligence," "National Artificial Intelligence Research and Development Strategic Plan," and "Artificial Intelligence, Automation, and the Economy," which determined the strategy for AI development in the nation. In July 2017, China's State Council issued a document entitled "A New Generation Artificial Intelligence Development Plan," designed to create an artificial intelligence industry and to turn China into a leading power in the field by 2030. At the 2018 annual meeting of the World Economic Forum, British Prime Minister Theresa May announced that she was going to make the UK a world leader in artificial intelligence. In 2019, Russian President V.V. Putin approved the "National Strategy for the Development of Artificial Intelligence for the period up to 2030," and then a project for the development of AI was drafted within the framework of the Digital Economy program. In recent years, a new global trend has been gaining popularity – the development of Artificial General Intelligence (AGI), whose functions approach the solution of tasks that are specific to natural intelligence. We discuss these issues in more detail below.

Simultaneously with the breakthrough of AI in business and government programs, there started protests and demands to limit the use of recognition tools that violated individual rights and posed risks of making wrong decisions. Thus, already at the signing of the Declaration of cooperation on Artificial Intelligence in April 2018, European states emphasized the need to take into account various social, economic, ethical and legal issues. In particular, the European Commission established a Group of Experts who published guidelines on AI ethics in April 2019. In September 2021, the UN Human Rights Council released a report with recommendations "for states and busi-

[3] IDC Forecasts Companies to Increase Spend on AI Solutions by 19.6% in 2022. *IDC*. 2022, February 15. Retrieved from https://www.idc.com/getdoc.jsp?containerId=prUS48881422

nesses to develop and implement safeguards to prevent and minimize harmful effects and promote the full use of the benefits that artificial intelligence can provide" [OHCHR 2021]. The head of the Council (UN High Commissioner for Human Rights) Michelle Bachelet, commenting on the report, called for a moratorium on artificial intelligence systems that threaten human rights, until governments can provide guarantees[4].

The report of the UN Human Rights Council states that decision-making processes in many AI systems are not transparent. The complexity of the informational environment, algorithms, and models underlying the development and operation of artificial intelligence systems, as well as deliberate secrecy of public and private actors are factors that undermine ways for the public to understand the impact of AI systems on human rights. Interestingly, it is machine learning that causes the greatest concern among defenders of human rights, since the result of such computations is unpredictable: machine learning systems will inevitably reduce essential transparency; they may be able to identify patterns and develop recipes that are difficult or impossible to account for. This is often referred to as the "black box" problem. Insufficient transparency makes an AI system hard for examination and may hinder effective accountability in cases where AI systems cause damage [OHCHR 2021]. Besides, it is a very sensitive issue that all machine learning depends very much on the data fed during the training itself. If such data are not checked for compliance with basic ethical values, or even worse, if they are specially selected so that they contradict the values, then machine learning systems will inevitably issue deliberately unethical recommendations.

The main stakeholders and beneficiaries of the AI fetish

Recently there was a paper by Iason Gabriel, "Toward a Theory of Justice for Artificial Intelligence" (the author is a Staff Research Scientist and an expert in philosophy at DeepMind, a company owned by Google) [Gabriel 2022]. The publication is devoted to the development of humanistic principles of using AItechnologies. Let us analyze the

[4] Urgent Action Needed over Artificial Intelligence Risks to Human Rights. *UN News.* 2021, September 21. Retreved from https://news.un.org/en/story/2021/09/1099972

philosophical and methodological foundations of the author's position, their influence on the ideas about the potential consequences of AI development, as well as the principal stakeholders of the proposed approach.

First of all, let us look at the ideological stance of the author, based on John Rawls's theory of justice [Rawls 1999]. The essence of this approach is depicted as follows: people are interested in increasing their own profits and decreasing the common share of benefits. This is a pronounced position of the ideology of social liberalism. To present the role of AI, the author treats the basic structure of society as a set of sociotechnical systems, whose functioning develops under the increasing influence of AI. This presentation of AI is a vivid manifestation of technocratic reductionism in the organization of social processes. Here, as a consequence, the author proposes a growing decline in the role of the state and society under the influence of AI. The leaders of the globalist project are interested in such results. But who are they, exactly? This needs specifying.

It is easy to see that the leaders of the globalist approach and the active proponents and defenders of the ideology of Western liberalism are the owners of huge IT corporations who, like an octopus, command the global informational space. The owners of such companies as Facebook, Twitter, Amazon, Google and their hired theorists, attributing the quality of agency to AI systems, seek to create the impression that they are endowed with certain "god-like" functions. After all, they do not just develop new software products but create tools that are supposedly able to establish justice and protect human rights. These tools are utopic. But proclaiming faith in them or imitation of such faith provides a good reason for protecting globalist ideas, and at the same time, the innermost interests of the world's largest IT manufacturers. Manipulating public consciousness, using all forms of social communication, and all AI tools, they declare that they will strive for common benefit. It is no coincidence that Facebook's founder Mark Zuckerberg decided to create a Metaverse based on his IT products, in which a new, "better arranged and prosperous" life will be built. However, it is extremely difficult to assume that Zuckerberg is really concerned about such high humanistic aspirations. Judging by all his business activities and his experience of manipulating mass consciousness in order to achieve a maximum profit, the tales of the "Metaverse"

serve only as a disguise for the implementation of the same mercenary goals.

The globalist approach, based on the ideology of Western liberalism with its pseudo-humanistic cliches, convincingly showed its inconsistency in the COVID-19 pandemic [Lepskiy 2020] and demonstrated its negative role in assessing and developing ways to overcome the growing threats to humanity, including threats related to digital transformation and the development of AI.

Prospects of AI in the context of the global crisis of the world civilization

It is this context of considering the AI development and its social significance, often pushed to the background (especially in publications of the type we discussed above), that is now of paramount and fateful importance for humanity. The ongoing global crisis of our consumer civilization leads to emergence of increasingly large-scale economic, political, and social conflicts, uncompromising struggle for resources, for spheres of influence, and ultimately for a redistribution of the entire global structure of socio-economic and political self-organization. And in this regard, AI becomes a weapon in the confrontation of various forces.

In this regard, as already noted above, special attention is drawn to the task of creating an AGI, the solution of which has become in recent years the subject of competition between the largest hightech companies, and also between the states leading in the field of AI, among which are our strategic opponents. This forces us to concentrate our efforts in this direction as much as possible, and to get ahead of our competitors. Recently, in Russia there was published the first book dedicated to AGI [Vedyakhin et al. 2021]. It analyzes in detail the main theoretical and methodological issues of the development of AGI and the scientific approaches necessary for this, identifies those specific functions that must be created and, most interestingly, the supposed large-scale changes that it is able to produce in the emerging new world order.

The authors identify two key capabilities of AGI, which are similar to natural intelligence. Unlike the narrow AI that solves one specific task, it must be integral, i.e., capable of solving many types of tasks. And it must acquire autonomy, i.e., the ability to act effectively and independently in various environments. All these increase the opera-

tional capabilities of AI systems, their use for military equipment and military operations, for solving problems of production, management, planning, organization of economic processes, optimization of a wide variety of spheres of public life and scientific research, which is extremely important for our country, under the current circumstances. And this is just as important for understanding and implementing those future historic changes in the development of the Earth's civilization, which are associated with the collapse of the principle and practice of the monopolar world.

At the same time, the development of AGI will raise new complex theoretical questions concerning its interactions with natural intelligence, the creation and prospects of hybrid intelligence, its possible competition with natural intelligence, possible risks and threats to man and society. Achieving a high degree of autonomy of AGI may lead to the appearance of such types and methods of its "amateur activity," which may pose a danger to man and society and will require new methods of ensuring security. Here we will face a new aspect of the same problem of *agency* of AI systems, which remains highly relevant.

Is AI a subject or an agent controlled by natural intelligence subjects?

Let us turn once again to the interpretations of theoretical questions about the agency of AI systems in relation to real human subjects. The approach of Iason Gabriel, the author of the above-mentioned article [Gabriel 2022], is to transfer to AI the same principles that are established for humans. The author writes that "AI increasingly shapes elements of the basic structure" of society, and, consequently, "the development and deployment of AI systems represent a new site for the operation of principles of distributive justice." According to the author, "AI interacts with the behavior of human decision-makers to shape the character of these practices, including how they distribute benefits and burdens across the population" [Gabriel 2022].

It turns out that for the author technologies seem to "come alive": "What is important for our purpose," he writes, "is that in modern societies, background justice is increasingly mediated algorithmically" [Gabriel 2022]. Here, Gabriel introduces the concept of background justice (apparently, mass-oriented justice) and assume that this justice is subject to an algorithm. Further, he states: "By making assessments

or predictions based upon an individual's past choices, and by providing decisions or recommendations that then shape that person's opportunity set, these systems exert a strong influence on the unfolding relationship between individual choices and collective outcomes" [Gabriel 2022]. As we can see, it is again assumed that the systems perform "intelligently" and create something. The author grants agency to AI technologies and demands that we extend the laws of social relations to AI, which "should support citizens' basic liberties, promote fair equality of opportunity, and provide the greatest benefit to those who are worst-off" [Gabriel 2022].

This anthropomorphic approach to artificial intelligence is fraught with grave consequences. Firstly, by transferring agency to AI, responsibility is implicitly removed from the human person who uses this technology, which hinders legal procedures. All technology is imperfect and cannot make a decision by itself. The error in the operation of the lie detector will never equal zero, and the same is true for using AI. It is the job of people to take into account the limitations of technology, and not just try to put them in some given framework. It is not by chance that in judicial practice, granted all the possibilities of criminology, humans do make the final decision. AI systems are not different from other technologies, so technologies should not remove responsibility from a person, and therefore should not have human characteristics (i.e., be fair, or humane, et al.).

The triumph of AI turned out to be so significant that many have decided that this technology can replace a human person, it only it needs placing in a certain framework. In fact, artificial intelligence has become a fetish of the 21st century, which some people extol as our future, while others fight against it. In fact, AI can recognize what the human eye cannot and identify a correspondence that a person cannot find. Yet itlags far behind the true cognitive capabilities of a human. Maybe in future it will be possible to create an AI integrated into the social environment, the so-called "general" AI, but now we are too far from this point and granting it agency is not only wrong, but also quite dangerous.

Ethical issues of AI development

The ethics of using AI are now widely discussed not only by public figures, politicians, but also by scholars. Much attention was drawn

to the public argument over the draft paper on the ethics of artificial intelligence by a Google employee, Timnit Gebru, one of the world's leading experts on the problems of bias in algorithms and data mining [Hao 2020]. The scientific and commercial positions fundamentally diverged, and as a result, the researcher left Google.

Discussions are also underway among Russian scientists. Along with constructive proposals, some recognized scholars offer very debatable and call for politicization of science and technology, endowing technology with the properties of "patriotism." Thus, on November 23, 2021, a meeting of the Presidium of the Russian Academy of Sciences was held, at which the academicians discussed, among other things, the ethical aspect of AI and the possibilities for AI to be trusted. One of the speakers stated, "From my point of view, artificial intelligence should not only be trusted, as was discussed today, but also should be patriotic, that is, it should primarily work for the interests of the country, and not against it" [Slavin 2021].

Today, many people protest against facial recognition technologies (in some US cities such recognition is prohibited, the European Parliament also proposed banning facial recognition technologies). However, it is not the recognition itself that is harmful, but its use for illegal purposes. People do not hide their illnesses from a doctor because they trusts this specialist and expect help. It is necessary to regulate the laws of use of technologies, including artificial intelligence, and to seriously punish those who harm a person or act illegally. Today, Russia appears poorly protected against fraudsters who use means of communication to deceive gullible citizens. All that the authorities can do today is to warn the population about new techniques of fraud. At the same time, the government is quite effective (also in the use of AI) in fighting its political opponents. The ethical problems of AI should be solved not by limiting the technologies, but primarily by limiting the actions of those people who misuse them.

Solving ethical issues of AI application is a very complex matter that cannot be accomplished only by adopting declarative codes of AI ethics. It is necessary to take into account that even our ethical principles, whatever ones we adopt – from the biblical commandments to the code of the builder of Communism – are only a few points in the infinite space of moral and ethical decisions in which we travel daily. Discussions of the ethics of AI are just beginning and professional com-

munities of philosophers and engineers should work closely together to develop answers to the new challenges of the time.

It should be emphasized that it is extremely important now to consider ethical problems in relation to the development of AI in a broader conceptual sense – from the viewpoint of the real functioning of ethical norms in our social life, the real state of morality in mass consciousness, in individual, collective, and institutional subjects. After all, there has always existed a gap between the knowledge of ethical norms and their implementation. Let us recall the Latin saying: *Video meliora, proboque, deteriora sequor* ("I see better things, and approve, but I follow worse"). Much too often, personal interest turned out to suppress moral precepts, and deception made a cozy nest for itself in the highest ethical instructions. There are no sufficient reasons to talk about moral progress in the development of mankind (detailed materials on this subject are presented in many philosophical studies [Dubrovsky 2007]). All these circumstances must be taken into account when we discuss the topic of "AI Ethics," both in relation to the creators of AI systems and to their users.

When trying to model the principles of ethics and implement them in AI operation, another challenge is that ethical norms cannot be organized as a clear hierarchical structure that allows an alternative choice. Here, the choice always depends on considering and evaluating the specific circumstances. Therefore, such modeling is only possible in specially selected cases.

At the same time, the problem of agency in AI design requires closer study. After all, in one way or another, the ability of an AI system to solve complex problems is associated with descriptions of some functions of natural intelligence. If it is not advisable to represent AI as a subject similar to a person, then how do we understand and define AI, when it is allowed to make decisions in certain situations and which defeats the world champions in chess and Go games? In our opinion, the most adequate approach may be the representation of AI as an agent endowed with a set of invariant simplified qualities that natural agents possess. These qualities include: deliberation, reflexivity, communicativeness, and simplified sociability. This representation of AI as an agent (or pseudo-subject) is consistent with the principle of distributed control in biology and psychology, called the principle of a dual subject [Lepskiy 1998]. This representation of AI, combined with systems of

principles and ontologies set in the concept of post-non-classical cybernetics of self-developing environments, allows using AI as a means of social innovation, while maintaining control over AI technologies as well as posing and solving the problem of integrating artificial and natural intelligence formations, yet preserving the basic qualities of natural intelligence [Lepskiy 2018; Lepskiy 2021].

Conclusion

AI is undergoing a phase of rapid growth. And yet, the impressive statistics of its successes should not mislead us: the current period is only the beginning of a total penetration of AI into our lives. That is why we should be very sensitive to possible cognitive distortions in the study of new phenomena. For example, anthropomorphism applied to AI can make us believe in the agency of the machine, which is false. The present paper, written by authors who specialize in philosophy and methodology, initiates a broader dialogue and a transition from creating codes of conduct for AI to creating the next generation of AI collaborating with humans and for them.

ЦИТИРУЕМАЯ ЛИТЕРАТУРА

Дубровский 2007 – *Дубровский Д.И.* О нравственном прогрессе и нравственном регрессе (К проблематике развития морального сознания) // Философские науки. 2007. № 11. С. 81–102.

Дубровский 2021 – *Дубровский Д.И.* Задача создания Общего искусственного интеллекта и проблема сознания // Философские науки. 2021. Т. 64. № 1. С. 13–44.

Лепский 1998 – *Лепский В.Е.* Концепция субъектно-ориентированной компьютеризации управленческой деятельности. – М.: Институт психологии РАН, 1998.

Лепский 2020 – *Лепский В.Е.* Рефлексия пандемии COVID-19: субъектно-ориентированный подход // Экономические стратегии. 2020. № 8 (174). С. 66–71.

Лепский 2021 – *Лепский В.Е.* Искусственный интеллект в субъектных парадигмах управления // Философские науки. 2021. Т. 64. № 1. С. 88–101.

Прорывные инновации… 2022 – Прорывные инновации: человек 2.0: доклад к XXIII Ясинской (Апрельской) международной научной конференции по проблемам развития экономики и общества, Москва, 4–8 апреля 2022 г. / под ред. Л.М. Гохберга, А.Р. Ефимова, Ю.В. Мильшиной. – М.: НИУ ВШЭ, 2022.

Сильный искусственный интеллект… 2021 – Сильный искусственный интеллект: На подступах к сверхразуму / А. Ведяхин и др. − М.: Интеллектуальная литература, 2021.

Славин 2021 − *Славин Б.Б.* Может ли искусственный интеллект быть справедливым // БИТ. 2021. № 10 (113). С. 32−35.

Gabriel 2022 − *Gabriel I.* Toward a Theory of Justice for Artificial Intelligence // AI & Society. 2022. Vol. 151. No. 2. P. 218−231.

Hao 2020 – *Hao K.* We read the paper that forced Timnit Gebru out of Google. Here's what it says // MIT Technology Review. 4 December 2020. – URL: https://www.technologyreview.com/2020/12/04/1013294/google-ai-ethics-research-paper-forced-out-timnit-gebru/

Lepskiy 2018 – *Lepskiy V.* Evolution of Cybernetics: Philosophical and Methodological Analysis // Kybernetes. 2018. Vol. 47. No. 2. P. 249–261.

Rawls 1999 − *Rawls J.* A Theory of Justice. − Cambridge, MA: Harvard University Press, 1999.

The Right to Privacy… 2021 − The Right to Privacy in the Digital Age: Report of the United Nations High Commissioner for Human Rights. A/HRC/48/31 // Office of the High Commissioner for Human Rights. 13 September 2021. − URL: https://www.ohchr.org/en/documents/thematic-reports/ahrc4831-right-privacy-digital-age-report-united-nations-high

REFERENCES

Dubrovsky D.I. (2007) On Moral Progress and Moral Regress. *Russian Journal of Philosophical Sciences = Filosofskie nauki.* No. 11, pp. 81−102 (in Russian).

Dubrovsky D.I. (2021) The Task of Creating a General Artificial Intelligence and the Problem of Consciousness. *Russian Journal of Philosophical Sciences = Filosofskie nauki.* Vol. 64, no. 1, pp. 13−44 (in Russian).

Gabriel I. (2022) Toward a Theory of Justice for Artificial Intelligence. *AI & Society.* Vol. 151, no. 2, pp. 218−231.

Gokhberg L.M., Efimov A.R., & Milshina Y.V. (Eds.) (2022) *Breakthrough Innovations: Man 2.0: Report to the 23rd Yasin (April) International Scientific Conference on Problems of Economic and Social Development, Moscow, April 4−8, 2022.* Moscow: National Research University Higher School of Economics (in Russian).

Hao K. (2020, December 4) We read the paper that forced Timnit Gebru out of Google. Here's what it says. *MIT Technology Review.* Retrieved from https://www.technologyreview.com/2020/12/04/1013294/google-ai-ethics-research-paper-forced-out-timnit-gebru/

Lepskiy V.E. (1998) *The Concept of Subject-Oriented Computerization of Managerial Activity.* Moscow: Institute of psychology Russian academy of sciences (in Russian).

Lepskiy V. (2018) Evolution of Cybernetics: Philosophical and Methodological Analysis. *Kybernetes.* Vol. 47, no. 2, pp. 249–261.

Lepskiy V.E. (2020) Reflection of the COVID-19 Pandemic: a Subject-Oriented Approach. *Ekonomicheskie strategii.* No. 8, pp. 66−71 (in Russian).

Lepskiy V.E. (2021) Artificial Intelligence in Subject-Oriented Control Paradigms. *Russian Journal of Philosophical Sciences = Filosofskie nauki.* Vol. 64, no. 1, pp. 88–101 (in Russian).

Office of the High Commissioner for Human Rights (OHCHR) (2021, September 13) *The Right to Privacy in the Digital Age: Report of the United Nations High Commissioner for Human Rights.* A/HRC/48/31. Retrieved from https://www.ohchr.org/en/documents/thematic-reports/ahrc4831-right-privacy-digital-age-report-united-nations-highRawls J. (1999) *A Theory of Justice.* Cambridge, MA: Harvard University Press.

Slavin B.B. (2021) Can Artificial Intelligence Be Fair? *BIT Journal.* No 10 (113), pp. 32−35 (in Russian).

Vedyakhin A.A. et al. (2021) *Strong Artificial Intelligence: On the Approaches to the Supermind.* Moscow: Intellektualnaya Literatura (in Russian).

DOI: 10.30727/0235-1188-2022-65-1-72-90
Оригинальная исследовательская статья
Original research article

Субъектность объяснимого искусственного интеллекта*

А.Н. Райков
Институт проблем управления РАН, Москва, Россия,
МИРЭА – Российский технологический университет, Москва, Россия

Аннотация

Статья посвящена определению путей развития способности систем искусственного интеллекта (ИИ) давать объяснения своим выводам. Эта тема не нова, однако именно сейчас нарастание сложности этих систем заставляет ученых активизировать исследования в этом направлении. Современные нейронные сети содержат сотни слоев нейронов, число параметров этих сетей достигает триллионов, генетические алгоритмы порождают тысячи поколений решений, семантика моделей ИИ усложняется, достигая квантового и нелокального уровней. Ведущие компании мира вкладывают огромные средства в создание объяснимого ИИ (*Explainable AI, XAI*), однако результат пока остается неудовлетворительным – человек зачастую не может понять «объяснений» ИИ, потому что последний принимает решения иначе, чем человек, а возможно, потому что получить хорошее объяснение невозможно в рамках классической парадигмы ИИ. С похожей проблемой ИИ столкнулся лет 40 назад, когда экспертные системы содержали всего несколько сотен логических правил-продукций. Проблема тогда была решена за счет усложнения логики и построения дополнительных баз знаний для объяснения выводов, даваемых ИИ. Сейчас, по-видимому, нужны иные подходы, прежде всего учитывающие внешнее окружение и субъектность систем ИИ. Настоящая работа акцентирует внимание на разрешении этой проблемы через погружение моделей ИИ в социально-экономическую среду, построение онтологий этой среды, а также учет профиля пользователя и формирование условий для целенаправленной сходимости решений и выводов ИИ к понятным пользователю целям.

Ключевые слова: философия искусственного интеллекта, онтологии, эпистемология, причинность, рефлексивно-активные системы, субъективная реальность.

* Работа поддержана Российским научным фондом (РНФ), грант № 21-18-00184 «Социогуманитарные основания критериев оценки инноваций, использующих цифровые технологии и искусственный интеллект».

Райков Александр Николаевич – доктор технических наук, ведущий научный сотрудник Института проблем управления РАН, профессор МИРЭА – Российского технологического университета.
ANRaikov@mail.ru
https://orcid.org/0000-0002-6726-9616

Для цитирования: *Райков А.Н.* Субъектность объяснимого искусственного интеллекта // Философские науки. 2022. Т. 65. № 1. С. 72–90. DOI: 10.30727/0235-1188-2022-65-1-72-90

Subjectivity of Explainable Artificial Intelligence*

A.N. Raikov
Institute of Control Sciences, Russian Academy of Sciences, Moscow, Russia,
MIREA – Russian Technological University, Moscow, Russia

Abstract

The article addresses the problem of identifying methods to develop the ability of artificial intelligence (AI) systems to provide explanations for their findings. This issue is not new, but, nowadays, the increasing complexity of AI systems is forcing scientists to intensify research in this direction. Modern neural networks contain hundreds of layers of neurons. The number of parameters of these networks reaches trillions, genetic algorithms generate thousands of generations of solutions, and the semantics of AI models become more complicated, going to the quantum and non-local levels. The world's leading companies are investing heavily in creating explainable AI (XAI). However, the result is still unsatisfactory: a person often cannot understand the "explanations" of AI because the latter makes decisions differently than a person, and perhaps because a good explanation is impossible within the framework of the classical AI paradigm. AI faced a similar problem 40 years ago when expert systems contained only a few hundred logical production rules. The problem was then solved by complicating the logic and building added knowledge bases to explain the conclusions given by AI. At present, other approaches are needed, primarily those that consider the external environment and the subjectivity of AI systems. This work focuses on solving this problem by immersing AI models in the social and economic environment, building ontologies of this environment, taking into

* The work was supported by the Russian Science Foundation, grant no. 21-18-00184 "Social and humanitarian foundations for evaluation criteria for innovations based on digital technologies and artificial intelligence."

account a user profile and creating conditions for purposeful convergence of AI solutions and conclusions to user-friendly goals.

Keywords: philosophy of artificial intelligence, ontologies, epistemology, causality, reflexive-active systems, subjective reality.

Alexander N. Raikov – D.Sc. in Technology, Leading Research Fellow, Institute of Control Sciences, Russian Academy of Sciences; Professor, MIREA – Russian Technological University.
ANRaikov@mail.ru
https://orcid.org/0000-0002-6726-9616

For citation: Raikov A.N. (2022) Subjectivity of Explainable Artificial Intelligence. *Russian Journal of Philosophical Sciences = Filosofskie nauki.* Vol. 65, no. 1, pp. 72–90. DOI: 10.30727/0235-1188-2022-65-1-72-90

Введение

Проблема доверия выводам систем ИИ появилась почти одновременно с рождением термина «искусственный интеллект» (ИИ), достаточно вспомнить работу Д. Пойи «Математика и правдоподобные рассуждения» [Polya 1954]. Уже тогда разработчиков систем ИИ интересовали правдоподобные рассуждения и достоверный вывод [Вагин и др. 2004]. Успешного разрешения проблемы «черного ящика ИИ», как она сейчас иногда обозначается, научно-технический мир достиг лет 40 назад, когда при создании средств ИИ основное внимание стало уделяться созданию логических экспертных систем. Тогда система ИИ на основе довольно замкнутой базы знаний, состоящей из 300–400 правил-продукций, выводила результат, но не могла его доходчиво объяснить. Такие системы не вызывали доверие, плохо продавались, разработчики тратили десятки миллионов долларов на гонорар инженерам по знаниям (когнитологам), предметным экспертам и программистам. С этой проблемой справились путем усложнения логик и создания дополнительных баз знаний, помогающих обобщить результат вывода и дать сносное объяснение.

За последние годы сложность ИИ резко возросла, число слоев в нейронных сетях достигает нескольких сотен, число параметров – триллионы, эволюционные методы вычислений порождают тысячи поколений (множеств) решений, объемы больших данных для обучения систем ИИ стремительно растут и пр. Меняются взгляды на семантику моделей систем ИИ, ее экспликация уходит на неформализуемый когнитивный уровень, учитывает квантовые

и релятивистские эффекты [Raikov 2021]. При этом в рамках развития ИИ возникают принципиально новые проблемы: обостряются угрозы в сфере кибербезопасности, создаются квантовые версии вычислителей, возрастает опасность злонамеренного и неэтичного использования ИИ, системы ИИ становятся решающим фактором производства конкурентоспособной продукции и др.

В таких условиях ведущие компании мира (*DARPA, Google, Tencent Research Institute* и др.) увеличивают инвестиции в создание объяснимого ИИ (*XAI, Explainable AI*). Ключевыми идеями исследований сейчас являются: послойный анализ работы ИИ, испытанные ранее продукционные (по сути, логические) схемы вывода, построения на графах знаний и пр. Среди возможных средств порождения объяснений в последнее время все больше используются графы знаний. Разветвленные связи в графах знаний делают их полезными для объяснений выводов систем ИИ. Например, эти графы строятся путем подсчета расстояний между целевым профилем пользователя и элементом в графе знаний, от которого зависит развитие ситуации. Авторы [Liu, Balsubramani, Zou 2019] предлагают генерировать объяснения по графам, сопровождая процесс обучением с подкреплением. В работе [Madry et al. 2017] предпринята попытка сделать так, чтобы нейронная сеть давала объяснение путем выявления причин взаимодействия пользователя с системой. Однако результаты пока не считаются удовлетворительными. Ответ на каждый новый вопрос по созданию объяснимого ИИ, делающий вроде бы «черный ящик» искусственной нейронной сети «белым», на самом деле обнаруживает внутри него множество других «черных ящиков». При этом экспоненциально растет число необходимых для получения объяснений вычислений, что зачастую заставляет отказываться от построения компоненты объяснения и ее включения в систему ИИ.

Внимание исследователей все больше привлекает учет субъектного фактора создания и использования систем ИИ, помещение этих систем в саморазвивающуюся рефлексивно-активную среду, построение онтологий контекста использования [Дубровский 2021; Лепский 2021].

В настоящей работе предлагается поменять акценты в парадигме создания *XAI,* обратить внимание на субъектные аспекты, внешнюю и глубинную стороны традиционных моделей построения ИИ. При этом к внешней стороне может относиться социально-экономическая и субъективная реальность, а к глу-

бинной – флюктуирующая, квантовая и релятивистская природа сознания человека.

XAI: состояние вопроса

Общепризнанного определения и взгляда на реализацию *XAI,* конечно, нет и, по-видимому, быть не может. Вместе с тем потребность в *XAI* нарастает, а значит следует развивать методологию разрешения проблемы. Имеются множественные интенции и предложения ученых и инженеров, которые сопровождают разработку *XAI,* например, объяснительный ресурс ИИ:

• зависит не столько от сложности логики и алгоритмов работы системы ИИ, сколько от пользователя и внешнего контекста, включая социум, экономику, пользователя, целенаправленность и др. [Лепский 2021];

• обеспечивает создание моделей ИИ, которые могут объяснить свои выводы, сохраняя нужную точность прогнозирования, обеспечивая пользователям понимание, доверие и управление новыми объектами [Chen et al. 2020];

• должен гарантировать, что решения и любые данные, обеспечивающие их, могут быть объяснены человеком непрофессионалом [Wang et al. 2020];

• направлен на послойную расшифровку работы системы ИИ для создания моделей и методов, которые одновременно точны и дают удовлетворительное объяснение [Veličković et al. 2019] и др.

До недавнего времени основное внимание построению компонент объяснения в системах ИИ уделялось естественно-научным, инженерным и логико-технологическим компонентам. В результате можно выделить следующие основные технологические аспекты XAI, которые определяют текущую реализацию прикладных исследований:

• учет внешнего окружения;

• верифицируемость объяснения на других примерах и моделях;

• зависимость объяснения от объяснительной модели ИИ, например, направленной на визуализацию, построение графики, использование примеров;

• принятие во внимание характеристик субъекта, нуждающегося в объяснении.

Остановимся сначала на первых двух из перечисленных аспектов. Они характеризуют техническую сторону цифровой среды

и практически не затрагивают природу субъектности, сознания, эмоций, духовную сферу. Так, в ряде работ обсуждаются вопросы так называемого обволакивающего интеллекта [Шалова 2016], подразумеваемого в значении окружающего интеллекта (*Ambient Intelligence, AmI*). *AmI* относится к цифровой среде, которая чувствительна и реагирует на присутствие людей. Окружающий интеллект разрабатывается с конца 1990-х годов как проекция электроники, телекоммуникаций и вычислений на будущее [True Visions… 2006; Nolin 2016]. Такие системы сейчас разрабатываются для обеспечения согласованной работы различных технических устройств, чтобы поддерживать людей в их повседневной деятельности, решения задач интуитивно понятным способом, используя информацию и ИИ, которые скрыты в сети, соединяющей эти устройства (например, Интернет вещей).

Построенные объяснения могут быть как релевантными, формализованными, так и смысловыми, неформализованными. Первые лежат в логической плоскости: само объяснение можно проверить, например, послойно анализируя процесс вывода построить логические причинно-следственные цепочки получения вывода. На систему объяснения в этом случае принципиальным образом влияют данные, которые используются для обучения и адаптации системы ИИ [Lin, Hung, Huang 2021; Leavy at al. 2020; Leavy, Siapera, O'Sullivan 2021]. Вторые могут быть оценены только через семантическую интерпретацию пользователем полученного от системы ИИ объяснения, например, пользователь может быть не согласен с результатом, поскольку у него есть свое понимание вопроса. Для второго случая данных недостаточно, поскольку данные представляют собой формализованную компоненту системы ИИ, только косвенно затрагивающую субъектные аспекты системы.

Вместе с тем проблема *XAI* – это не только логическое и технологическое предоставление удовлетворительных объяснений, она тесно связана с неформализуемыми аспектами феномена сознания, субъективной реальности [Дубровский 2021], противоречивого эпистемологического и этического выбора [Kaul 2022]. Ее разрешение должно учитывать способы, которыми люди достигают договоренностей в сферах политических, экономических, социальных отношений, а также соглашаются с тем, чтобы ими впоследствии руководствоваться.

Ссылаясь на требование прозрачности при логическом принятии решений, исследователи отмечают потребность понять, что

такое объяснение вообще [Rauber, Trasarti, Gianotti 2019, 10–11]. Понимание объяснения далеко не однозначно и варьируется в зависимости от ситуации и дисциплины. Довольно емкий обзор по теме объяснения [Mueller et al. 2019] представляет собой широкий набор ссылок на работы по *XAI* в различных дисциплинах. В обзоре анализируются ключевые концепции *XAI* и различные виды систем ИИ (экспертные системы, системы рассуждений на основе прецедентов, системы машинного обучения, байесовские классификаторы, статистические модели и деревья решений). Рассматривается разнообразие приложений: классификации жестов, изображений, текста; отладка программ; синтез музыкальных рекомендаций; финансовый учет; стратегические игры; формирование команд; роботизация; диагностика болезней; построение гипотез, касающихся отношения объяснения к фундаментальным когнитивным процессам и пр. По всей видимости, обзор не столько определяет пути развития *XAI*, сколько демонстрирует неочевидность дальнейшего развития темы.

В большом обзоре работ по *XAI* [Adadi, Berrada 2018] выявляются коммерческие, этические и нормативные причины, по которым необходимы объяснения. Рассматриваются различные цели объяснений, например, чтобы оправдать, контролировать, улучшать и открывать. Обсуждаются методы объяснений с точки зрения локального и глобального, внутреннего и апостериорного, модельно-специфического и модельно-независимого (см. в частности, таблицу, обобщающую ключевые концепции XAI [Adadi, Berrada 2018, 52141]). Облако слов *XAI,* которое они предоставляют [Adadi, Berrada 2018, 52140], имеет интерпретируемый язык машинного обучения (*ML*) и объяснимый ИИ как наиболее известные термины в литературе.

Работа [Arrieta et al. 2020] представляет обзор концепций, таксономий, возможностей и проблем XAI по таким направлениям, как классификация моделей ML в зависимости от их уровня объяснимости [Arrieta et al. 2020, 90], таксономия литературы и тенденций в сфере объяснимости для моделей машинного обучения [Arrieta et al. 2020, 93]. Авторы отмечают, что интерпретируемость «черного ящика» машинного обучения важна для обеспечения беспристрастности при принятии решений, устойчивости к злонамеренным возмущениям, а также в качестве гарантии того, что только значимые переменные определяют результат [Arrieta et al. 2020, 83]. Авторы пролагают, что

феномен понятности является наиболее важным атрибутом *XAI* [Arrieta et al. 2020, 84–85].

Вместе с тем ответ на вопрос о потребности в объяснении применительно к системам ИИ не является однозначным. Трактовки объяснимости иногда вызывают скептицизм в отношении ее полезности. Предстоит выяснить:

• Может ли высокая прозрачность работы системы ИИ привести к информационной перегрузке?

• Может ли визуализация привести к чрезмерному доверию или неправильному чтению?

• Могут ли системы машинного обучения на самом деле предоставлять хорошие обоснования на естественном языке?

• Действительно ли разные люди в любом случае нуждаются в разных объяснениях? [Heaven 2020].

Ставятся вопросы, связанные с определением роли объяснения при оценке ответственности в законодательстве и судебной практике, отмечается потребность достигнуть компромисса между полезностью и стоимостью объяснений, сложностью и вычислительными ресурсами. Ведь любая уточняющая информация в виде объяснения может быть представлена в виде набора абстрактных причин или оправданий некоего результата, а не описания процесса принятия решения в целом [Doshi-Velez, Kortz 2017, 4]. Для них генерация объяснений – это вопрос дизайна системы, и они предлагают рассматривать системы объяснений отдельно от систем ИИ [Doshi-Velez, Kortz 2017, 16–17], чтобы создать возможности для отраслей, где функционируют системы объяснений в терминах, интерпретируемых человеком, без ущерба для точности исходного предиктора.

В обзоре, посвященном принципам объяснения и человеко-машинным системам ИИ, делается акцент на необходимости ориентации на человека, у которого есть ожидания в отношении объяснения, при создании системы объяснения [Mueller et al. 2020]. Работа в области психологии, посвященная познанию и предубеждениям, использует данные человеческого мышления, чтобы аргументировать их вклад в интерпретируемые модели сложных систем ИИ [Byrne 2019, 6280].

В итоге можно отметить в целом позитивное отношение к дальнейшему развитию *XAI*. Тема постоянно экстраполируется во многие смежные области, включая социально-экономические. Это было ожидаемо, поскольку большинство исследователей в

области ИИ имеют опыт работы в области вычислений, математики, технической кибернетики, естественных наук и, возможно, психологии или философии. Как известно, характер знаний существенно зависит от того, кто является его потребителем, в какой среде оно сформировано и как ведется. Поэтому природа и структура объяснений строятся с учетом не столько вероятности выбора, который делает система ИИ, сколько внешних, в частности социально-экономических, причин его порождения, а также ожиданий и переживаний пользователя системы ИИ.

Каузальность (причинность) как базис объяснения

Особое место в построении компонент объяснения занимает тезис каузальности – причинно-следственной связи событий. Причинность, по-видимому, является наиболее фундаментальным явлением, отражающим всеобщую связь и единство во Вселенной. Она связывает мысли с действиями, а действия с последствиями, движение планет с образующими их атомами, успех лечения людей от используемой методики и др. На частый вопрос «Почему?» помогают ответить причинно-следственные связи между событиями. Подобные связи объясняются научными законами и закономерностями.

Причинность характеризуется частым совместным появлением событий. При этом далеко не всегда по частоте совместного появления событий можно однозначно судить о наличии между ними причинно-следственной связи. Например, традиционный ИИ не может точно преобразовать корреляцию событий в такую связь и не может обеспечить высокий уровень прозрачности или доверия к результатам вывода системы ИИ. Однако именно корреляция и частотность событий, встречающихся в больших массивах информации, являются основой для построения явных и выявления неявных причинно-следственных связей между событиями.

В чем причина того, что коляска движется, если ее толкает человек, или в космическом пространстве звезды вращаются вокруг центра масс и образуются кратные звезды? Философы исследуют этот вопрос на протяжении тысячелетий, по крайней мере со времен Платона и Аристотеля. Проблема рассматривается абстрактно и конкретно, с разных сторон. Число каузальных утверждений явно нарастает, но вопрос остается.

Наиболее распространенное мнение, что в основе причинно-следственной связи лежит регулярность: одно событие (вещь)

постоянно связано с другим. Классическая и наиболее распространенная тока зрения восходит к Дэвиду Юму. Он считал, что если события *A* и *B* постоянно появляются вместе и *A* происходит до *B*, то этого все же недостаточно для того, чтобы заключить, что *A* является причиной *B*. Причина появляется тогда, когда *A* и *B* также должны быть смежными, быть рядом друг с другом, то есть иметь пространственную связь.

Вместе с тем понятие смежности нельзя не подвергнуть сомнению. Например, из фундаментальной физики известен эффект квантовой запутанности (энтэлгмента). Можно считать доказанным, что две частицы, находящиеся в разных концах почти бесконечной Вселенной, могут быть связаны таким образом, что изменение квантового состояния одной из них приводит к мгновенному и независимому от расстояния изменению состояния другой [Einstein, Podolsky, Rosen 1935; The BIG Bell… 2018]. Это, однако, происходит в нарушение законов теории относительности: причинно-следственная связь движется быстрее скорости света. И остается вопрос, является ли изменение состояния одной из частиц случаем реальной причинности. Пока остается непонятной природа явления и причины флюктуации частицы, состояние которой изменилось в первую, вторую очередь или совместно с другой. Появился некий научный парадокс, и, как следствие, временной приоритет и смежность по Юму могут быть оспорены.

Понятие причинности претерпевает изменения со временем и в различных контекстах. Ранее причинность больше мыслились как действия агента-человека или робота, сейчас понятие причинности трансформируется в причинность, которая становится предметом объяснения, свободного от эмоциональной окраски, например восхищения, похвалы или осуждения.

Вопросы причинности классифицируются по нескольким направлениям в зависимости от подхода: плюрализм, примитивизм, физикализм и др. Признается, что переход к плюрализму является просто признанием поражения всех иных, скорее специфичных направлений. Например, плюралист использует различные теории и понимает причинность как множество разных вещей или событий. Причинно-следственная связь в отдельных теориях и подходах не поддается анализу. Возможно, это происходит из-за ограниченности выбранного подхода [Mumford, Anjium 2013].

Вместе с тем в аналитической философии, например с учетом взглядов Локка, что-то должно быть взято за основу

[Mumford, Anjium 2013, 88–89], хотя любую часть аналитического подхода можно подвергнуть сомнению. Если бы в мире была бесконечная сложность, например, доходящая до структуризации фотонов, нейтрино и кварков, то в конце концов не было бы ничего базового. Но это неизвестно наверняка. Возможно, существует базовый, неразложимый далее элемент природы, который мы не знаем, но способны в будущем познать. Этот взгляд может подойти для отмеченного выше примитивистского подхода. Некоторые примитивы, вероятно, будут правильными, так что нет ничего явно неправильного в том, чтобы использовать примитивистский подход. Возможно, это больше, чем просто возможный вариант, особенно для частных утилитарных соображений.

Тема объяснимости выводов ИИ явно шире и выходит за рамки вопроса каузальности, охватывает субъективную реальность в саморазвивающихся междисциплинарных полисубъектных (рефлексивно-активных) средах [Лепский 2021]. Бытие субъектов в таких средах может быть задано системой онтологий, которая обеспечивает сборку пребывающих в среде субъектов в целое. Разработана и опробована система онтологий, в которую входят онтологии обеспечения жизнедеятельности, преодоления точек разрыва, стратегического целеполагания, разработки стратегий и проектов, внедрения и инновационного обеспечения стратегий и проектов [Lepskiy 2019; Лепский 2021]. Показано, что постановка задач для ИИ должна осуществляться с использованием системы онтологий бытия субъектов и в контексте поддержки рефлексивной активности субъектов.

Таким образом, в контексте создания *XAI* можно думать о причинности как об одной из фундаментальных сил и элементов Вселенной. Причинность – это то, что удерживает объекты вместе посредством атомарных и молекулярных связей. Она производит изменение одной вещи с помощью другой. Это придает любому действию значимость. И тогда есть ли основания думать, что мы можем объяснить причинность в некаузальных терминах?

Дисциплинарные и предметные особенности объяснения

Отметим особый подход экономических теорий к феномену объяснения [Kaul 2022]. Экономисты иногда считают объяснения, даваемые экономическими теориями и моделями, формой, которую принимает теория, и воспринимают это как само собой разумеющееся. Большинству людей, которые сталкиваются

с объяснениями, предлагаемыми этими теориями, сама идея объяснения может показаться нелогичной. Экономисты строят модели, которые направлены именно на абстрактное объяснение экономической ситуации и поэтому далеко не всегда могут быть опровергнуты эмпирическими данными. Вместе с тем объяснительный приоритет модельного экономического построения работает совсем не только для получения количественных характеристик динамики и прогноза развития ситуации, которая, по сути, в экономике наиболее важна. В указанной работе [Kaul 2022] отмечаются разнообразные взгляды экономистов на получение ответов на вопросы относительно феномен объяснения:

• примирение противоборствующих сторон относительно роли ценностных или нормативных различий;

• признание необходимости телеологии в объяснении при неспособности эмпиризма определять человеческие цели;

• различие и развитие природы телеологических и каузальных объяснений в конкретных областях экономики; использование модели рационального выбора;

• объяснение событий в различных подотраслях экономики;

• рассмотрение дедуктивного, индуктивное и абдуктивного рассуждения в неоклассической экономике и др.

Все чаще встречаются работы, пытающиеся объяснить экономические явления с помощью инструментов, которые вроде бы далеки от экономики, например с помощью методов квантовой физики [Orrell, Houshmand 2022; Райков 2009]. Однако, по-видимому, вопросы адекватности применения методов физики к экономическим системам все еще нуждаются в осмыслении. Таким образом, экономика имеет очень своеобразный подход к объяснению по отношению к ее регулярной практике построения экономической теории с целью обеспечить достоверность и доверие к экспертизе экономической ситуации, и его нельзя признать удовлетворительным для использования в полной мере при создании *XAI*.

Очевидно, что люди относятся к сверхсложным системам, выходящим за границы, диктуемые технической кибернетикой. В отличие от естественных наук, где формулы, закономерности, физические законы и пр. могут быть получены путем наблюдения, измерены и воспроизведены разными способами, в субъектной рефлексивно-активной среде человеческое поведение внутри коллективов и между коллективами не поддается пря-

мому наблюдению, измерению и формализованному описанию [Лепский 2021], и, как следствие, однозначному объяснению. Поэтому для объяснения явлений привлекаются упомянутые выше онтологии.

В менеджменте, стратегическом планировании, где решение проблем осуществляется с учетом идентификации сотен факторов, материальных и субъектных, для объяснения проблемной ситуации используются методы стратегического анализа и когнитивного моделирования [Raikov 2022]. Как и в подходе с онтологиями, решаемая проблема коллективно декомпозируется на конечное число целей, факторов, препятствий, ограничений, функций, задач и пр., они выстраиваются специальным образом, например с помощью метода анализа иерархий, затем результат загружается в компьютер и проводится моделирование. Результат моделирования не всегда поддается объяснению, однако сам аналитический процесс, состоящий из множества шагов, делается под непосредственным контролем команды пользователей системы ИИ и должен вызывать доверие.

Предметные особенности проблемы объяснения неизменно проявляются в повседневной и деловой среде. Возьмем актуальную задачу борьбы с вредными насекомыми, например с кукурузным мотыльком. От него страдают порядка 250 растений, а урожай может сократиться на 25% и более. По всей видимости, системы ИИ могли бы помочь более успешно справиться с решением проблемы, поскольку обладают способностью собирать, накапливать и анализировать большие объемы данных, делать прогнозы. Однако ИИ далеко не всегда может давать объяснения своим выводам, что резко снижает доверие к ним и может сделать применение ИИ нерентабельным. Достаточно отметить сложность и скрытность поведения мотылька, которое пока не поддается успешному моделированию и прогнозированию [Абросимов, Райков 2022].

Приведенный пример с мотыльком видится много более простым, чем проблема объяснения с учетом субъектных аспектов. Вместе с тем в разрешении проблемы с мотыльком эти аспекты тоже имеют место, поскольку борьбу с вредителем растений запускают люди, а куколка прячется в стволе растения. Процесс едва ли поддается наблюдению, и, как следствие, охвату системами сбора информации, а затем их аналитической обработке, прогнозированию и объяснению с помощью систем ИИ.

Заключение

Объяснительная функция приобретает все более важное значение в дальнейшем развитии и применении систем ИИ, что вызвано их усложнением и, как следствие, снижением доверия к генерируемым ими рекомендациям.

Помимо технологической составляющей, решающее значение для повышения эффективности систем ИИ приобретает учет субъективной реальности, погружение системы ИИ в полисубъектную рефлексивно-активную среду.

Росту доверия к выводам систем ИИ способствует непосредственное включение людей в процесс получения результата. Для такого включения используются специальные подходы, основанные на построении онтологий и структуризации проблем для создания условий их целенаправленного разрешения.

ЦИТИРУЕМАЯ ЛИТЕРАТУРА

Абросимов, Райков 2022 – *Абросимов В.К., Райков А.Н.* Интеллектуальные сельскохозяйственные роботы. – М.: Карьера Пресс, 2022.

Вагин и др. 2004 – *Вагин В.Н., Головина Е.Ю., Загорянская А.А., Фомина М.В.* Достоверный и правдоподобный вывод в интеллектуальных системах / под ред. В.Н. Вагина и Д.А. Поспелова. – М.: Физматлит, 2004.

Дубровский 2021 – *Дубровский Д.И.* Задача создания Общего искусственного интеллекта и проблема сознания // Философские науки. 2021. Т. 64. № 1. С. 13–44.

Лепский 2021 – *Лепский В.Е.* Искусственный интеллект в субъектных парадигмах управления // Философские науки. 2021. Т. 64. № 1. С. 88–101.

Райков 2009 – *Райков А.Н.* Протуберанцы макроэкономики // Экономические стратегии. 2009. № 7. С. 42–49.

Шалова 2016 – *Шалова С.Х.* Обзор и анализ исследований в области систем обволакивающего интеллекта // Инженерный вестник Дона. 2016. № 4 (43). С. 125.

Adadi, Berrada 2018 – *Adadi A., Berrada M.* Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI) // IEEE Access. 2018. Vol. 6. P. 52138–52160.

Arrieta et al. 2020 – *Arrieta A.B., Díaz-Rodríguez N., Del Ser J., Bennetot A., Tabik S., Barbado A., García S., Gil-López S., Molina D., Benjamins R., Chatila R.* Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges towards Responsible AI // Information Fusion. 2020. Vol. 58. P. 82–115.

Byrne 2019 – *Byrne R.M.J.* Counterfactuals in Explaining Artificial Intelligence (XAI): Evidence from Human Reasoning // Proceedings of the

Twenty-Eighth International Joint Conference on Artificial Intelligence (IJCAI-19). – California: *International Joint Conferences on Artificial Intelligence,* 2019. P. 6276–6282.

Chen et al. 2020 – *Chen M., Wei Z., Huang Z., Ding B., & Li Y.* Simple and Deep Graph Convolutional Networks // Proceedings of Machine Learning Research. 2020. Vol. 119. P. 1725–1735.

Doshi-Velez, Kortz 2017 – *Doshi-Velez F., Kortz M.* Accountability of AI Under the Law: The Role of Explanation // Berkman Klein Center Working Group on Explanation and the Law, Berkman Klein Center for Internet & Society Working Paper. 2017. – URL: http://nrs.harvard.edu/urn-3:HUL.InstRepos:34372584

Einstein, Podolsky, Rosen 1935 – *Einstein A., Podolsky B., Rosen N.* Can Quantum-Mechanical Description of Physical Reality Be Considered Complete? // Physical Review. 1935. Vol. 47. No. 10. P. 777–780.

Heaven 2020 – *Heaven W.D.* Why Asking an AI to Explain Itself Can Make Things Worse // MIT Technology Review. 2020. January 29. – URL: https://www.technologyreview.com/2020/01/29/304857/why-asking-an-ai-to-explainitself-can-make-things-worse/

Kaul 2022 – *Kaul N.* 3Es for Al: Economics, Explanation, Epistemology // Frontiers in Artificial Intelligence. Vol. 5. Article 833238.

Leavy et al. 2020 – *Leavy S., Meaney G., Wade K., Greene D.* Mitigating Gender Bias in Machine Learning Sata Sets // International Workshop on Algorithmic Bias in Search and Recommendation. – Cham: Springer. P. 12–26.

Leavy, Siapera, O'Sullivan 2021 – *Leavy S., Siapera E., O'Sullivan B.* Ethical Data Curation for AI: An Approach based on Feminist Epistemology and Critical Theories of Race // AIES'21: Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society. – New York: The Association for Computing Machinery, 2021. P. 695–703.

Lepskiy 2018 – *Lepskiy V.* Evolution of Cybernetics: Philosophical and Methodological Analysis // Kybernetes. 2018. Vol. 47. No. 2. P. 249–261.

Lin, Hung, Huang 2021 – *Lin Y.T., Hung T.W., Huang L.T.L.* Engineering Equity: How AI Can Help Reduce the Harm of Implicit Bias // Philosophy and Technology. 2021. Vol. 34. No. 1. P. 65–90.

Liu, Balsubramani, Zou 2019 – *Liu R., Balsubramani A., Zou J.* Learning Transport cost from subset correspondence // arXiv. 2019. – URL: https://arxiv.org/pdf/1909.13203.pdf

Madry et al. 2017 – *Madry A., Makelov A., Schmidt L., Tsipras D., Vladu A.* Towards Deep Learning Models Resistant to Adversarial Attacks // arXiv. 2017. – URL: https://arxiv.org/pdf/1706.06083.pdf

Mueller et al. 2019 – *Mueller S.T., Hoffman R.R., Clancey W, Klein G.* Explanation in Human-AI systems: A Literature Meta-Review Synopsis of Key Ideas and Publications and Bibliography for Explainable AI // DARPA XAI Literature Review. 2019. – URL: https://arxiv.org/pdf/1902.01876.pdf

Mueller et al. 2020 – *Mueller S. T., Veinott E. S., Hoffman R.R., Klein G., Alam L., Mamun T., Clancey W.J.* Principles of Explanation in Human-AI Systems // Association for the Advancement of Artificial Intelligence. 2020. – URL: https://arxiv.org/pdf/2102/2102.04972.pdf

Mumford, Anjium 2013 – *Mumford S., Anjium R.L.* Causation. A Very Short Introduction. – Oxford: Oxford University Press, 2013.

Nolin 2016 – *Nolin J., Olson N.* The Internet of Things and Convenience // Internet Research. 2016. Vol. 26. No. 2. P. 360–376.

Orrell, Houshmand 2022 – *Orrell D., Houshmand M.* Quantum Propensity in Economics // Frontiers in Artificial Intelligence. 2022. Vol. 4. Art. 772294.

Polya 1954 – *Polya G.* Mathematics and Plausible Reasoning. – Princeton, NJ: Princeton University Press, 1954.

Raikov 2021 – *Raikov A.* Cognitive Semantics of Artificial Intelligence: A New Perspective. – Singapore: Springer, 2021.

Raikov 2022 – *Raikov A.* Automating Cognitive Modelling Considering Non-Formalisable Semantics // Intelligent Sustainable Systems (Lecture Notes in Networks and Systems. Vol. 334). – Singapore: Springer, 2022.

Rauber, Trasarti, Gianotti 2019 – *Rauber A., Trasarti R., Gianotti F.* Transparency in Algorithmic Decision Making // ERCIM News. 2019. Vol. 116. P. 10–11. – URL: https://ercim-news.ercim.eu/en116/special/transparency-in-algorithmic-decision-making-introduction-to-the-special-theme

The BIG Bell… 2018 – *The BIG Bell Test Collaboration.* Challenging Local Realism with Human Choices // Nature. 2018. Vol. 557. No. 7704. P. 212–216.

True Visions… 2006 – True Visions: The Emergence of Ambient Intelligence / ed. by E.H.L. Aarts, J.L. Encarnação. – Berlin: Springer, 2006.

Veličković et al. 2019 – *Veličković P., Ying R., Padovano M., Hadsell R., Blundell C.* Neural Execution of Graph Algorithms // 33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, 2019. – URL: https://grlearning.github.io/papers/88.pdf

Wang et al. 2020 – *Wang J., Li Z., Long Q., Zhang W., Song G., & Shi C.* Learning Node Representations from Noisy Graph Structures // 2020 IEEE International Conference on Data Mining (ICDM). – Los Alamitos, CA: IEEE Computer Society. P. 1310–1315.

REFERENCES

Abrosimov V.K. & Raikov A.N. (2022) *Intelligent Agricultural Robots.* Moscow: Kar'era Press (in Russian)

Vagin V.N., Golovina E.Y., Zagoryanskaya A.A., & Fomina M.V. (2004) *Authentic and Plausible Inference in Intelligent Systems* (V.N. Vagina & D.A. Pospelov, Eds.). Moscow: Fizmatlit (in Russian).

Dubrovsky D.I. (2021) The Task of Creating General Artificial Intelligence and the Problem of Consciousness. *Russian Journal of Philosophical Sciences = Filosofskie nauki.* Vol. 64, no. 1, pp. 13–44 (in Russian).

Lepskiy V.E. (2021) Artificial Intelligence in Subjective Control Paradigms. *Russian Journal of Philosophical Sciences = Filosofskie nauki.* Vol. 64, no. 1, pp. 88–101 (in Russian).

Raikov A.N. (2009) Prominences of Macroeconomics. *Ekonomicheskie strategii.* 2009. No. 7, pp. 42–49 (in Russian).

Aarts E.H.L. & Encarnação J.L. (Eds.) (2006) *True Visions: The Emergence of Ambient Intelligence*. Berlin: Springer.

Adadi A. & Berrada M. (2018) Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). *IEEE Access.* Vol. 6, pp. 52138–52160.

Arrieta A.B., Díaz-Rodríguez N., Del Ser J., Bennetot A., Tabik S., Barbado A., García S., Gil-López S., Molina D., Benjamins R., Chatila R. (2020) Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges towards Responsible AI. *Information Fusion.* Vol. 58, pp. 82–115.

Byrne R.M.J. (2019) Counterfactuals in Explaining Artificial Intelligence (XAI): Evidence from Human Reasoning. *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence (IJCAI-19)*, pp. 6276–6282.

Byrne R.M.J. (2019) Counterfactuals in Explaining Artificial Intelligence (XAI): Evidence from Human Reasoning. In: *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence (IJCAI-19)* (pp. 6276–6282). California: *International Joint Conferences on Artificial Intelligence*.

Chen M., Wei Z., Huang Z., Ding B., & Li Y. (2020) Simple and Deep Graph Convolutional Networks. *Proceedings of Machine Learning Research.* Vol. 119, pp. 1725–1735.

Doshi-Velez F. & Kortz M. (2017) Accountability of AI Under the Law: The Role of Explanation. *Berkman Klein Center Working Group on Explanation and the Law, Berkman Klein Center for Internet & Society Working Paper.* Retrieved from http://nrs.harvard.edu/urn-3:HUL.InstRepos:34372584

Einstein A., Podolsky B., & Rosen N. (1935) Can Quantum-Mechanical Description of Physical Reality Be Considered Complete? *Physical Review.* Vol. 47, no. 10, pp. 777–780.

Heaven W. D. (2020) Why asking an AI to explain itself can make things worse. *MIT Technology Review.* January 29. Retrieved from https://www.technologyreview.com/2020/01/29/304857/why-asking-an-ai-to-explainit-self-can-make-things-worse/

Kaul N. (2022) 3Es for Al: Economics, Explanation, Epistemology. *Frontiers in Artificial Intelligence.* Vol. 5, article 833238.

Leavy S., Meaney G., Wade K., & Greene D. (2020) Mitigating Gender Bias in Machine Learning Sata Sets. In: *International Workshop on Algorithmic Bias in Search and Recommendation* (pp. 12–26). Cham: Springer.

Leavy S., Siapera E., & O'Sullivan B. (2021) Ethical Data Curation for AI: An Approach based on Feminist Epistemology and Critical Theories of Race. In: *AIES'21: Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society* (pp. 695–703). New York: The Association for Computing Machinery.

Lepskiy V. (2018) Evolution of Cybernetics: Philosophical and Methodological Analysis. *Kybernetes*. Vol. 47, no. 2, pp. 249–261.

Lin Y.T., Hung T.W., & Huang L.T.L. (2021) Engineering Equity: How AI Can Help Reduce the Harm of Implicit Bias. *Philosophy and Technology*. Vol. 34, no. 1, pp. 65–90.

Liu R., Balsubramani A., Zou J. (2019) Learning transport cost from subset correspondence. *arXiv*. Retrieved from https://arxiv.org/pdf/1909.13203.pdf

Madry A., Makelov A., Schmidt L., Tsipras D., & Vladu A. (2017) Towards deep learning models resistant to adversarial attacks. *arXiv*. Retrieved from https://arxiv.org/pdf/1706.06083.pdf

Mueller S.T., Hoffman R.R., Clancey W. Klein G. (2019) Explanation in Human-AI systems: A Literature Meta-Review Synopsis of Key Ideas and Publications and Bibliography for Explainable AI. *DARPA XAI Literature Review.* Retrieved from https://arxiv.org/pdf/1902.01876.pdf

Mueller S.T., Veinott E.S., Hoffman R.R., Klein G., Alam L., Mamun T., & Clancey W.J. (2020) Principles of Explanation in Human-AI Systems. *Association for the Advancement of Artificial Intelligence*. Retrieved from https://arxiv.org/pdf/2102.04972.pdf

Mumford S. & Anjium R.L. (2013) *Causation. A Very Short Introduction.* Oxford: Oxford University Press.

Nolin J. & Olson N. (2016) The Internet of Things and Convenience. *Internet Research*. Vol. 26, no. 2, pp. 360–376.

Orrell D. & Houshmand M. (2022) Quantum Propensity in Economics. *Frontiers in Artificial Intelligence.* Vol. 4, art. 772294.

Polya G. (1954) *Mathematics and Plausible Reasoning.* Princeton, NJ: Princeton University Press.

Raikov A. (2021) *Cognitive Semantics of Artificial Intelligence: A New Perspective.* Singapore: Springer.

Raikov A. (2022) Automating Cognitive Modelling Considering Non-Formalisable Semantics. In: Nagar A.K., Jat D.S., Marín-Raventós G., & Mishra D.K. (Eds) *Intelligent Sustainable Systems* (Lecture Notes in Networks and Systems. Vol. 334). Singapore: Springer.

Rauber A., Trasarti R., & Gianotti F. (2019) Transparency in Algorithmic Decision Making. *ERCIM News.* Vol. 116, pp. 10–11. Retrieved from https://ercim-news.ercim.eu/en116/special/transparency-in-algorithmic-decision-making-introduction-to-the-special-theme.

The BIG Bell Test Collaboration. (2018) Challenging Local Realism with Human Choices. *Nature*. Vol. 557, no. 7704, pp. 212–216.

Veličković P., Ying R., Padovano M., Hadsell R., & Blundell C. (2019) Neural Execution of Graph Algorithms. In: *33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, 2019.* Retrieved from https://grlearning.github.io/papers/88.pdf

Wang J., Li Z., Long Q., Zhang W., Song G., & Shi C. (2020) Learning Node Representations from Noisy Graph Structures. In: *2020 IEEE International Conference on Data Mining* (pp. 1310–1315). Los Alamitos, CA: IEEE Computer Society.

Социогуманитарные основания цифровой трансформации

DOI: 10.30727/0235-1188-2022-65-1-91-108
Оригинальная исследовательская статья
Original research article

Философско-методологические основания совершенствования цифровой трансформации и внедрения искусственного интеллекта*

В.Е. Лепский
Институт философии РАН, Москва, Россия

Аннотация

В настоящее время нарастает процесс цифровой трансформации и внедрения искусственного интеллекта (ИИ) в широкий спектр социальных систем. Как правило, уделяется недостаточно внимания оценке социальных последствий от такого рода инноваций. Базовые причины связаны с доминированием модели техногенной цивилизации, воплощением которой является технократический подход, и использованием этого подхода в интересах глобалистского проекта. В разработках и внедрении цифровых технологий и ИИ возникает онтологический парадокс, для преодоления которого актуальна проблема разработки адекватных философско-методологических оснований оценки социальных инноваций, использующих цифровые технологии и ИИ. В статье обосновывается целесообразность использования трех типов научной рациональности (классика, неклассика, постнеклассика) для преодоления ограничений западной модели техногенной цивилизации и использование соответствующего этой рациональности субъектного подхода. Принципиально важно, что три типа научной рациональности соответствуют ключевым этапам эволюции кибернетики и ИИ. Эволюция ИИ проанализирована с этих позиций, и предложен подход к преодолению онтологического парадокса в цифровых трансформациях и внедрении ИИ. В контексте развития представлений о научной рациональности рассмотрена специфика инновационных моделей, использующих цифровые технологии и ИИ. Обоснована проблема становления интегративной области знания как эргономики цифровых трансформаций и ИИ, что позволит учесть богатый эрго-

* Работа поддержана Российским научным фондом (РНФ), грант № 21-18-00184 «Социогуманитарные основания критериев оценки инноваций, использующих цифровые технологии и искусственный интеллект».

номический опыт многокритериальной социогуманитарной оценки использования средств вычислительной техники и программного обеспечения: продуктивность, безопасность, удовлетворенность и развитие. Рассмотрены базовые позиции конфигуратора оценки инноваций, использующих цифровые технологии и ИИ, включающие научно-методическое и организационное обеспечение и заинтересованных субъектов.

Ключевые слова: цифровые технологии, философия искусственного интеллекта, этика искусственного интеллекта, философия техники, научная рациональность, кибернетика, эргономика.

Лепский Владимир Евгеньевич – доктор психологических наук, главный научный сотрудник сектора междисциплинарных проблем научно-технического развития Института философии РАН.

velepskiy@mail.ru

https://orcid.org/0000-0002-6893-0234

Для цитирования: *Лепский В.Е.* (2022) Философско-методологические основания совершенствования цифровой трансформации и внедрения искусственного интеллекта // Философские науки. 2022. Т. 65. № 1. С. 91–108. DOI: 10.30727/0235-1188-2022-65-1-91-108

Philosophical and Methodological Foundations for Improving Digital Transformation and Implementing Artificial Intelligence*

V.E. Lepskiy

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

Abstract

Nowadays, there is an evolving process of digital transformation and the introduction of artificial intelligence (AI) into a wide range of social systems. Usually, insufficient attention is paid to assessing the social consequences of such innovations. The underlying causes of that are related to the dominance of the technogenic model of civilization, the embodiment of which is the technocratic approach, and the use of this approach in the interests of the globalist project. In the development and implementation of digital technologies and AI, an ontological paradox arises, for overcoming which it is required to develop adequate philosophical and methodological

* The work was supported by the Russian Science Foundation, grant no. 21-18-00184 "Social and humanitarian foundations for evaluation criteria for innovations based on digital technologies and artificial intelligence."

foundations for assessing social innovations based on digital technologies. The article discusses the expediency of using three types of scientific rationality (classics, non-classics, post-non-classics) to overcome the limitations of the Western model of technogenic civilization and the use of a subjective approach corresponding to this rationality. It is fundamentally important that the three types of scientific rationality correspond to the key stages in the evolution of cybernetics and AI. The evolution of AI is analyzed from these positions and an approach is proposed to overcome the ontological paradox in digital transformations and the implementation of AI. In the context of the development of ideas on scientific rationality, the author considers the specifics of innovative models based on digital technologies and AI. The article examines the problem of the formation of an integrative field of knowledge as the ergonomics of digital transformations and AI, which will allow to take into account the rich ergonomic experience of a multi-criteria socio-humanitarian assessment of the use of computer technology and software: productivity, safety, satisfaction, and development. In the conclusion, the article considers the basic positions of the *configurator,* that is, of the devise for assessing innovations based on digital technologies and AI, including assessing of scientific, methodological and organizational issues and persons concerned.

Keywords: digital technologies, philosophy of artificial intelligence, ethics of artificial intelligence, philosophy of technology, scientific rationality, cybernetics, ergonomics.

Vladimir E. Lepskiy – D.Sc. in Psychology, Chief Research Fellow, Department of Interdisciplinary Problems in the Advance of Science and Technology, Institute of Philosophy, Russian Academy of Science.
velepskiy@mail.ru
https://orcid.org/0000-0002-0590-4020

For citation: Lepskiy V.E. (2022) Philosophical and Methodological Foundations for Improving Digital Transformation and Implementing Artificial Intelligence. *Russian Journal of Philosophical Sciences = Filosofskie nauki*. Vol. 65, no. 1, pp. 91–108. DOI: 10.30727/0235-1188-2022-65-1-91-108

Введение

В настоящее время оценка результатов внедрения цифровых технологий и искусственного интеллекта (ИИ) в социальные системы и разработка соответствующих нормативных документов, как правило, базируются на сложившихся подходах, обычно используемых для оценки цифровых технологий при проектировании технических систем. Проблема в том, что категориальный аппарат и базовые понятия из парадигм и онтологий социальных

систем пытаются использовать в парадигмах и онтологиях систем ИИ («этика ИИ», «доверие ИИ» и др.). Фактически в разработках и внедрении ИИ имеет место онтологический парадокс, для преодоления которого необходима разработка адекватных философско-методологических и социогуманитарных оснований оценки социальных инноваций, использующих цифровые технологии и ИИ.

Для системной оценки такого рода инноваций необходимо учитывать широкий спектр разнообразных субъектных позиций: миропроектов и стратегических проектов, государств и их объединений, цивилизаций, научных подходов, а также подходов, сложившихся в прикладных областях знания, практических сферах деятельности и др. Актуализированные субъектные позиции находятся в определенных отношениях и влияют друг на друга в зависимости от их сочетания и сложившихся ситуаций. Как следствие, важным шагом является разработка философско-методологической платформы (конфигуратора [Лефевр 1973]) взаимосвязанных позиций для анализа социальных инноваций и выработки критериев их оценки. Это позволит создать инструментарий, инвариантный к различным сферам использования цифровых технологий и ИИ, этот инструментарий можно адаптировать для оценки конкретных инноваций.

В данной статье проанализированы актуальные основания неконтролируемого бума цифровых трансформаций и ИИ, связанные с доминированием западной модели техногенной цивилизации и глобалистским проектом.

Предлагаются отдельные базовые философско-методологические основания для построения конфигуратора (системной платформы) оценки инноваций в социальных системах с использованием цифровых технологий и ИИ. Обосновывается целесообразность использования представлений о трех типах научной рациональности [Степин 2003] как системе субъектных парадигм [Lepskiy 2021], отвечающей принципу соответствия Н. Бора [Bohr 1976] и идеям формирования новых научных парадигм Т. Куна [Kuhn 1962]. Принципиально важно, что три типа научной рациональности соответствуют ключевым этапам эволюции кибернетики и ИИ [Lepskiy 2018], задают основания для критического анализа модели техногенной цивилизации и ориентиры для поиска подходов к построению модели посттехногенной цивилизации. Обосновывается актуальность проблемы становле-

ния эргономики цифровой трансформации и ИИ, которая должна обеспечивать междисциплинарную и трансдисциплинарную организацию и оценку социальных инноваций, использующих цифровые технологии и ИИ.

Кризис техногенной цивилизации

В начале XXI века наблюдается бум цифровой трансформации и ИИ. Вызовы и угрозы, создаваемые этим бумом, далеко не всегда должным образом контролируются и имеют порой весьма негативные последствия для будущего человечества. Бесконтрольность процессов цифровой трансформации и внедрения ИИ определяют две наиболее значимые причины:

- сложившаяся под влиянием западной цивилизации доминирующая модель техногенной цивилизации;
- активизация лидеров глобалистского проекта, использующих в своих интересах цифровую трансформацию и ИИ на основе модели техногенной цивилизации.

Как следствие, имеет место технократический подход к оценке последствий использования цифровых технологий и ИИ, игнорирование социальных ценностей и социальных аспектов внедрения такого рода технологий. Преодоление негативных последствий сложившейся ситуации возможно на основе использования адекватных философско-методологических оснований, которые позволили бы преодолеть кризис модели техногенной цивилизации и создать альтернативные глобалистскому новые миропроекты, ориентированные на гармонию и развитие человечества.

Кризис техногенной цивилизации прежде всего связан с ограниченностью ее базовых ценностей (научно-технического прогресса и науки), игнорированием социальных ценностей [Степин 1989], ориентацией на унификацию локальных цивилизаций, свертывание многоцивилизационного мира. Ограниченность такого подхода проявляется в нарастании угроз для человека, общества, экологии и собственно техносферы.

В техногенной цивилизации игнорируется субъектно-ориентированный подход, что проявляется в ориентации на атомизацию обществ, на снижение уровня субъектности от мировых организаций и государств до отдельных граждан. Это проявляется и в ориентации на создание унифицированного цифрового мира, фактически цифровая трансформация выступает как инструмент рефлексивного управления со стороны организаторов глобалистского проекта.

Как следствие, актуальна проблема создания модели посттехногенной цивилизации, ориентированной на всесторонний учет социальных ценностей в организации процессов цифровой трансформации. Эта проблема была поставлена В.С. Степиным [Степин 1989], который предложил гуманистическое видение научно-технического прогресса и соответствующие философско-методологические основания для становления посттехногенной цивилизации.

Целесообразно включить ценности науки и научно-технического прогресса в систему социальных ценностей: сохранения и развития человека, человечества, биосферы и техносферы. Принципиально важно отметить, что цифровая трансформация и ИИ оказывают влияние на все выделенные социальные ценности.

Сохранение и развитие человека. В многочисленных научных исследованиях выделяются негативные последствия бесконтрольного использования цифровых технологий и ИИ на деятельностную, коммуникативную и рефлексивную активность человека, возникновение угрозы расчеловечевания:

- разрушение целостности [Лекторский 2010] субъектности;
- деформация мышления (клиповое мышление и др.);
- деформация потребностно-мотивационной сферы, примитивизация ценностных ориентаций;
- блокировка рефлексии и критического анализа поступающей информации;
- снижение креативности;
- открытость к манипулятивным воздействиям, формирование повышенной конформности;
- кибербуллинг, провоцирование асоциальных форм поведения и др.;
- неосознаваемое повышение уровня риска принимаемых решений;
- блокировка эмпатии, деформация коммуникативных процессов;
- разрушение традиционных механизмов идентификации;
- формирование зависимости от цифровой реальности, отрыв от естественной социальной реальности, погружение в виртуальную реальность, интернет-аддикция и др.

Сохранение и развитие человечества. Использование цифровых технологий и ИИ как инструментов разрушения государств, унификации цивилизаций и переход рычагов управления в руки

транснациональных корпораций и международных финансовых структур способствует нарастанию угроз сохранению и развитию человечества.

Важным аспектом оценки цифровой трансформации и ИИ является учет цивилизационных и культурных идентичностей заказчиков, разработчиков и пользователей ИИ. Например, принципиально отличаются подходы к ИИ у специалистов в США и Европы с одной стороны и в Китае – с другой. У первых доминирует либеральный подход с ярко выраженным превалированием интересов индивидов над общественными интересами, у вторых – ориентация на коллективизм с жестким контролем над индивидами. Оба подхода вступают в противоречие со спецификой российской цивилизации. Как следствие актуален поиск адекватного подхода с учетом российской цивилизационной специфики. Эти соображения дают основания для утверждения, что некорректно рассмотрение универсальной этики ИИ.

Активизация лидеров глобалистского проекта, использующих в своих интересах цифровые трансформации и ИИ

Современная эпоха характеризуется свертыванием проекта глобализации и закатом однополярного мира. Активизируются лидеры глобалистского проекта, которые, используя модель техногенной цивилизации, сконцентрировали свои усилия на использовании технологий цифровых трансформаций и ИИ. Успешность такого рода воздействий на общество связана с возможностями организовывать весьма эффективные процессы рефлексивного управления разнообразными субъектами мирового сообщества и населением в целом [Schwab, Malleret 2020].

Цифровая трансформация и ИИ весьма успешно используются для дальнейшей поляризации обществ на «избранных» и «изгоев» за счет увеличения разрыва в доходах, широкого использования цифровых технологий, ориентированных на расчеловечивание большей части населения, разрушение института семьи, разрушение системы образования и др. Одним из ярких примеров является проект метавселенной, платформы, которая представляет собой мощный инструмент воздействия на человечество в интересах глобалистского проекта.

Одновременно снижается роль государства, из-под контроля государства выводятся финансовые потоки и механизмы валютного

регулирования, другие экономические механизмы и т.п. В целом формируются условия для монопольного управления мировым сообществом лидерами глобалистского проекта, представителями транснациональных корпораций и мировых банковских структур. Цифровые технологии и ИИ оказываются эффективными инструментами реализации глобалистского проекта.

Базовые философско-методологические основания

В качестве базовых философско-методологических оснований оценки инноваций, использующих цифровые технологии и ИИ, мы предлагаем использовать систему научной рациональности: классическую, неклассическую, постнеклассическую, предложенную В.С. Степиным [Степин 2003, 619–636]. Это обосновывается следующими соображениями.

Во-первых, систему трех типов научной рациональности можно рассматривать как систему парадигм [Lepskiy 2021], отвечающую принципу соответствия Н. Бора [Bohr 1976] и идеям формирования новых научных парадигм Т. Куна [Kuhn 1962], что обеспечивает связность парадигм, эволюция которых соотносится с типами научной рациональности.

Во-вторых, три типа научной рациональности соответствуют ключевым этапам эволюции кибернетики [Lepskiy 2018, Lepskiy 2021]. В свою очередь, эволюция кибернетики имеет тесные связи с эволюцией цифровых технологий и ИИ. Как следствие, по мере эволюции кибернетики, цифровых технологий и ИИ обеспечивается соотнесение областей знания, научных подходов и трендов, связанных с типами научной рациональности:

• наблюдатель-актор в управлении (внешний, внешний и встроенный, распределенный);

• парадигмы управления («субъект – объект», «субъект – субъект», «субъект – метасубъект» / «субъект –полисубъектная среда»)

• базовые философские подходы (позитивизм, философский конструктивизм, гуманистический философский конструктивизм);

• базовые субъектные подходы (деятельностный, субъектно-деятельностный, субъектно-ориентированный);

• базовые виды активности (деятельностная, коммуникативная, рефлексивная);

• подходы к механизмам интеграции областей знания (монодисциплинарный, междисциплинарный, трансдисциплинарный);

• виды управления (классическое, рефлексивное, через воздействия на саморазвивающуюся среду);

• базовые модели в управлении (аналитические, функционально-структурные, человеко-размерные среды);

• базовые механизмы управления (иерархии, сети, среды);

• базовые представления о знаниях в управление (информация, личностное знание, активные субъектно-соотнесенные знания);

• базовые этические регуляторы в управление (этика целей, коммуникативная этика, этика стратегических субъектов);

• базовые виды рефлексии (личностная рефлексия, коммуникативная рефлексия, метарефлексия).

В-третьих, в контексте постнеклассической рациональности проведен конструктивный анализ кризиса модели техногенной цивилизации [Степин 1989] и намечены ориентиры для поиска подходов к построению модели посттехногенной цивилизации.

Учет эволюции ИИ в контексте развития представлений о научной рациональности

Перспективные направления развития цифровых технологий и ИИ связаны с современными представлениями об эволюции сред гибридной реальности (субъектных, цифровых, физических).

В эволюции научной рациональности задаются различные системные основания для представления о средах гибридной реальности, опираясь на которые современная кибернетика исследует организацию и функционирование сред гибридной реальности с позиции трех взаимосвязанных типов научной рациональности (классика, неклассика, постнеклассика).

Классической научной рациональности соответствует классическая кибернетика (Н. Винер). В базовой методологической парадигме «субъект – объект» цифровые технологии и искусственный интеллект в основном обеспечивают моделирование объекта управления. Модель объекта управления называется цифровым двойником [Grieves 2014].

Неклассической научной рациональности соответствует кибернетика второго порядка (Ф. Ферстер). В базовой методологической парадигме «субъект – субъект» цифровые технологии и ИИ в основном обеспечивают моделирование активных объектов управления, которыми являются субъекты и псевдосубъекты [Wark 2019], а также моделирование их взаимодействий. Модель

объекта управления (субъекта) целесообразно называть «цифровой субъект» [Goriunova 2019].

Постнеклассической научной рациональности соответствует кибернетика третьего порядка (В.Е. Лепский). Базовая методологическая парадигма – «субъект – метасубъект». Под метасубъектом понимается саморазвивающаяся полисубъектная (рефлексивно-активная) среда, а его моделью выступает цифровой метасубъект.

Функционирование цифровых двойников, цифровых субъектов и цифровых метасубъектов в средах гибридной реальности задается системой онтологий и системой принципов саморазвивающихся полисубъектных сред (табл. 1).

Тип научной рациональности	**Базовые субъектные парадигмы**	**Базовые модели и парадигмы ИИ**	**Базовые кибернетические парадигмы**
Классическая	Субъект – Объект	Цифровой двойник. Частные парадигмы ИИ (*морфологическая парадигма, логическая парадигма, нейрокибернетическая парадигм*а, *имитационная парадигма и др.)*	Кибернетика первого порядка (наблюдаемых систем)
Неклассическая	Субъект – Субъект	Цифровой субъект. Сильный, общий ИИ, метавселенная	Кибернетика второго порядка (наблюдающих систем)
Постнеклассическая	Субъект – Метасубъект	Цифровой метасубъект. Глобальный, средовой ИИ	Кибернетика третьего порядка (саморазвивающихся полисубъектных (рефлексивно-активных) сред)

Табл. 1. Связь парадигм ИИ и кибернетики с парадигмами научной рациональности

Ориентация на преодоления «онтологического парадокса»

При исследовании возможностей и ограничений применения цифровых технологий и искусственного интеллекта (ИИ) возникают проблемы, связанные с разрывом парадигм, задающих

представления об ИИ без учета парадигм жизнедеятельности и развития социальных систем [Лепский 2021]. Инструменты ИИ разрабатываются в парадигмах ИИ (имеются в виду морфологическая парадигма, логическая парадигма, нейрокибернетическая парадигма, имитационная парадигмы, парадигмы слабого, сильного, общего ИИ и др.) и в контексте этих же парадигм организуются инновации в социальных системах. Такой подход соответствует логике техногенной цивилизации. Как следствие, возникают многочисленные негативные последствия внедрения ИИ и упускаются полезные решения, учитывающие парадигмы и онтологии обеспечения жизнедеятельности и развития социальных систем.

Актуальны проблемы соотнесения парадигм и онтологий социальных систем и ИИ, организации соответствующих технологических уровней и интерфейса между ними. Разработка этой проблемы ведется на основе субъектно-ориентированного подхода в рамках постнеклассической кибернетики саморазвивающихся полисубъектных (рефлексивно-активных) сред [Lepskiy 2018].

В моделях таких сред конвергенция естественнонаучных и гуманитарных подходов возможна на основе разработки двухуровневой технологической модели, включающей взаимосвязанные концептуально-технологический (гуманитарный) и инструментально-технологический (естественнонаучный) уровни (рис. 2). Конвергенция будет осуществляться на основе системы онтологий бытия активных элементов естественного и искусственного интеллекта. Принципиально важно отметить, что создаваемую методологическую платформу следует рассматривать как научную парадигму, а следовательно, она должна удовлетворять сложившимся в науке критериям новых научных парадигм, в частности

Рис. 1. Преодоление онтологического парадокса в разработке и использовании ИИ на технологическом уровне

сложившемуся в естественных науках принципу соответствия Бора и общенаучным идеям становления новых парадигм Т. Куна.

Фактически речь идет о том, что для использования ИИ в процессах управления необходима организация двух уровней технологического обеспечения, включающего в себя ИИ. Первый уровень – концептуально-технологический, непосредственно связан с парадигмами и онтологиями обеспечения жизнедеятельности и развития социальных систем. На этом уровне может быть проведено соотнесение со специализированными технологиями ИИ. Второй уровень – инструментально-технологический, его составляют технологии ИИ, созданные на основе любых парадигм ИИ. Актуальная, сложная проблема – установление интерфейса между этими двумя уровнями в интересах обеспечения жизнедеятельности и развития социальных систем [Lepskiy 2018].

Учет специфики инновационных моделей, использующих цифровые технологии и ИИ

В разработках цифровых технологий и ИИ доминирует культура и модели, соответствующие классической научной рациональности, которая проявляется в использовании функциональной и линейной моделей инноваций [Лепский 2016]. Система ИИ, соответствующая эпохе классической научной рациональности, представляет компьютерное устройство, программа, функционирующая автономно от человека в соответствии с определенным заданием. Вместе с тем взаимодействие систем ИИ с человеком порождают новую, гибридную реальность. Совсем не скоро, да и вообще вряд ли когда-нибудь, эмоциональный потенциал человека перекочует в машину. Еще труднее предполагать это в отношении бессознательного, трансцендентного, феномена сознания. Считается, что возможности языка и логики для осуществления процессов мышления ограничены, хотя и составляют их существенную часть. Таким образом, модели с доминированием классической научной рациональности оставляют ряд принципиально важных вопросов о субъекте развития, механизмах их идентификации и сборки, организации пространства коммуникаций и доверия и др. без рассмотрения.

Неклассической научной рациональности соответствуют нелинейные модели инноваций [Лепский 2016], где инновационный процесс не ограничивается только сферой цифровых технологий и ИИ и включает институциональные, организационные и управленческие инновации. Нелинейная модель множественных

источников инноваций ориентирована на механизм развития с максимальным учетом разнообразия участников инновационного процесса путем создания условий для их творческого взаимодействия. В настоящее время предпринимаются попытки внедрить нелинейные модели инноваций, использующие цифровые технологии и ИИ, но они пока не являются значимыми для данной области инноваций.

Постнеклассической научной рациональности соответствуют модели саморазвивающихся полисубъектных инновационных сред. В данный момент нам не известны подходы с использованием таких моделей к инновациям на основе цифровых технологий и ИИ [Лепский 2016].

Рассмотренные модели – функциональная и линейная (классика), нелинейная (неклассика) и модель саморазвивающихся полисубъектных инновационных сред (постнеклассика) – оказывают принципиально разное влияние на критерии оценки инноваций, использующих цифровые технологии и ИИ. Прежде всего, это связано с субъектными позициями, актуализируемыми в этих моделях, и, соответственно, с ценностными, целевыми ориентациями и этическими регуляторами у этих субъектов, с онтологиями, обеспечивающими деятельностные, коммуникативные и рефлексивные процессы инновационного развития. Современные подходы к инновационному развитию должны опираться на современные представления о научной рациональности, каковыми являются представления в рамках постнекласической рациональности.

Проблема становления эргономики цифровой трансформации и ИИ

Еще в 1980-е годы была сформулирована проблема становления эргономики средств вычислительной техники и программного обеспечения [Березкин и др. 1985]. Сегодня крайне актуальна проблема становления эргономики цифровых трансформаций и искусственного интеллекта.

В настоящее время в мировой и отечественной практике разработки и внедрения цифровых технологий и ИИ отсутствует социогуманитарное обеспечение этих процессов. При проектировании социотехнических систем комплексный учет человеческого фактора брала на себя эргономика. В конце XX века эргономисты, обладающие развитыми практиками комплексного учета человеческого фактора и поддерживающие культуру инженерии,

успешно содействовали решению разнообразных сложных социогуманитарных проблем (качества жизни, человеческого потенциала, организации сообществ в Интернете, информационно-психологической безопасности и др.).

На социогуманитарное обеспечение цифровых трансформаций и ИИ могли бы претендовать и другие области знания, но у эргономики есть характеристики, которые обеспечивают ей преимущество:

• ведущая ориентация на междисциплинарный подход;

• учет специфики полисубъектных систем (индивиды, группы, сообщества, организации, этносы, государства, общества и др.);

• многокритериальный подход при обеспечении конкретных проектов (продуктивность, безопасность, развитие, удовлетворенность и др.);

• высокая культура интеграции специалистов гуманитарного и естественнонаучного профиля;

• опыт крупномасштабных проектов с использованием вычислительной техники и программного обеспечения;

• высокая квалификация специалистов по системной интеграции;

• обеспечение прозрачности процессов разработки и внедрения цифровых технологий и ИИ, что существенно в условиях российских реалий для снижения коррупции.

Для оценки цифровых технологий и ИИ также важен опыт многокритериальной социогуманитарной оценки использования средств вычислительной техники и программного обеспечения в эргономике. Оцениваются продуктивность, безопасность, удовлетворенность и развитие: эффективное реагирование (продуктивность) на позитивные возможности инноваций, контролирующее реагирование (безопасность) на потенциальные угрозы от инноваций, адекватное реагирование на удовлетворенность различных слоев населения инновациями, развивающее реагирование на инновации, которое связано со способностью субъектов создать или иметь проект своего развития, видения будущего, и с этих позиций оценить инновации.

Становление эргономики цифровых технологий и ИИ – крайне актуальная проблема, в стране еще остались специалисты с опытом эргономического обеспечения крупномасштабных автоматизированных систем организационного управления, способных инициировать решение этой проблемы и руководить процессом решения.

Высокая сложность междисциплинарной оценки инноваций, использующих цифровые технологии и ИИ, обосновывает переход к трансциплинарному подходу, обеспечение которого может осуществить эргономика цифровой трансформации и ИИ.

Конфигуратор оценки инноваций, использующих цифровые технологии и ИИ

Для организации социогуманитарной платформы оценки инноваций, использующих цифровые технологии и ИИ, воспользуемся предложенной В.А. Лефевром идеей системного конфигуратора: исследователь проводит обоснованный отбор некоторых принципиально разных представлений об объекте исследования. Объект как бы проецируется на несколько экранов. Каждый экран задает свое собственное членение на элементы, порождая тем самым определенную структуру. Экраны связаны друг с другом так, что у нас имеется возможность соотносить различные картины. Подобное устройство, синтезирующее различные системные представления, Лефевр назвал конфигуратором [Лефевр 1973].

Разработка социогуманитарной платформы предполагает выделение системы позиций для оценки инноваций, описание этих позиций, их связей и др. Предварительно можно выделить базовые группы таких позиций. К научно-методическому и организационному обеспечение относятся:

- эргономика цифровых технологий и ИИ;
- философско-методологические основания и соответствующие специфике цифровых технологий и ИИ области знания;
- концептуальные документы и стандарты;
- центры экспертизы (международные, национальные, региональные, отраслевые и др.);
- культура и традиции прикладных сфер использования цифровых технологий и ИИ (экономика, образование, медицина и др.);
- цивилизационная специфика разработки и использования цифровых технологий и ИИ и др.

Субъектами такой социогуманитарной платформы являются:

- заинтересованные метасубъекты (цивилизации, общества, миропроектанты, государства, организации и др.);
- разработчики цифровых технологий и ИИ;
- пользователи цифровых технологий и ИИ и др.

Заключение

Основные причины недостаточного внимания к оценке социальных последствий цифровой трансформации и внедрения искусственного интеллекта в широкий спектр социальных систем связаны с доминированием модели техногенной цивилизации, воплощением которой является технократический подход и использование этого подхода в интересах глобалистского проекта. Такой подход порождает онтологический парадокс, для преодоления которого предлагается разработка философско-методологических оснований на основе системы трех типов научной рациональности с доминированием субъектного подхода, соответствующего постнеклассической научной рациональности. В развитие этого подхода проанализированы соответствующие модели инноваций, обосновано становление социогуманитарной эргономики цифровой трансформации и ИИ, предложены базовые позиции конфигуратора оценки инноваций, использующих цифровые технологии и ИИ, его научно-методическое и организационное обеспечение, указаны заинтересованные субъекты.

ЦИТИРУЕМАЯ ЛИТЕРАТУРА

Березкин и др. 1985 – *Березкин Б.С., Лепский В.Е., Мунипов В.М., Смолян Г.Л.* Эргономическое обеспечение проектирования программных средств // Эргономическое обеспечение проектирования средств вычислительной техники и АСУ / Труды ВНИИТЭ. Сер. Эргономика. Вып. 30. – М.: ВНИИТЭ, 1985. С. 8–19.

Лекторский 2010 – *Лекторский В.А.* Субъект в истории философии: проблемы и достижения // Методология и история психологии. 2010. Т. 5. Вып. 1. С. 5–18.

Лепский 2016 – *Лепский В.Е.* Инновационное развитие России: философский анализ // Философия науки и техники. 2016. Т. 21. № 1. С. 169–187.

Лепский 2021– *Лепский В.Е.* Искусственный интеллект в субъектных парадигмах управления // Философские науки. 2021. Т. 64. № 1. С. 88–101.

Лефевр 1973 – *Лефевр В.А.* Конфликтующие структуры. – М.: Советское радио, 1973.

Степин 1989 – *Степин В.С.* Научное познание и ценности техногенной цивилизации // Вопросы философии. 1989. № 10. С. 3–18.

Степин 2003 – *Степин В.С.* Теоретическое знание. – М.: Прогресс-Традиция, 2003.

Bohr 1976 – *Bohr N.* Collected Works. Vol. 3: The Correspondence Principle (1918–1923) / ed. by J.R. Nielsen. – Amsterdam: North-Holland Publishing, 1976.

Goriunova 2019 – *Goriunova O.* Digital Subjects: An Introduction // Subjectivity. 2019. Vol. 12. No. 1. P. 1–11.

Grieves 2014 – *Grieves M.* Digital Twin: Manufacturing Excellence through Virtual Factory Replication. – URL: https://www.researchgate.net/publication/275211047

Kuhn 1962 – *Kuhn T.S.* The Structure of Scientific Revolutions. – Chicago: University of Chicago Press, 1862.

Lepskiy 2018 – *Lepskiy V.* Evolution of Cybernetics: Philosophical and Methodological Analysis // Kybernetes. 2018. Vol. 47. No. 2. P. 249–261.

Lepskiy 2021 – *Lepskiy V.* Systems Analysis of the Foundations for the Formation of new Paradigms of Control // IFAC-PapersOnLine. 2021. Vol. 54. No. 13. P. 622–626.

Schwab, Malleret 2020 – *Schwab K., Malleret T.* COVID-19: The Great Reset. – Geneva: Forum Publishing, 2020.

Wark 2019 – *Wark S.* The Subject of Circulation: On the Digital Subject's Technical Individuations // Subjectivity. 2019. Vol. 12. No. 1. P. 65–81.

REFERENCES

Berezkin B.S., Lepskiy V.E., Munipov V.M., & Smolyan G.L. (1985) Ergonomic software engineering. In: *Ergonomic Support of Computer and ACS Design* (pp. 8–19). Moscow: VNIITE (in Russian).

Bohr N. (1976) *Collected Works. Vol. 3: The Correspondence Principle (1918–1923)* (J. R. Nielsen, Ed.). Amsterdam: North-Holland Publishing.

Goriunova O. (2019) Digital Subjects: An Introduction. *Subjectivity.* Vol. 12, no.1, pp. 1–11.

Grieves M. (2014) *Digital Twin: Manufacturing Excellence through Virtual Factory Replication.* Retrieved from https://www.researchgate.net/publication/275211047

Kuhn T.S. (1962) *The Structure of Scientific Revolutions*. Chicago: University of Chicago Press.

Lefebvre V.A. (1973) *Conflicting Structures.* Moscow: Sovetskoe radio (in Russian).

Lektorskiy V.A. (2010) Subject in the History of Philosophy: Problems and Achievements. *Metodologiya i istoriya psikhologii.* Vol 5, no. 1, pp. 5–18 (in Russian).

Lepskiy V.E. (2016) Innovative Development of Russia: A Philosophical Analysis. *Filosofiya nauki i tekhniki.* Vol. 21, no. 1, pp. 169–187.

Lepskiy V. (2018) Evolution of Cybernetics: Philosophical and Methodological Analysis. *Kybernetes.* Vol. 47, no. 2, pp. 249–261.

Lepskiy V.E. (2021a) Artificial Intelligence in Subjective Control Paradigms. *Filosofskie nauki = Russian Journal of Philosophical Sciences.* Vol. 64, no. 1, pp. 88–101 (in Russian).

Lepskiy V. (2021b) Systems Analysis of the Foundations for the Formation of new Paradigms of Control. *IFAC-PapersOnLine.* Vol. 54, no. 13, pp. 622–626.

Schwab K. & Malleret T. (2020) *COVID-19: The Great Reset.* Geneva: Forum Publishing.

Stepin V.S. (1989) Scientific Knowledge and Values of Technogenic Civilization. *Russian Studies in Philosophy.* No 10, pp. 3–18. (in Russian).

Stepin V.S. (2003) *Theoretical Knowledge.* Moscow: Progress-Traditsiya (in Russian).

Wark S. (2019) The Subject of Circulation: On the Digital Subject's Technical Individuations. *Subjectivity.* Vol. 12, no. 1, pp. 65–81.

DOI: 10.30727/0235-1188-2022-65-1-109-123
Оригинальная исследовательская статья
Original research article

Проблема аутопойезиса техногенной цивилизации и формирование ценностных основ применения цифровых технологий*

Е.В. Малахова
Национальный исследовательский ядерный университет «МИФИ», Москва, Россия,
Институт философии РАН, Москва, Россия

Аннотация

В работе ставится проблема способности к самовоспроизводству современной техногенной цивилизации и того, каким образом механизмы этого воспроизводства способны влиять на формирование аксиологических оснований использования порожденных этой цивилизацией цифровых технологий. Самовоспроизводство цивилизационных структур рассматривается через их постоянное повторение в процессе коммуникации, что приводит философов к необходимости введения специальных понятий для обозначения перечисленного. В существующих философских и социологических исследованиях, использующих системный подход, для описания подобных операций уже утвердилось использование термина «аутопойезис», введенного для этих целей в работах Н. Лумана. При рассмотрении аутопойезиса техногенной цивилизации для определения последней автор опирается на работы В.С. Степина. В результате проведенного исследования было выявлено, что существующие внутренние противоречия техногенной цивилизации, приводящие к ее кризису, в том числе ценностному, во многом происходят из-за того, что в ней одномоментно присутствуют взаимно некомплементарные аутопойетические структуры: 1) относящиеся к индустриальной и постиндустриальной эпохе и связанные с научным мировоззрением; 2) структуры, перешедшие в современность из доиндустриальной эпохи и до сих пор хранящие большую часть ценностного и культурного багажа современного общества. Таким образом, при формировании ценностных основ применения современных средств цифровизации, техногенная цивилизация вынуждена преодолевать разрывы в собственных аутопойетических

* Работа поддержана Российским научным фондом (РНФ), грант № 21-18-00184 «Социогуманитарные основания критериев оценки инноваций, использующих цифровые технологии и искусственный интеллект».

структурах, упомянутых выше, и создавать ценностные системы, свободные от указанных противоречий.

Ключевые слова: социальная философия, аксиология, техногенная цивилизация, ценности, цифровизация, научное мировоззрение, постнеклассическая научная парадигма.

Малахова Елена Владимировна – кандидат философских наук, доцент кафедры международных отношений Национального исследовательского ядерного университета «МИФИ», докторант сектора междисциплинарных проблем научно-технического развития Института философии РАН.

e.v.malahova@mail.ru

https://orcid.org/0000-0002-1829-8234

Для цитирования: *Малахова Е.В.* Проблема аутопойезиса техногенной цивилизации и формирование ценностных основ применения цифровых технологий // Философские науки. 2022. Т. 65. № 1. С. 109–123. DOI: 10.30727/0235-1188-2022-65-1-109-123

The Problem of Autopoiesis of Technogenic Civilization and the Formation of Value Base for the Use of Digital Technology*

E.V. Malakhova

National Research Nuclear University MEPhI, Moscow, Russia,

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

Abstract

The paper discusses the self-reproduction ability of the existing technogenic civilization and the issues of the influence of self-reproduction mechanisms on the formation of axiological grounds for the use of digital technologies generated by this civilization. The self-reproduction of civilizational structures is considered through their constant repetition in the process of communication. In existing philosophical and sociological studies based on systems approach, the term *autopoiesis*, introduced for these purposes in the works of N. Luhmann, has already been used to describe such processes. Considering the autopoiesis of the technogenic civilization, the articles relies on the works of V.S. Stepin to determine the main features of that civilization. As a result of the conducted research, it was revealed that the existing internal contradictions of the technogenic civilization that

* The work was supported by the Russian Science Foundation, grant no. 21-18-00184 "Social and humanitarian foundations for evaluation criteria for innovations based on digital technologies and artificial intelligence."

can lead and are already leading to its crisis, including the value one, are caused by simultaneous presence of mutually non-complementary autopoietic structures: (1) the ones related to the industrial and post-industrial era and to the scientific worldview; (2) and structures that have passed into modernity from the pre-industrial era, and yet still are have value and cultural significance in modern society. The author concludes that, forming the value foundations of the use of modern digital technologies, the technogenic civilization is forced to overcome the abovementioned gaps in its own autopoietic structures and create value systems free from these contradictions.

Keywords: social philosophy, axiology, technogenic civilization, values, digitalization, scientific worldview, postnonclassical scientific paradigm.

Elena V. Malakhova – Ph.D. in Philosophy, Associate Professor, Department of International Relations, National Research Nuclear University MEPhI; Postdoctoral Research Fellow, Department of Interdisciplinary Problems in the Advance of Science and Technology, Institute of Philosophy, Russian Academy of Sciences.

e.v.malahova@mail.ru

https://orcid.org/0000-0002-1829-8234

For citation: Malakhova E.V. (2022) The Problem of Autopoiesis of Technogenic Civilization and the Formation of Value Base for the Use of Digital Technology. *Russian Journal of Philosophical Sciences = Filosofskie nauki.* Vol. 65, no. 1, pp. 109–123. DOI: 10.30727/0235-1188-2022-65-1-109-123

Введение

Исследования цивилизационной проблематики нередко выходят на первый план в моменты кризисов и масштабных социальных процессов, ведущих к заметным социокультурным сдвигам. В такие периоды возникают закономерные вопросы о причинах и возможных результатах происходящего не только на микро- и макроуровне социологических моделей, но и на всеобъемлющем уровне, который охватывается по преимуществу философскими и культурологическими концепциями.

Именно на этих высших уровнях рассмотрения социальных процессов можно говорить о понятии цивилизации и ставить вопрос о том, каковы черты наличествующего цивилизационного типа [Степин 2011, 78–93], происходят ли текущие процессы строго в его рамках или же стремятся выйти за них, с тем, чтобы положить начало некоему новому типу цивилизационного развития.

В настоящей статье поставлены вопросы о том, каким образом может существовать, постоянно воспроизводясь, техногенная цивилизация; каким образом в ней существуют различные типы структур, которые в будущем способны привести к ее видоизменению; и наконец, как механизмы воспроизводства техногенного цивилизационного типа взаимосвязаны с формированием ценностных основ современных технологических инноваций, в особенности, цифровых.

Таким образом, настоящее исследование включает в себя два уровня.

1) Цивилизационный уровень, на котором рассматривается, каким образом техногенная цивилизация воспроизводится и видоизменяется. Здесь понятие аутопойезиса используется по преимуществу для описания цивилизации с системной точки зрения.

2) Аксиологический уровень, на котором можно увидеть, как ранее описанные механизмы цивилизационного воспроизводства связаны с ценностной составляющей, которая в свою очередь связана с оценкой технологических инноваций как важнейшей специфической составляющей техногенной цивилизации. На этом уровне также будут рассмотрены сбои и разрывы цивилизационного самовоспроизведения, приводящие к кризису техногенной цивилизации.

Методология исследования

Понятие и значение аутопойезиса. В качестве теоретико-методологического базиса в настоящей статье используется концепция аутопойезиса, разработанная Н. Луманом [Луман 2011].

Луман не только создал одну из наиболее влиятельных концепций для современной социологии и социальной философии. Важно отметить: он рассматривает общество с точки зрения системного подхода, получившего в XX веке чрезвычайную популярность из-за своего высочайшего эвристического потенциала. Более того, Луман одновременно рефлексирует по поводу применимости, возможностей и границ этого подхода. В частности, он стремится преодолеть критически оцениваемые особенности системного и структурного анализа общества, доставшиеся «в наследство» от структурного функционализма – а именно, отношение к структурам как к чему-то статичному и определенному,

когда такой подход противопоставлялся рассмотрению динамических аспектов социального развития.

Основное понятие, на которое опираемся как на относящееся к способности социокультурных структур к самоповторению и самовоспроизводству, – *аутопойезис* – Луман напрямую связывает не только с социальными структурами, но и с понятием коммуникации, основополагающим для всех его концептуальных построений.

Изначально понятие аутопойезиса было введено У. Матураной и Ф. Варелой [Maturana, Varela 1980] для объяснения феноменов биологического самовоспроизводства и эволюции живых организмов. В этом значении данная теория, несмотря на объяснительный потенциал, нередко подвергалась критике именно из-за философских, с точки зрения ряда ученых, коннотаций. Однако в том, в чем биологи видели недостаток, ученые-гуманитарии смогли открыть достоинство и использовать концепцию аутопойетического воспроизведения в отношении социальных и коммуникативных структур, как это сделал Луман [Šubrt 2019]. Он же показал, что критика данного понятия с содержательной позиции некорректна, так как оно может и должно применяться именно с чисто методологической позиции. Так, Луман пишет, что понятие аутопойезиса используется не для того, чтобы объяснить, какие именно структуры существуют и почему, но для того, чтобы установить объяснительный принцип, который может применяться в практически любых структурах, по крайней мере, в социальной сфере, хотя сам ученый и не делает подобного различения [Луман 2011, 68].

Вообще, использование понятия аутопойезиса – это, на первый взгляд, очень изящное методологическое решение для описания самореферентных структур, не имеющих опоры на что-то извне, однако связанных с окружающей средой множеством связей и способных приспосабливаться к этой среде, при необходимости меняясь так, чтобы при этом достаточно долго сохранять самотождественность [Chettiparamb 2020]. Это понятие также позволяет решить проблему, на которую в свое время обратил внимание Вебер: вопрос свободы исследователя от ценностей, «свободы от оценки» в социальных науках. Решение проблемы,

если обратиться к терминологии Степина, состоит в переходе из классической парадигмы научной рациональности к неклассической и в перспективе к постнеклассической [Степин 1989]. Анализируя неклассическую и постнеклассическую парадигмы, ученый остается в рамках собственной ценностной системы, которая встроена в его исследовательскую парадигму, и определяет принципы научной этики, конституирует целеполагание. В рамках аутопойетической концепции, как настаивает Луман, единый инвариантный принцип должен применяться как для объясняемой, так и для объясняющей систем, так как они существуют и воспроизводятся одинаково. Сам исследователь, в конце концов, тоже не более (и не менее), чем система, социальные параметры которой возникают и воспроизводятся ровно по тем же законам, что и весь остальной изучаемый им социум [Clarke 2019].

Очень любопытно, что с аутопойезисом оказывается связано понятие неопределенности, которая нарастает, когда система (в данном случае, опять же, именно социальная) отделяется от породившей ее среды и начинает самовоспроизводиться, приобретая, по терминологии Лумана, операциональную закрытость – как невозможность использовать какие бы то ни было операции, кроме собственных [Луман 2011, 69]. Таким образом, система утрачивает связи с операциями внешней среды, и для того, чтобы ничем более извне не детерминированные операции могли воспроизводиться, им приходится приписывать, иногда весьма произвольным или случайным образом, категорию смысла, порождая такую форму операциональной деятельности, как коммуникация.

Кроме того, Луман настаивает на том, что аутопойетические социальные системы способны эволюционировать и видоизменяться, двигаясь в сторону усложнения или редукции. Конечно, можно предположить, что таким эволюционизмом концепция Лумана отчасти обязана идеям Матураны и Варелы. Но это все же одна из возможностей объяснить видоизменение социальных структур в рамках единой методологии, основанной на системном подходе и стремящейся охватить равно социальную статику и динамику.

Таким образом, использование понятия аутопойезиса дает возможность говорить о самовоспроизводстве сложных социальных систем, к высшему уровню которых относятся и цивилизационные

типы. Однако, поскольку такой подход формален в том смысле, что не затрагивает содержательной части воспроизводимых систем, для рассмотрения последней мы должны будем обратиться к ценностной составляющей цивилизационных типов. Именно она представляет собой системообразующий принцип, через определенное отношение к реальности формируя нормативные комплексы, задавая возможность целеполагания и модусы коммуникации.

Понятие ценности и ценностного отношения

Несмотря на то, что в данной статье не ставится задача углубленного рассмотрения аксиологической проблематики (она достойна как минимум отдельной работы), тем не менее без прояснения того, что понимается под ценностью и ценностным отношением, продолжать дальнейшее рассуждение было бы невозможно.

Определение ценности во многом затрудняется тем, что это не только философский термин, но и общеупотребительное понятие, имеющее в зависимости от контекста огромное множество коннотаций. При развитии философской аксиологии как учения о ценностях было сделано много попыток определить ценности через блага, цели, нормы, интересы, потребности – и со временем все это предсказуемо не выдерживало критики [Шохин 2006]. Таким образом, поэтапно отказываясь от универсалистских и объективистских трактовок ценности, можно прийти к персоналистскому пониманию ценности как сосредоточенной именно в субъекте. Но если рассуждать таким образом, то ценность невозможна не только без субъекта как такового, но и без той операции, при помощи которой субъект ее создает и воссоздает – операции оценивания как придания ценности.

В результате, ценность оказывается своего рода неразрывной связью субъекта и объекта через операцию придания ценности [Каган 1997]. Кроме того, субъект не всегда может быть исключительно индивидуальным (отдельной личностью); он часто оказывается также коллективным [Лепский 2016]. Если говорить о ценностях в цивилизационном разрезе, то это преимущественно будут ценности коллективных субъектов. Последние, постоянно аутопойетически воспроизводя собственные ценности, под-

держивают существование отдельных культурных сегментов и цивилизации в целом как системы.

Аутопойезис техногенной цивилизации

Ключевые особенности техногенного цивилизационного типа. Техногенная цивилизация – это достаточно молодой (вероятно, самый молодой из существующих) цивилизационный тип. Ее появление потребовало существенного подготовительного периода, пришедшегося, в основном, на Ренессанс и начало Нового времени, когда зарождается то, что можно назвать научным мировоззрением [Степин, Горохов, Розов 1995].

Связь техногенной цивилизации и научного мировоззрения более чем очевидна. Бесспорно, что существенное количество сведений о реальности задолго до этого было накоплено в рамках философского, а иногда даже и мифологического мировоззрения, но для развития технологий и массового внедрения их в практику потребовался именно мировоззренческий сдвиг в сторону признания научного знания объективно истинным по преимуществу.

Техногенная цивилизация качественно отлична от других традиционных цивилизационных типов. Несходства носят самый фундаментальный характер, так как основаны на различных системах ценностей.

Когда мы говорим, что техногенная цивилизация основана на научном мировоззрении, то в первую очередь обращаем внимание на то, что в ней должны превалировать исходные для этого мировоззрения ценности, которые и будут определять векторы дальнейшего развития, а также обоснования для целеполагания и оценки полученных результатов. Для научного мировоззрения основные ценности – это, в первую очередь, истина (в классическом смысле термина) и научная новизна. Возможно, именно отсюда проистекает стремление воспринявшей все это техногенной цивилизации к унификации и экспансивности. Научная истина объективна, поэтому рассматривается как всеобщая, унифицированная и не зависящая от индивидуальных и даже социокультурных факторов. Можно сказать, что все это могло бы относиться в большей мере к классической и отчасти неклассической науке, в то время как в постнеклассической научной парадигме декларируется включе-

ние в науку той социально-ценностной составляющей, которая даже теперь имеет пока что внешнюю по отношению к научному знанию природу. При этом до сих пор не ясно до конца, каким образом этого можно достичь, чтобы не размывались границы самой науки, что, по мнению самого же Степина и его соавторов было бы концом научного знания [Степин, Горохов, Розов 1995, 372]. Тем не менее техногенная цивилизация складывалась в большей степени под влиянием классической парадигмы и до сих пор испытывает на себе значительное воздействие последней, поскольку внедрение и распространение ценностей науки в рамках массово распространенного мировоззрения, как правило, существенно запаздывает по отношению к их появлению в самой науке.

Ценностные различия цивилизационных типов. Культура в целом, как известно, развивается достаточно неравномерно. Если наука и техника сейчас находятся в ее авангарде, то другие сферы и социальные институты, действующие на основе вненаучных ценностей, видоизменяются намного медленнее. Более того, попытки внедрить туда ценности науки наталкиваются на противодействие, иногда неосознанное, иногда целенаправленное, именно из-за несоответствия ценностных систем научных и вненаучных, когда последние являются, по большей части, донаучными.

Здесь необходимо провести хотя бы краткий сравнительный анализ цивилизационных типов и их превалирующих ценностных систем.

Итак, если основные максимы техногенной цивилизации связаны с научным мировоззрением и ценностями науки – истиной и новизной, – то это само по себе уже заставляет задаться вопросом о том, откуда должны происходить все остальные ценности, необходимые для социального и личностного бытия. Отвечая на этот вопрос, необходимо обратиться к более ранним ценностным системам, развивавшимся до науки, а позднее – параллельно с ней. Эти ценностные системы будут принадлежать традиционным цивилизациям, и как следствие, техногенная цивилизация испытает существенные трудности в создании вненаучных ценностей.

Символическое ядро культуры традиционных цивилизаций при их зарождении было сформировано мифологическим мировоззрением, и несмотря на то, что последнее постепенно уходит

в прошлое, его влияние на структуры языка и культуры, а также на понимание ценностей как чего-то онтологически устойчивого, существующего как бы вне времени и человеческого сознания, до сих пор отчасти сохраняется, заставляя периодически выдвигать ценности той или иной эпохи или социальной группы на роль «всеобщих», «вечных» и «неизменных». Мифологическое мировоззрение не подразумевает прогресса и вообще относится к изменениям настороженно. Идеалом и образцом, к которому нужно стремиться, будет являться прошлое – тоже, естественно, мифологизированное. Идея прогресса подобному типу мировоззрения глубоко чужда.

Однако есть ряд аспектов, которые свойственны мифологическому мировоззрению и благодаря которым оно до сих пор сохраняет привлекательность, и отдельные его составляющие продолжают жить и воспроизводиться в культуре. В первую очередь, это все, что связано со смысло- и целеполаганием. Особенность любых мифов заключается не столько в том, чтобы дать возможность что-то объяснить на уровне причинно-следственных связей – с этим как раз намного лучше справляется наука. Но мифы способны дать объяснение с точки зрения придания происходящим в реальности событиям смысла, пояснить не столько как или почему они происходят, сколько дать ответ на вопрос – зачем. С точки зрения науки постановка вопроса о целях некорректна. Но в большинстве областей личностной и социальной активности человек мыслит именно так, потому что иное грозит нам экзистенциальным тупиком, выбраться из которого будет не проще, чем герою «Мифа о Сизифе».

Таким образом, когда речь идет о противопоставлении техногенной цивилизации традиционным, не стоит забывать о том, что элементы этих цивилизационных типов в реально существующих обществах оказываются соседствующими в едином временном отрезке.

Аутопойетические разрывы техногенной цивилизации и возможности их преодоления. Если рассматривать цивилизацию как социальную суперсистему, то она должна обладать как минимум двумя взаимно обусловленными свойствами – а именно, быть аутопойетической и рекурсивной. То есть должны постоянно

самовоспроизводиться ее структуры, с одной стороны, и каждая часть системы должна воспроизводить те же основные структуры, что и вся система в целом. Однако если техногенная цивилизация является неоднородной и может доминировать, то неизбежно включит в себя элементы других цивилизационных типов. В этом случае можно получить более одного варианта воспроизводства и самовоспроизводства структур. Именно так чаще всего и происходит, когда техногенная цивилизация возникает там, где на тот момент (всегда) существуют традиционные цивилизации.

Для техногенной цивилизации возможны как минимум два варианта взаимодействия с традиционными. В первом – можно наблюдать ситуацию, когда экспансивная техногенная цивилизация, распространяясь, вступает во взаимодействие с традиционной, где до этого науки как социального института могло и не сложиться, а технологии будут находиться на доиндустриальном уровне. В таком варианте техногенная цивилизация, по мнению Степина, ведет себя достаточно агрессивно, стремясь поглотить иные культуры, которые, чтобы избежать подобного, вынуждены приспосабливаться к этой «экспансии» как к воздействию внешней среды.

При этом традиционные культуры не отторгают техногенное влияние целиком, замыкаясь в себе, шанс сохраниться есть у тех их частей, которые, с одной стороны, не связаны с техногенной цивилизацией никакими связями по принципу структурной сопряженности (то есть не соприкасаются с ней, не «помогают» ей и не «препятствуют»), и с другой стороны, выполняют достаточно важные социальные функции, которые нельзя ни упразднить, ни заменить тем, что способна предложить техногенная цивилизация.

Во втором варианте возможного взаимодействия техногенной и традиционной цивилизаций развитие событий происходит тем хорошо известным образом, когда техногенная цивилизация появляется постепенно, можно сказать, «изнутри» традиционной, как это и произошло изначально в Европе в Новое время. При этом процесс взаимодействия двух цивилизационных типов на первый взгляд выглядит несколько менее драматичным, чем в первом варианте, так как занимает гораздо больше времени. Здесь при

длительном поступательном внедрении научного мировоззрения и технического прогресса они вполне органично входят в ткань культуры. Однако, поскольку научное мировоззрение, как было показано, не способно продуцировать все необходимые ценности, а только их небольшую часть, остальные ценности неизбежно должны были стать наследием традиционного этапа цивилизационного развития, в рамках которого постоянный прогресс (свойственный техногенному этапу) вообще не подразумевался.

В этом случае аутопойезис также как и в первом варианте взаимодействия, будет происходить не в одной, а в двух или нескольких системах, связанных через структурную сопряженность и влияющих друг на друга, одновременно с этим наука и техника будут выступать своего рода «локомотивом», тянущим за собой остальные системы. Рекурсия также будет проявляться в рамках каждой из взаимосвязанных друг с другом систем, поэтому периодически будут столкновения различных рекурсивных систем там, где предполагается внедрение особенностей одной системы в другую. Это столкновение становится очевидным, когда происходит массовое внедрение новых технологических достижений в социальные практики, изначально подчиненные преимущественно вненаучным ценностным системам. Именно это явление можно наблюдать при бурном развитии цифровых технологий.

Формирование ценностных основ применения цифровых технологий. Если говорить о создаваемых и уже созданных техногенной цивилизацией технологиях, в особенности, цифровых, то здесь в первую очередь необходимо понять, что именно будет являться ценностью: технология, способы ее применения или цели использования.

В рамках научного мировоззрения возможна ситуация, когда технология может стать ценностью и может выступать в роли объекта, связанного ценностным отношением с субъектом, а также конституировать целеполагание, становясь целью. Хотя такой подход органичен именно для техногенной цивилизации, которая неоднородна и в которой взаимодействуют и нередко конкурируют культурные пласты, принадлежащие разным цивилизационным типам. Для значительной части сегментов современной цивилизации технология – не цель, а средство, и соответственно,

оценивать надо цели, которые могут быть поставлены и достигнуты, и формы и границы использования технологии. Однако оценка целей и средств, внешних по отношению к технологиям как таковым, неизбежно должна осуществляться исходя из вне- и донаучных ценностей, о которых было сказано ранее. Здесь можно снова столкнуться с теми же ценностными, а затем и этическими проблемами, которые уже давно связаны не только с технологиями, но и с научными достижениями в целом.

Одна из таких трудноразрешимых проблем постнеклассической научной парадигмы (о чем предупреждал еще Степин) – каким образом можно включить в научную и технологическую проблематику социально-ценностной составляющей так, чтобы не произошло разрушение границ этих областей и не ограничивало бы их дальнейшее развитие? Последнее вполне вероятно, если считать технологии, а вслед за ними и научные разработки, только средством.

Попробуем предположить, что преодоление этого противоречия между ценностями научно-технологической сферы и активно развивающихся сейчас цифровых технологий, с одной стороны, и остальных социальных подсистем, с другой, будет возможно при построении системы более общего порядка, в которую, как своего рода метасистему, войдут все указанные сферы. В этом случае все их «частные» ценностные системы будут подчиняться ценностям более высокого порядка, что также позволит снять противоречия между сосуществующими в рамках одной цивилизации рекурсивными аутопойетическими культурными системами. Подробный разбор этих предполагаемых ценностных систем высшего порядка заслуживает отдельного исследования. В рамках настоящего исследования предполагается, что для таких ценностных систем фокус внимания должен сместиться с целей и средств на субъекта (как индивидуального, так и коллективного) как наиболее фундаментальную и наименее изменчивую из необходимых составляющих ценности.

Заключение

Суммируя, можно сказать, что аутопойезис техногенной цивилизации осуществляется параллельно в нескольких общественных

подсистемах, при том, что наука и технологии являются ведущими в данном цивилизационном типе.

В случае, когда разрыв между быстрым научно-техническим прогрессом и медленным развитием остальных подсистем становится критическим, в обществе из-за возросших несоответствий возникают кризисные состояния, в ходе которых все подсистемы экстренно «подтягиваются» к научно-технической системе, вырабатывая или видоизменяя ценностные комплексы для того, чтобы «приспособиться» к изменившимся условиям.

Смеем предположить, что только тогда можно говорить о грядущей посттехногенной цивилизации, когда все отдельно существующие подсистемы перестанут скачкообразно приспосабливаться одна к другой и сольются воедино на новом метауровне, образовав единую, целостную систему с единым ценностно-мировоззренческим базисом.

ЦИТИРУЕМАЯ ЛИТЕРАТУРА

Князева 2005 – *Князева Е.Н.* Творческий путь Франсиско Варелы: от теории автопоэзиса до новой концепции в когнитивной науке // Вопросы философии. 2005. № 8. С. 91–104.

Каган 1997 – *Каган М.С.* Философская теория ценностей. – СПб.: Петрополис, 1997.

Лепский 2016 – *Лепский В.Е.* Аналитика сборки субъектов развития. – М.: Когито-Центр, 2016.

Луман 2011 – *Луман Н.* Общество общества. – М: Логос, 2011.

Степин 1989 – *Степин В.С.* Научное познание и ценности техногенной цивилизации // Вопросы философии. 1989. № 10. С. 3–18.

Степин 2011 – *Степин В.С.* Цивилизация и культура. – СПб.: СПбГУП, 2011.

Степин, Горохов, Розов 1995 – *Степин В.С., Горохов В.Г., Розов М.А.* Философия науки и техники. – М.: Контакт-Альфа, 1995.

Шохин 2006 – *Шохин В.К.* Философия ценностей и ранняя аксиологическая мысль. – М.: Изд-во РУДН, 2006.

Chettiparamb 2020 – *Chettiparamb A.* Autopoietic Interaction Systems: Microdynamics of Participation and Its Limits // International Planning Studies. Vol. 25. No. 4. P. 427–440.

Clarke 2019 – *Clarke B.* Finding Cybernetics // World Futures. Vol. 75. No. 1–2. P. 17–28.

Maturana, Varela 1980 – *Maturana H., Varela F.* Autopoiesis: The Organization of the Living // *Maturana H., Varela F.* Autopoiesis and Cognition: The Realization of the Living. – Dordrecht: D. Reidel, 1980. P. 63–134.

Šubrt 2019 – *Šubrt J.* Niklas Luhmann's system theory: A critical analysis // RUDN Journal of Sociology 2019 Vol. 19. No. 4. P. 607–616.

REFERENCES

Chettiparamb A. (2020) Autopoietic Interaction Systems: Microdynamics of Participation and Its Limits. *International Planning Studies.* Vol. 25, no. 4, pp. 427–440.

Clarke B. (2019) Finding Cybernetics. *World Futures.* Vol. 75, no. 1–2, pp. 17–28.

Knyazeva E.N. (2005) The Creative Path of Francisco Varela: From the Theory of Autopoiesis to a New Concept in Cognitive Science. *Voprosy filosofii.* No. 8, pp. 91–104 (in Russian).

Kagan M.S. (1997) *Philosophical Theory of Values.* Saint Petersburg: Petropolis (in Russian).

Lepskiy V.E. (2016) *Analytics of the Assembly of Development Subjects.* Moscow: Kogito-Tsentr (in Russian).

Luhmann N. (1997) *Die Gesellschaft der Gesellschaft.* Frankfurt: Suhrkamp (Russian translation: Moscow: Logos, 2011).

Maturana H. & Varela F. (1980) Autopoiesis: The Organization of the Living. In: Maturana H. & Varela F. *Autopoiesis and Cognition: The Realization of the Living* (pp. 63–134). Dordrecht: D. Reidel.

Shokhin V.K. (2006) *Philosophy of Values and Early Axiological Thought.* Moscow: RUDN Publishing House (in Russian).

Stepin V.S. (1989) Scientific Knowledge and Values of Technogenic Civilization. *Voprosy filosofii.* No. 10, pp. 3–18 (in Russian).

Stepin V.S. (2011) *Civilization and Culture*. Saint Petersburg: SPbGUP (in Russian).

Stepin V.S., Gorokhov V.G., & Rozov M.A. (1995) *Philosophy of Science and Technology.* Moscow: Kontakt-Alpha (in Russian).

Šubrt J. (2019) Niklas Luhmann's System Theory: A Critical Analysis. *RUDN Journal of Sociology.* Vol. 19, no. 4, pp. 607–616.

DOI: 10.30727/0235-1188-2022-65-1-124-143
Оригинальная исследовательская статья
Original research article

Ценностные ориентации технологий искусственного интеллекта в США и Китае: философский анализ*

А.М. Савельев
Институт философии РАН, Москва, Россия,
Аналитический центр при Правительстве Российской Федерации, Москва, Россия

Д.А. Журенков
Институт философии РАН, Москва, Россия,
Всероссийский научно-исследовательский институт «Центр», Москва, Россия

А.Е. Пойкин
Всероссийский научно-исследовательский институт «Центр», Москва, Россия

Аннотация

Искусственный интеллект (ИИ) в XXI веке уже перестал восприниматься как исключительно технологическое явление, все больше и больше приобретая черты социального и гуманитарного феномена, развивающегося в сложном контексте культурных, ценностных, мировоззренческих и морально-этических сторон жизни человека. Влияние технологий, связанных с ИИ, на современное общество пока еще сложно оценить в полной мере, что не мешает исследователям, энтузиастам и политическим лидерам делать попытки определить ценностные рамки, которые обеспечат использование ИИ в интересах развития общества. С ростом интереса к ИИ все больше технологически развитых стран мира создают свои стратегии по развитию и использованию этого технологического чуда XXI века. Эти новаторские документы часто кажутся расплывчатыми и неопределенными, но тем не менее они позволяют оценить, как политические и научно-технологические элиты этих стран видят ценностные ориентации развития технологий ИИ как на национальном, так и на международном уровне. В статье с позиций постнеклассической научной рациональности проведен философский анализ ценностных ориентаций развития

* Работа поддержана Российским научным фондом (РНФ), грант № 21-18-00184 «Социогуманитарные основания критериев оценки инноваций, использующих цифровые технологии и искусственный интеллект».

технологий искусственного интеллекта в США и Китае – современных научно-технических лидеров в этой сфере – на основе стратегических документов, определяющих развитие и применение ИИ в этих странах. Авторы статьи делают вывод, что на современном постнеклассическом этапе развития науки ценностная компонента не просто является одной из интегральных компонент научно-технической деятельности, но может носить определяющий характер в определении целей и задач развития высоких технологий на государственном уровне.

Ключевые слова: философия искусственного интеллекта, этика, социальная философия, постнеклассическая научная рациональность, ценности, философия техники, научно-технический прогресс.

Савельев Антон Максимович – аспирант Института философии РАН, ведущий советник Аналитического центра при Правительстве Российской Федерации.
anton.saveliev@gmail.com
https://orcid.org/0000-0002-3687-9147

Журенков Денис Александрович – аспирант Института философии РИН, руководитель Центра диверсификации организаций ОПК ФГУП «ВНИИ «Центр»
dzhurenkoff@mail.ru
https://orcid.org/0000-0002-3968-5815

Пойкин Артем Евгеньевич – заместитель начальника отдела Центра диверсификации организаций ОПК ФГУП «ВНИИ «Центр»
art-tem1@mail.ru
https://orcid.org/0000-0001-6185-8922

Для цитирования: *Савельев А.М., Журенков Д.А., Пойкин А.Е.* Ценностные ориентации технологий искусственного интеллекта в США и Китае: философский анализ // Философские науки. 2022. Т. 65. № 1. С. 124–143. DOI: 10.30727/0235-1188-2022-65-1-124-143

Value Orientations of Artificial Intelligence Technologies in USA and China: A Philosophical Analysis

A.M. Saveliev
Institute of Philosophy, Russian Academy of Sciences, Moscow, Russia, Analytical Center under the Government of the Russian Federation, Moscow, Russia

* The work was supported by the Russian Science Foundation, grant no. 21-18-00184 “Social and humanitarian foundations for evaluation criteria for innovations based on digital technologies and artificial intelligence.”

D.A. Zhurenkov
Institute of Philosophy, Russian Academy of Sciences, Moscow, Russia,
All-Russian Scientific Research Institute "Center", Moscow, Russia

A.E. Poikin
All-Russian Scientific Research Institute "Center", Moscow, Russia

Abstract

Artificial Intelligence (AI) in the 21st century is no longer perceived as a purely technological phenomenon, more and more becoming a social and humanitarian phenomenon that develops in a complex context of cultural, value, philosophical, and ethical aspects of human life. The impact of AI-related technologies on contemporary society is still difficult to assess fully, which does not prevent enthusiastic researchers and political leaders from attempting to define a value framework that will ensure the use of AI for societal development. As interest in AI grows, more and more technologically advanced countries in the world are creating their own strategies for the development and use of this technological marvel of the 21st century. These pioneering documents often seem vague and indefinite, but nevertheless they allow us to assess how the political and scientific and technological elites of these countries see the value orientations of AI technology development, both nationally and internationally. The article presents a philosophical analysis of the value orientations of AI technology development in the USA and China – modern scientific and technological leaders in this field – on the basis of strategic documents defining the development and application of AI in these countries from the position of post-non-classical scientific rationality. The authors of the article conclude that, at the contemporary post-non-classical stage of science development, the value component is not only one of the integral components of scientific and technological activities but may be decisive in determining the goals and objectives of high-tech development at the state level.

Keywords: philosophy of artificial intelligence, ethics, social philosophy, post-non-classical scientific rationality, values, philosophy of technology, scientific and technological progress.

Anton M. Saveliev – postgraduate student, Institute of Philosophy, Russian Academy of Sciences; Leading Adviser, Analytical Center under the Government of the Russian Federation.
anton.saveliev@gmail.com
https://orcid.org/0000-0002-3687-9147

Denis A. Zhurenkov – postgraduate student, Institute of Philosophy, Russian Academy of Sciences; Head of the Center for Diversification of

Defense Industry Organizations, All-Russian Scientific Research Institute "Center."
dzhurenkoff@mail.ru
https://orcid.org/0000-0002-3968-5815

Artem E. Poikin – Deputy Head of Department, Center for Diversification of Defense Industry Organizations, All-Russian Scientific Research Institute "Center."
art-tem1@mail.ru
https://orcid.org/0000-0001-6185-8922

For citation: Savelyev A.M., Zhurenkov D.A., & Poikin A.E. (2022) Value Orientations of Artificial Intelligence Technologies in the USA and China: A Philosophical Analysis. *Russian Journal of Philosophical Sciences = Filosofskie nauki.* Vol. 65, no. 1, pp. 124–143.
DOI: 10.30727/0235-1188-2022-65-1-124-143

Введение

Современная наука не отказывает искусственному интеллекту (ИИ) в праве иметь философское и ценностное измерение. Более того, ИИ сам зачастую понимается как обоюдоострое оружие прогресса – почти что наравне с технологией расщепления атомного ядра, чей разрушительный потенциал во многом определил контуры политического устройства мира во второй половине XX века. Такие выдающиеся ученые и общественные интеллектуалы, как Стивен Хокинг и Мартин Рис, а также инноватор Илон Маск и исследователь ИИ Стюарт Рассел, весьма красноречиво говорили о разрушительной силе, которой обладает ИИ, упоминая прежде всего о риске полного уничтожения человечества, если технологии сильного ИИ станут неуправляемыми или попадут в неумелые руки. В 2009 году на конференции в Асиломаре ведущие исследователи ИИ [Horvitz, Selman 2009] выразили свою растущую обеспокоенность ценностной и морально-этической стороной развития ИИ, что в итоге привело к подписанию открытого письма[1] и созданию Асиломарских принципов ИИ[2] – набора из

[1] An Open Letter: Research Priorities for Robust and Beneficial Artificial Intelligence // Future of Life Institute. 2015. – URL: https://futureoflife.org/2015/10/27/ai-open-letter/

[2] Asilomar AI Principles // Future of Life Institute. 2017. – URL: https://futureoflife.org/2017/08/11/ai-principles/

23 руководящих принципов, описывающих ценностную, этическую и общественную проблематику развития ИИ и этические нормы для развития искусственного интеллекта во благо человечества. В свою очередь потенциальная опасность применения ИИ в качестве оружия открыто обсуждалась на ведущей конференции по ИИ – конференции Ассоциации по развитию искусственного интеллекта (*AAAI*) 2015 года – и на семинаре по ИИ и этике в рамках той же конференции. Международное сообщество также разделяет определенное недоверие к ИИ, но в целом согласно с тем, что разработать и формализовать какие-либо четкие ориентиры во избежание серьезных опасностей довольно трудно: технология ИИ слишком сложна [Floridi, Cowls 2022]. Тем не менее данный аспект не мешает технологически развитым странам создавать масштабные программы развития и технологий ИИ, чье социальное и гуманитарное обеспечение подкрепляется соответствующими документами стратегического характера. Зачастую данные стратегии носят крайне расплывчатый и общий характер, однако сам факт их существования, а также наличие в них ценностной и моральной составляющей, позволяет судить об ИИ как о сложном феномене, чьи границы уже давно распространились за рамки научно-технической парадигмы [Архипов, Наумов 2017а; Архипов, Наумов 2017б]. Авторы статьи предприняли попытку провести общий философский анализ стратегических документов, определяющих перспективы развития технологий ИИ в США и Китае на предмет выявления в них тех ключевых ценностных ориентаций, которые, с точки зрения руководства этих стран, должны определить развитие технологий искусственного интеллекта в ближайшем будущем. Выбор этих стран как объектов исследования отнюдь не случаен – американское и китайское руководство рассматривают ИИ не просто как изолированный набор передовых и необходимых технологических решений, но как инструмент комплексного, парадигмального преобразования общественных, экономических, управленческих и в конечном итоге мировоззренческих основ существующего миропорядка. Таким образом, философский анализ ИИ как преобразующего инструмента, провоцирующего парадигмальные сдвиги в обществе, выдвигает новые требования к методологии такого анализа, которая должна учитывать проблематику ИИ как феномена нового постнеклассического этапа развития науки, в условиях которого индивидуальный и коллективный субъект научного познания не

только решает сугубо деятельностные задачи научного поиска, но и осуществляет рефлексию над ценностными основаниями научной деятельности [Степин 2009, 249–295].

Искусственный интеллект на современном этапе его развития с уверенностью стоит рассматривать в контексте постнеклассической научной рациональности (постнеклассики), где он предстает в качестве сложной интегрированной системы [Лекторский 2016], включающей в себя явления и процессы технологического, инженерного, научного, социального, биологического, психологического и, как следствие, ценностного характера, где проблематика неизбежно смещается к парадигме «человек – машина» [Лекторский 2001]. Таким образом, ценностные ориентации развития технологий ИИ в контексте постнеклассики стоит рассматривать с позиций отражения этики, норм, ценностей и традиций коллективного субъекта инновационной деятельности (в контексте настоящего исследования – страны и нации) в подходах, как к самому научному исследованию, так и к предполагаемым его результатам, к внедрению технологий в социальную и культурную, экономическую и политическую ткань общества [Friedler, Scheidegger, Venkatasubramanian 2021].

Искусственный интеллект – этическое и ценностное измерение

Как уже отмечалось ранее – искусственный интеллект в условиях постнеклассики без сомнения является, в т.ч. и мировоззренческим феноменом, заслуживающим всестороннего анализа с позиций философии науки. В этой связи перед современным исследователем стоит задача поиска надлежащих исходных оснований для философско-методологического анализа сложного феномена ИИ. В настоящее время тема ценностного и этического осмысления ИИ как никогда актуальна, что заставляет представителей общественной, политической и научной элиты во всем мире разрабатывать соответствующие теоретические концепции, призванные, так или иначе, урегулировать ценностный и этический аспект разработки и применения технологий ИИ. Плюрализм подходов к этому вопросу, безусловно, обогащает научное знание, но не делает задачу философско-методологического анализа проще. Более того, многообразие ценностно-этических концепций существенно затрудняет выработку общепризнанных норм разработки и использования ИИ. К подобной мысли пришел и профессор философии и информационной этики Оксфордского

университета Л. Флориди (1964) – один из авторитетных исследователей в области современной философии техники, а именно – «философии информационных технологий» [Floridi 2002]. Флориди проанализировал шесть важнейших инициатив, направленных на совершенствование ценностно-этического обеспечения развития ИИ:

• «Асиломарские принципы ИИ», разработанные под эгидой Института будущего жизни в 2017 году;

• «Монреальская декларация ответственного ИИ», разработанная под эгидой Монреальского университета в 2017 году[3];

• Общие принципы, предложенные во втором издании книги «Этически обоснованный дизайн ИИ: взгляд на благополучие человека в автономных и интеллектуальных системах»[4]; этот документ является результатом совместных усилий 250 экспертов под патронажем Института инженеров по электротехнике и электронике (*IEEE*);

• Этические принципы, предложенные в докладе об искусственном интеллекте, робототехнике и «автономных» системах, опубликованном Европейской группой по этике в науке и новых технологиях Европейской комиссии в марте 2018 года;

• «Пять всеобъемлющих принципов для кодекса ИИ», предложенные в докладе Комитета по искусственному интеллекту Палаты лордов Великобритании в 2018 году;

• «Принципы Партнерства в ИИ» многосторонней организации *Partnership on AI*, состоящей из ученых, исследователей, общественных объединений, создающих и использующих технологии ИИ.

Л. Флориди совместно со своим коллегой Д. Коулзом проанализировали все эти документы и обнаружили, что в совокупности они содержат 47 основных принципов того, как ИИ может быть использован с пользой для общества. Согласно концепции, принятой вышеупомянутыми исследователями, все эти принципы имеют высокую степень когерентности с четырьмя основными принципами, обычно используемыми в биоэтике: *«делай благо»* (*beneficence*), *«не навреди»* (*non-maleficence*), *«уважение автономии субъекта»* (*autonomy*) и *«справедливость»* (*justice*) [Floridi, Cowls 2022; Beauchamp, Saghai 2012]. По мнению Фло-

[3] Montreal Declaration for a Responsible Development of Artificial Intelligence. – URL: https://www.montrealdeclaration-responsibleai.com/the-declaration

[4] Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems. Version 2 // IEEE. – URL: https://standards.ieee.org/wp-content/uploads/import/documents/other/ead_v2.pdf

риди, биоэтика больше всего похожа на информационную этику в вопросах средового взаимодействия субъектов в условиях новой сложной экосистемы современного цифрового общества [Floridi 2008; Floridi 2013; Floridi 2019].

Однако Флориди и Каулз утверждают, что дополнительно необходим новый принцип: принцип «*обоснованности*», который понимается одновременно как «понятность/объяснимость» ИИ для неспециалистов и его подотчетность пользователю [Floridi, Cowls 2022]. Приняв данный исходный посыл, основанный на работах Л. Флориди и Д. Коулза, можно сконструировать общую матрицу ценностных оснований технологий искусственного интеллекта и их онтологических особенностей (см. табл. 1).

Принцип ценностного основания	**Онтологические особенности**
«делай благо»	Содействие благополучию индивидуальных и коллективных субъектов, сохранение достоинства и поддержание гармонии в масштабе всего мира.
«не навреди»	Неприкосновенность жизни, безопасность и «осторожность в возможностях».
«уважение автономии субъекта»	Право субъекта принимать решение, право полностью или частично делегировать принятие решения иным субъектам.
«справедливость»	Содействие процветанию, солидарности, предотвращение несправедливости.
«обоснованность»	Обеспечение работы других принципов через понятность и подотчетность.

Табл. 1. Общая матрица ценностных оснований технологий искусственного интеллекта

1. ***«Делай благо»***: содействие благополучию, сохранение достоинства и поддержание гармонии в масштабе всего мира. Этот принцип подчеркивает основополагающее значение содействия благополучию людей в планетарном масштабе с помощью технологий ИИ, дальнейшего процветания человечества и сохранения социально ответственной окружающей среды для будущих поколений.

2. ***«Не навреди»***: неприкосновенность жизни, безопасность и «осторожность в возможностях» [Floridi, Cowls 2019]. Этот принцип предостерегает от различных негативных последствий

чрезмерного или неправильного использования технологий ИИ, особенно когда речь идет о неприкосновенности личной жизни и военном применении ИИ. При этом пока не ясно, как себя проявят угрозы, связанные с ИИ в будущем: будет ли речь идти преимущественно о неправомерном использовании ИИ самими людьми, или же опасность будет исходить от самих технологий как таковых.

3. ***«Уважение автономии субъекта»***: принимать решения и делегировать принятие решений. «Когда мы используем ИИ и его "умные" возможности, мы добровольно уступаем часть своих полномочий по принятию решений технологическим артефактам – искусственным агентам», – утверждает Флориди. Этот важный принцип говорит о балансе между правом принятия решений, которое люди сохраняют за собой, и правом, которое они делегируют искусственным цифровым агентам. Люди должны сохранять за собой право решать, как поступать: пользоваться свободой выбора там, где это необходимо, и уступать ее в тех случаях, когда на то есть веские причины, утверждает Флориди.

4. ***«Справедливость»***: содействие процветанию, солидарности, предотвращение несправедливости. Хотя справедливость может показаться довольно широким понятием, все энтузиасты ИИ и мыслители согласны с тем, что устранение несправедливой дискриминации, равно как и необходимость всеобщего процветания, должны стоять в центре использования технологий ИИ.

5. ***«Объяснимость/обоснованность»***: обеспечение работы других принципов через понятность и подотчетность. Проще говоря, этот принцип должен ответить на вопрос: является ли человечество «пациентом», получающим «лечение» в виде «горькой пилюли искусственного интеллекта», «врачом», назначающим его, или возможно, и тем, и другим. Таким образом, принцип объяснимости должен включать в себя понимание того, как работает ИИ, равно как и понимание меры ответственности тех, кто работает с ИИ [Floridi, Cowls 2022].

Данная матрица ценностных оснований технологий искусственного интеллекта, безусловно, не является бесспорной – она имеет довольно общий и размытый характер. Однако эти недостатки оборачиваются достоинствами в рамках наших исследовательских задач: стратегические документы, посвященные применению технологий ИИ в государственном масштабе, сами по себе носят

расплывчатый и общий характер. Таким образом, данная матрица может быть временно принята в качестве потенциально приемлемого инструмента оценки того, как ценностные и этические ориентации подразумеваются и реализуются в стратегиях развития ИИ. Чтобы немного облегчить эту задачу, авторы хотели бы предложить следующую визуализацию выбранного инструмента оценки (см. табл. 2).

Оценка	**Шкала основных принципов ценностной ориентации**	**Шкала принципа объяснимости**
0	принцип не упомянут вообще	не упомянуты поддерживающие механизмы и отсутствует дальнейшее представление
1	принцип упомянут хотя бы один раз в документе	поддерживающие механизмы упомянуты, но носят чисто декларативный характер
2	принцип раскрыт, по крайней мере, в одной из стратегических целей документа	поддерживающие механизмы подробно изложены и подкреплены юридическими инициативами
3	принцип представлен в многочисленных стратегических целях документа	поддерживающие механизмы четко прописаны, подкреплены законодательными инициативами и имеют соответствующие программы, поддерживаемые государством

Табл. 2. Шкала ценностных ориентаций ИИ в анализируемых документах

В рамках данной статьи оценка представленности тех или иных ценностных оснований в анализируемых документах будет проводиться по двум шкалам – шкале основных принципов ценностной ориентации (*«делай благо», «не навреди», «автономия», «справедливость»*) и шкале принципа *«обоснованности»*. Окончательный вердикт будет вынесен на основе простого семантического анализа, в котором выбранные выше единицы оценки могут встретить или не встретить свое семантическое представление в анализируемом тексте. При этом авторы данной статьи не ставят перед собой задачу определить точную частоту таких встреч. Вместо этого мы попытаются исследовать базовую семантическую корреляцию между принципами, упомянутыми в матрице ценностных оснований, и положениями анализируемых документов.

Данный методологический подход может показаться слишком расплывчатым, но авторы опасаются, что любой другой более специализированный подход может оказаться слишком узким для документов, которые кажутся расплывчатыми как по букве, так и по духу.

Ценностные ориентации технологий ИИ в США

Отличительной особенностью США является отсутствие как системного правового регулирования в сфере развития ИИ, так и общенациональной государственной стратегии в этой сфере. Ближе всего к национальной стратегии в области искусственного интеллекта Соединенных Штатов по своей сути соответствует «Исполнительный приказ по ИИ», подписанный президентом Д. Трампом 11 февраля 2019 года[5]. В данном документе провозглашена т.н. «американская инициатива по ИИ» – рамочная национальная стратегия Соединенных Штатов в области искусственного интеллекта. Эта стратегия предусматривает согласованные усилия по продвижению и защите американских технологий и инноваций в области ИИ в рамках пяти приоритетов: 1) устойчивое инвестирование в НИОКР в области ИИ, 2) использование средств федерального правительства для развития ИИ, 3) устранение барьеров на пути инноваций в области ИИ, 4) расширение возможностей американских работников с помощью образования, ориентированного на ИИ, и возможностей обучения, и 5) содействие созданию международной среды, поддерживающей американские инновации в области ИИ и их ответственное использование.

«Исполнительный приказ по ИИ» во многом дополняется приоритетами Национального стратегического плана НИОКР в области ИИ (*National Artificial Intelligence Research and Development Strategic Plan*) (далее – Стратегический план):

• осуществлять долгосрочные инвестиции в исследования ИИ, в приоритетном порядке инвестировать в следующее поколение ИИ, которые будут способствовать открытиям и проницательности и позволят Соединенным Штатам оставаться мировым лидером в области ИИ;

---

[5] Maintaining American Leadership in Artificial Intelligence // Executive Office of the President. E.O. 13859. February 11, 2019. – URL: https://www.federalregister.gov/documents/2019/02/14/2019-02544/maintaining-american-leadership-in-artificial-intelligence

• разработать эффективные методы взаимодействия человека и ИИ, создать системы ИИ, которые эффективно дополняют и расширяют человеческие возможности;

• учитывать этические, правовые и общественные последствия применения ИИ; разработать системы ИИ, которые способны решать этические, правовые и общественные проблемы с помощью технологических инструментов;

• обеспечить безопасность и надежность систем ИИ – они должны быть надежными, безотказными, безопасными и заслуживающими доверия;

• создать общедоступные массивы данных и среды для обучения и тестирования ИИ; разработать и обеспечить доступ к высококачественным базам данных и средам обучения ИИ, а также к ресурсам для тестирования технологий ИИ и подготовки профильных специалистов;

• разработать широкий спектр методов оценки ИИ, включая технические стандарты и контрольные показатели;

• учитывать и удовлетворять государственные и общественные потребности в кадрах для НИОКР в области ИИ, повысить эффективность подготовки кадров в сфере НИОКР в области ИИ;

• повысить эффективность государственно-частных партнерств для ускорения прогресса в области ИИ, продвигать возможности для постоянных инвестиций в НИОКР в области ИИ и перехода достижений в практические возможности в сотрудничестве с научными кругами, промышленностью, международными партнерами и другими нефедеральными структурами.

Рассмотрим то, как ценностные основания ИИ трактуются в данном документе.

1. ***«Делай благо»***, – пожалуй, является наиболее ярко выраженным принципом, поскольку все восемь стратегических приоритетов упомянуты как приносящие потенциальную пользу «почти всем аспектам общества, включая экономику, здравоохранение, безопасность, право, транспорт и даже саму технологию». Преимущества для общества от всестороннего внедрения ИИ – ускорение восстановления нормальной жизни людей после стихийных бедствий, улучшение медицинской диагностики (например, в выявлении особо опасных видов рака), появление новых рабочих мест на рынке и т.д. По базовой шкале этот принцип представлен на оценку **«3»**. Тем не менее Стратегический план не предусматривает правовых инициатив и планов, которые могли бы оказать

поддержку реализации принципа благодеяния. Таким образом, по шкале обоснованности можно выставить только оценку **«1,5».**

2. ***«Не навреди»*** – косвенно упоминается в одной из целей Стратегического плана: «осознать и подробно рассмотреть те этические, правовые и общественные последствия, которые влечет за собой внедрение ИИ». Американские стратегические документы применительно к этой цели призывают к этичному и безопасному внедрению ИИ для предотвращения любого возможного вреда людям. Кроме того, «не навреди» косвенно подразумевается в другой цели этого документа: «Обеспечить безопасность и надежность систем ИИ». Утверждается, что ИИ должен быть «безопасным по [изначальному] замыслу», где безопасность обеспечивается на протяжении всего жизненного цикла технологии ИИ. По базовой шкале мы можем присвоить оценку **«2,5»**, но по шкале обоснованности возможна только оценка **«1»**, поскольку все средства, поддерживающие данный ценный ценностный принцип, носят исключительно декларативный характер.

3. ***«Уважение автономии субъектов»*** – представлено в цели «разработать эффективные методы взаимодействия человека и ИИ», где говорится, что развертывание систем ИИ должно рассматриваться «как один из вариантов исходного дизайна ИИ, который позволяет его операторам решать, стоит ли им вообще прибегать к запуску системы с ИИ или нет». По основной шкале можно смело присвоить оценку **«1»**, но полное отсутствие вспомогательных средств и дальнейшего юридического представления позволяет выставить только оценку **«0»** по шкале обоснованности.

4. ***«Справедливость»*** – хорошо представлена как в значении справедливого и честного использования ИИ (о чем достаточно пространно говорится в Стратегическом плане с позиции, преимущественно, общественных и государственных субъектов), так и в контексте «социальной справедливости» (цель – «лучше понять национальные потребности в кадрах для НИОКР в области ИИ»). В последнем случае речь идет о необходимости поддержки американских исследователей в области ИИ, а также студентов старших курсов. По основной шкале нами поставлена оценка **«3»**, но по шкале обоснованности – только **«1,5»**: в документе заявлены многочисленные средства поддержки, хотя и без какого-либо юридического и финансового обеспечения.

Ценностные ориентации технологий ИИ в Китае

Китай является одним из самых активных игроков в сфере ИИ, что находит свое отражение не только в технологиях и рыночной

инфраструктуре, но и в сфере законодательного регулирования. В настоящее время Китай обладает самой разветвленной системой законодательных актов и государственных планов, декларирующих приоритеты развития ИИ-технологий, но это многообразие компенсируется достаточно размытым характером этих документов [Савельев, Журенков 2019].

Искусственный интеллект в приоритетах китайского руководства на ближайшие годы (до 2025 года) рассматривается не только в контексте повышения национальной конкурентоспособности и технологической независимости на стратегическом уровне, но и как «средство совершенствования социального управления и развития». Так, «План развития искусственного интеллекта нового поколения», принятый Государственным советом КНР в 2017 году, – основополагающий стратегический документ Китая в области ИИ – утверждает, что искусственный интеллект «способен своевременно распознавать групповые когнитивные и психологические изменения, а также поможет [ответственным лицам] проявлять инициативу в принятии общественно важных решений»[6].

В плане указаны шесть ключевых задач, которые необходимо решить для достижения вышеупомянутых целей, в частности: 1) создание открытой и дружественной инновационной системы технологий ИИ; 2) развитие высокотехнологичной и высокоэффективной интеллектуальной экономики; 3) создание безопасного и благоприятного интеллектуального общества; 4) усиление военно-гражданской интеграции в области ИИ; 5) создание всеобъемлющей, безопасной и эффективной интеллектуальной инфраструктуры; и 6) перспективное планирование нового поколения крупных проектов, связанных с ИИ.

Несмотря на свою детальность и последовательность, китайский стратегический план является, пожалуй, самым трудным документом для анализа на предмет ценностных ориентаций. Он носит декларативный характер, что сильно его роднит с аналогичными стратегическими документами США, однако ориентация на преимущественно экономические и технологические приоритеты ставит в один ряд с бизнес-планами, а не с государственными стратегическими документами.

---

[6] A Next Generation Artificial Intelligence Development Plan // State Council of the People's Republic of China . 2017. – URL: https://d1y8sb8igg2f8e.cloudfront.net/documents/translation-fulltext-8.1.17.pdf

1. ***«Делай благо»*** – принцип, который редко упоминается, но явно подразумевается в китайском стратегическом плане. Практически, все приоритеты и задачи стратегического плана призваны принести стране экономическое процветание КНР на годы вперед, а также установить новое качество жизни для китайских граждан. Это включает в себя применение инновационных технологий ИИ в образовании, здравоохранении, пенсионном обеспечении и других первостепенных нуждах общества. В документе напрямую говорится о том, что человека нужно ставить «на первое место», «следовать общечеловеческим ценностям», «уважать права человека», соблюдать «национальную и региональную этику». В плане есть несколько ключевых приоритетных областей всестороннего внедрения ИИ во благо китайского общества, а именно: интеллектуальное образование, интеллектуальное медицинское обслуживание, интеллектуальные системы здравоохранения и ухода за престарелыми. Тем не менее план не предусматривает никаких средств для реализации этих благородных целей и приоритетов. Нет ни программ, ни проектов, ни законодательных инициатив для их поддержки. По основной шкале можно поставить оценку **«3»**. По шкале обоснованности – только **«1»**.

2. ***«Не навреди»*** – этот ценностный принцип является одним из главных приоритетов стратегического плана: «Разработать законы, правила и этические нормы, способствующие развитию ИИ, что требует безопасного и этичного использования ИИ». С этой целью документ выступает за разработку правовых основ ИИ, а также за исследования в области науки о поведенческих паттернах ИИ. Тем не менее стратегический план никак не раскрывает смысловое содержание этих положений и не предлагает никаких мер по их реализации. По базовой шкале мы можем поставить оценку **«1»**, но по шкале обоснованности возможна только оценка **«0»**.

3. Принцип ***«уважения автономии субъекта»*** – вообще не представлен. В рассматриваемой стратегии не упоминаются какие-либо модели принятия решений и делегирования полномочий для пользователей ИИ. Утверждается, что ИИ может существенно упростить процесс принятия решений и повысить его эффективность, но из документа не ясно, каким образом предполагается этого достичь. Мы присваиваем **«0»** по обеим шкалам.

4. ***«Справедливость»*** представлена в основном в понимании ответственного и законного использования технологий ИИ, о чем гласит один из приоритетов стратегического плана – «разработка

законов, правил и этических норм, способствующих развитию ИИ», – предполагающий создание этической и моральной системы рамочного взаимодействия людей и машин. Отметим, что в плане прописываются и принципы законности при обработке личной информации, защита частной жизни и безопасности личных данных. Таким образом, нельзя делать вывод о том, что частное и индивидуальное в КНР считается менее важным и ценным, чем общественное благо – в официальных документах данной проблематике уделяется достаточно много внимания, вопреки распространённым стереотипам. Есть и национальная специфика – призывается строить «сообщество единой судьбы», «поощрять социальную справедливость» и в разделе «ответственности» отметим «создание механизма подотчетности ИИ». В то же время теоретические контуры этической и моральной системы рамочного взаимодействия людей и машин описаны приблизительно, а способы ее создания и укоренения в обществе вообще не названы. Мы присваиваем оценку **«1»** по основной шкале и оценку **«0»** по шкале обоснованности.

Заключение

Рассмотренные нами стратегические документы США и Китая в области технологий ИИ позволяют заключить, что его ценностное измерение сейчас находится в центре внимания технологически развитых государств, стремящихся упрочить свое положение в мире при помощи новых цифровых технологий. ИИ в Китае и США рассматривается в первую очередь как инструмент ускорения (или даже провокации) благоприятных социальных перемен.

Стоит отметить, что ценностные ориентации обеих анализируемых стран не демонстрируют существенного антагонизма практически по всем вопросам. Разница в выставленных нами оценках вызвана по большей части крайне размытым представлением тех или иных ценностных ориентиров в анализируемых документах.

	«Не навреди»	**«Делай благо»**	**«Не навреди»**	**«Принцип уважения автономии субъекта»**	**«Принцип справедливости»**
США	**3**	**2,5**	**1**	**3**
КНР	**3**	**1**	**0**	**1**

Табл. 3. Сравнительная матрица ценностных ориентаций в технологиях ИИ в США и Китае

Существует мнение, что основное отличие подходов Китая и Запада (США в настоящем исследовании) лежат в различии ценностей – коллективных и индивидуальных. Это закладывает различные этические подходы к разработке ИИ, разное понимание справедливости, безопасности, конфиденциальности. При этом, американские стратегические документы в своих целях в области ИИ зачастую оперируют именно коллективными ценностными ориентациями, такими как «общественное благо», «социальная справедливость», «ответственность перед обществом», уделяя особое внимание защите интересов коллективных субъектов – профессиональных, научных и культурных сообществ, а также социально незащищенных слоев. Не так однозначна и ориентация КНР на «коллективные ценности» – в рамках стратегических инициатив китайского правительства создаются меры регулирования алгоритмов (например, рекламы и продаж), которые обеспечивают безопасность личных данных пользователей, прозрачность использования данных в Интернете и т.д. Что действительно отличает ценностные подходы к технологиям ИИ в Китае, так это включение идеологических концептов в этику разработки и внедрения ИИ. Руководящая идеология прописывается в законодательстве, в алгоритмах как экологическая норма прописано «продвижение социалистических ценностей»[7]. Вторым важным отличием китайского подхода является включение национальной культурной парадигмы в ценностные ориентации технологического развития (что полностью игнорируется в американских стратегических документах). Например, инженеры в области ИИ и высоких технологий в Китае должны руководствоваться такими ценностями как «процветание, демократия, вежливость и гармония», «свобода, равенство, справедливость и верховенство закона», «патриотизм, преданность, честность, дружба». Особо выделяются принципы ответственности, «предшествующей свободе», обязательств, «предшествующих правам», коллективное, «предшествующее индивидуальному», «гармония, предшествующая конфликту»[8].

---

[7] См.: Руководящие заключения по укреплению комплексного управления алгоритмами предоставления информационных услуг в Интернете // Министерство промышленности и информатизации КНР. 2021. – URL: https://wap.miit.gov.cn/xwdt/gxdt/art/2021/art_a8af2b48620b4905b365fc73cd81a1ec.html

[8] См.: Национальный учебник по инженерной этике для выпускников ВУЗов КНР (2019). – URL: http://www.tup.tsinghua.edu.cn/booksCenter/book_06831902.html

Что действительно объединяет оба подхода к ценностным ориентациям ИИ в США и Китае, так это излишне декларативный характер, как самих ценностных ориентаций, так и тех методов, которыми они должны быть воплощены в жизнь. Американские и китайские эксперты в области ИИ в полной мере осознают угрозы, которые несет ИИ, равно как и те потенциальные блага, которые он может дать обществу. Однако у них отсутствует понимание тех конкретных мер и шагов, при помощи которых можно защитить человечество от неправомерного использования ИИ, а зачастую и понимание того, что считать неправомерным использованием данных передовых технологий. Этот печальный факт лишь подчеркивает важность принципа «обоснованности», введенного Л. Флориди. Любые, даже самые замечательные и ясно сформулированные, ценностные ориентации становятся лишь благими пожеланиями без конкретных механизмов их внедрения и обеспечения. В условиях нарастающей конвергенции биологических и цифровых субъектов, вызванной практически повсеместным внедрением технологий ИИ, прежние подходы утрачивают свое значение, ведь ИИ уже давно не ограничен стенами научно-исследовательских институтов и вычислительных центров. Представитель современного общества все больше и больше приобретает черты инфорга – гибридного субъекта, существующего и развивающегося одновременно в реальном и цифровом мире. В этих принципиально новых условиях уже мало создать некий набор тех или иных ценностных и морально-этических правил для пользования технологией во благо [Umpleby, Medvedeva, Lepskiy 2019]. Напротив, речь должна идти о новой области знаний – социогуманитарной эргономике технологий ИИ – дисциплине, которая бы учитывала биологическую, психологическую социальную, когнитивную и духовную природу субъекта, частью жизнедеятельности которого являются технологии ИИ.

ЦИТИРУЕМАЯ ЛИТЕРАТУРА

Архипов, Наумов 2017а – *Архипов В.В., Наумов В.Б.* Искусственный интеллект и автономные устройства в контексте права: о разработке первого в России закона о робототехнике // Труды СПИИРАН. 2017. № 6 (55). С. 46–62.

Архипов, Наумов 2017б – *Архипов В.В., Наумов В.Б.* О некоторых вопросах теоретических оснований развития законодательства о ро-

бототехнике: аспекты воли и правосубъектности // Закон. 2017. № 5. С. 157–170.

Лекторский 2001 – *Лекторский В.А.* Эпистемология классическая и неклассическая. – М.: Эдиториал УРСС, 2001.

Лепский 2016 – *Лепский В.Е.* Технологии управления в информационных войнах (от классики к постнеклассике) – М.: Когито-Центр, 2016.

Савельев, Журенков 2019 – *Савельев А.М., Журенков Д.А.* Национальные стратегии развития систем искусственного интеллекта: опыт стран-лидеров и ситуация в России // Научный вестник оборонно-промышленного комплекса России. 2019. № 3. С. 75–82.

Степин 2003 – *Степин В.С.* Теоретическое знание. – М.: Прогресс-Традиция, 2003.

Beauchamp, Saghai 2012 – *Beauchamp T.L., Saghai Y.* The Historical Foundations of the Research-Practice Distinction in Bioethics // Theoretical Medicine and Bioethics. 2012. Vol. 33. No. 1. P. 45–56.

Floridi 2002 – *Floridi L.* What is the Philosophy of Information? // Metaphilosophy. Vol. 33. No. 1–2. P. 123–145.

Floridi 2008 – *Floridi L.* Foundations of Information Ethics // The Handbook of Information and Computer Ethics / ed by K.E. Himma, H.T. Tavani. – Hoboken: Wiley, 2008. P. 3–24.

Floridi 2013 – *Floridi L.* The Ethics of Information. – Oxford: Oxford University Press, 2013.

Floridi 2019 – *Floridi L.* What the Near Future of Artificial Intelligence Could Be // Philosophy & Technology. Vol. 32. No. 1. P. 1–16.

Floridi, Cowls 2022 – *Floridi L., Cowls J.* A Unified Framework of Five Principles for AI in Society // Machine Learning and the City: Applications in Architecture and Urban Design. – Hoboken: Wiley, 2022. P. 535–545.

Friedler, Scheidegger, Venkatasubramanian 2021 – *Friedler S.A., Scheidegger C., Venkatasubramanian S.* The (Im)Possibility of Fairness: Different Value Systems Require Different Mechanisms for Fair Decision Making // Communications of the ACM. 2021. Vol. 64. No. 4. P. 136–143.

Umpleby, Medvedeva, Lepskiy 2019 – *Umpleby S.A., Medvedeva T.A., Lepskiy V.* Recent Developments in Cybernetics, from Cognition to Social Systems // Cybernetics and Systems. 2019. Vol. 50. No. 4. P. 367–382.

REFERENCES

Arkhipov V.V. & Naumov V.B. (2017a) Artificial Intelligence and Autonomous Devices in Legal. *SPIIRAS Proceedings.* No. 6, pp. 46–62 (in Russian).

Arkhipov V.V. & Naumov V.B. (2017b) On Some Issues of the Theoretical Basis for the Development. *Zakon.* No. 5, pp. 157–170 (in Russian).

Beauchamp T.L. & Saghai Y. (2012) The Historical Foundations of the Research-Practice Distinction in Bioethics. *Theoretical Medicine and Bioethics.* Vol. 33, no. 1, pp. 45–56.

Floridi L. (2002) What is the Philosophy of Information? *Metaphilosophy.* Vol. 33, no. 1–2, pp. 123–145.

Floridi L. (2008) Foundations of Information Ethics. In: Himma K.E. & Tavani H.T. (Eds.) *The Handbook of Information and Computer Ethics* (pp. 3–24). Hoboken: Wiley.

Floridi L. (2013) *The Ethics of Information.* Oxford: Oxford University Press.

Floridi L. (2019) What the Near Future of Artificial Intelligence Could Be. *Philosophy & Technology.* Vol. 32, no. 1, pp. 1–16.

Floridi L. & Cowls J. (2022) A Unified Framework of Five Principles for AI in Society. In: Carta S. (Ed.) *Machine Learning and the City: Applications in Architecture and Urban Design* (pp. 535–545). Hoboken: Wiley.

Friedler S.A., Scheidegger C., & Venkatasubramanian S. (2021) The (Im) Possibility of Fairness: Different Value Systems Require Different Mechanisms for Fair Decision Making. *Communications of the ACM.* Vol. 64, no. 4, pp. 136–143.

Lektorsky V.A. (2001) *Classical and Nonclassical Epistemology.* Moscow: Editorial URSS (in Russian).

Lepskiy V.E. (2016) *The Technology of Management and Information Wars (From Classical to Post-Non-Classicalneklassike).* Moscow: Kogito-Tsentr (in Russian).

Savelyev A.M. & Zhurenkov D.A. (2019) National Strategies for the Development of Artificial Intelligence Systems: The Experience of the Leading Countries and the Situation in Russia. *Scientific Bulletin of the Military Industrial Complex of Russia.* No. 3, pp. 75–82 (in Russian).

Stepin V.S. (2003) *Theoretical Knowledge*. Moscow: Progress-Traditsiya (in Russian).

Umpleby S.A., Medvedeva T.A., & Lepskiy V.E. (2019) Recent Developments in Cybernetics, from Cognition to Social Systems. *Cybernetics and Systems.* Vol. 50, no. 4, pp. 367–382.

DOI: 10.30727/0235-1188-2022-65-1-144-159
Оригинальная исследовательская статья
Original research article

Социогуманитарные критерии инновационности цифровых технологий: анализ международного опыта стандартизации*

Б.Б. Славин
Финансовый университет при Правительстве РФ, Москва, Россия

Аннотация

Настоящее исследование посвящено вопросам социогуманитарных оснований измерения и оценки инноваций в области цифровых технологий на основе анализа международного опыта подготовки стандартов, в частности на основе «Руководства Осло», выпускаемого Организацией экономического сотрудничества и развития и являющегося фундаментом для стандартов *ISO*. Важным этапом в понимании инноваций с точки зрения цифровизации является признание инновационным не только продукта, но и процесса, который приводит к повышению эффективности. Бóльшая часть задач в области цифровых технологий как раз связана с повышением эффективности процессов, и, следовательно, цифровизация становится важным критерием инновационности организаций. Одним из назначений цифровых технологий является интеграция различных видов деятельности, в силу чего интегративное развитие (в том числе и с использованием цифровых платформ) тоже является критерием инновационности. Цифровой формат уникален тем, что позволяет получать новые знания благодаря анализу данных и процессов, а следовательно, аналитическая составляющая, включая технологии искусственного интеллекта, также является показателем инновационности. Таким образом, к критериям инновационности цифровых технологий следует отнести их использование для повышения эффективности деятельности организаций, для внедрения коммуникативных цифровых платформ и для создания разнообразных инструментов анализа данных, включая искусственный интеллект.

Ключевые слова: философия техники, цифровизация, измерение инноваций, гуманитарная экспертиза, критерии инновационности, интегративное развитие, социальная ценность.

* Работа поддержана Российским научным фондом (РНФ), грант № 21-18-00184 «Социогуманитарные основания критериев оценки инноваций, использующих цифровые технологии и искусственный интеллект».

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

bbslavin@gmail.com

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

Для цитирования: *Славин Б.Б.* Социогуманитарные основания критериев оценки инноваций, использующих цифровые технологий: анализ международного опыта стандартизации // Философские науки. 2022. Т. 65. № 1. С. 144–159. DOI: 10.30727/0235-1188-2022-65-1-144-159

Social and Humanitarian Grounds of Criteria for Assessment of Digital Technology Innovations: An Analysis of International Standardization Experience*

B.B. Slavin

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

Abstract

This research discusses the social and humanitarian grounds of measurement and assesment of innovations in the field of digital technologies based on the analysis of the international experience of the standards, in particular on the basis of the Oslo Manual issued by the Organization for Economic Cooperation and Development, which is the foundation for ISO standards. From the point of view of digitalization, an important stage in understanding innovation is the recognition of innovation not only as a product, but also a process that leads to increased efficiency. Most of the tasks in the field of digital technologies are precisely related to increasing the efficiency of processes, and, consequently, digitalization is becoming an important criterion for the innovativeness of organizations. One of the purposes of digital technologies is the integration of various types of activities, therefore integrative development (including using digital platforms) is also a criterion of innovation. The digital format is unique since it allows to gain new knowledge from data analysis and processes, and therefore the analytical component, including artificial intelligence technologies, is also an indicator of innovation. The article concludes that the criteria for the innovativeness of digital technologies should include their use to improve the efficiency of organizations, to adopt communicative digital platforms, and to create a variety of data analysis tools, including artificial intelligence.

* The work was supported by the Russian Science Foundation, grant no. 21-18-00184 "Social and humanitarian foundations for evaluation criteria for innovations based on digital technologies and artificial intelligence."

Keywords: philosophy of technology, digitalization, assessment of innovations, humanitarian expertise, innovativeness criteria, integrative development, social value.

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. (2022) Social and Humanitarian Grounds of Criteria for Assessment of Digital Technology Innovations: An Analysis of International Standardization Experience. *Russian Journal of Philosophical Sciences = Filosofskie nauki*. Vol. 65, no. 1, pp. 144–159.
DOI: 10.30727/0235-1188-2022-65-1-144-159

Введение

Стандарты отражают многократно повторенный и признанный лучшим опыт в той или иной области. Концепция стандартов говорит об их социогуманитарной природе, и именно так необходимо подходить к исследованию роли стандартов в деятельности организаций. Обычно процесс стандартизации ассоциируется с чем-то уже устоявшимся и не являющимся инновационным. И это действительно так, поскольку стандартизировать можно только то, что можно стабильно воспроизводить с гарантированным результатом. Однако если сами по себе инновации имеют дело с чем-то неизведанным и неповторимым, то инновационный процесс вполне поддается стандартизации. Более того, чем более инновационной становится экономика, тем нужнее стандарты. Любые стандарты базируются на измерении, а следовательно, стандарты в области инноваций требуют инструментов для измерений.

Сегодня драйвером инноваций являются информационные технологии. В них сочетаются как свойства производственных технологий, товаров, так и свойства услуг.

Именно понимание особой роли информационных технологий привело к тому, что в конце прошлого века особое внимание было уделено измерению инноваций в области цифровизации. Еще в 1995 году Ринальдо Евангелиста и Джорджо Сирилли писали: «Растущее экономическое и технологическое значение сектора услуг в современных обществах требует более систематического

сбора данных об инновационной деятельности в этих отраслях» [Evangelista, Sirilli 1995, 207]. Более того, сегодня инновации становятся уделом не только бизнеса, но и некоммерческой деятельности, взаимодействия органов власти с гражданами. По мнению Фреда Голта, «до 2018 года существовала потребность только в одном секторе – предпринимательском, но теперь существует определение инноваций, применимое ко всем секторам экономики, таким как сектор государственного управления (управления органов власти в сочетании с совокупностью государственных корпораций), сектор домашних хозяйств и сектор некоммерческих организаций, обслуживающих домашние хозяйства» [Gault 2020]. Все это привело к тому, что в конце второго десятилетия появились новые стандарты в области управления инновациями с учетом цифровизации экономики.

Так, например, управлению инновациями посвящены стандарты серии 56000 Международного института стандартизации (*International Organization for Standardization – ISO*). Стандарт *ISO* 56000:2020 «Innovation management – Fundamentals and vocabulary» содержит глоссарий в области инноваций. Стандарт *ISO* 56002:2019 «Innovation management – Innovation management system – Guidance» посвящен основам инновационного управления и локализован как ГОСТ Р ИСО 56002–2020 «Инновационный менеджмент. Системы инновационного менеджмента. Руководящие указания». Стандарт *ISO* 56003:2019 «Innovation management – Tools and methods for innovation partnership – Guidance» посвящен инструментам и методам инновационного партнерства (например, при взаимодействии организаций и стартапов), русский перевод – ГОСТ Р ИСО 56003–2020 «Инновационный менеджмент. Методы и средства организации инновационного партнерства. Руководящие указания».

Стандарт *ISO/TR* 56004:2019 «Innovation Management Assessment – Guidance», посвященный оценке качества управления инновациями, аналогично предыдущему имеет русский перевод, хотя и с отличным от *ISO* номером в названии: ГОСТ Р 59062-2020/*ISO*/*TR* 56004:2019 «Оценка инновационного менеджмента. Руководящие указания». Стандарт *ISO* 56005:2020 «Innovation management – Tools and methods for intellectual property management – Guidance» посвящен рекомендациям в области инструментов и методов управления интеллектуальной собственностью. Ряд стандартов *ISO* этой серии находится еще в разработке. Среди

них – стандарт *ISO/AWI* 56001 «Innovation management – Innovation management system – Requirements», посвященный системам управления инновациям, стандарт *ISO/FDIS* 56006 «Innovation management – Tools and methods for strategic intelligence management – Guidance», где рассматриваются инструменты и методы стратегического интеллектуального управления, и стандарт *ISO/AWI* 56007 «Innovation management – Tools and methods for idea management – Guidance», в котором представлены инструменты и методы управления идеями. Пока ведется работа над стандартом *ISO/AWI* 56008 «Innovation management – Tools and methods for innovation operation measurements – Guidance», в нем будут изложены рекомендации в области использования инструментов и методов измерения инновационной деятельности.

Основу всех стандартов *ISO* в области инноваций в части оценки инноваций во многом определяет справочное руководство «Измерение научно-технической деятельности. Принципы сбора и интерпретации данных о технологических инновациях», которое выпускает Организация экономического сотрудничества и развития (ОЭСР) с 1992 года. Данное руководство получило название *Oslo Manual* (Руководство Осло) и входит в общее семейство стандартов сбора статистики *Frascati Manual,* разработанных еще для ОЭСР в 1963 голу группой *NESTI* (*National Experts on Science and Technology Indicators*) в итальянском городке Фраскати. В четвертом издании Руководства Осло [OECD, Eurostat 2018] учтена современная практика инноваций, в том числе обсуждается роль цифровизации для инновационного развития предприятий, важность работы с данными и использования цифровых платформ.

Руководство Осло по измерению инноваций

В Руководстве Осло основополагающим считается принцип возможности и необходимости измерения инноваций. В документе дано определение: «Инновация – это новый или улучшенный продукт или процесс (или их комбинация), который значительно отличается от предыдущих продуктов или процессов подразделения и который был предоставлен потенциальным пользователям (как продукт) или введен в эксплуатацию подразделением (как процесс)» [OECD, Eurostat 2018]. В редакции 2018 года типы инноваций сокращены до двух: продукт и процесс, убраны маркетинговые и организационные инновации, которые по своей сути являются процессными. Основную роль в инновациях играет

знание, а не только новизна и полезность, явно обозначенная в формулировке. Важно, что понятие инновации относится как к самой деятельности, так и к продукту такой деятельности. Соответственно, предприятие может быть вполне инновационным, даже если выпускает продукцию, которую к инновациям отнести сложно, например, какой-нибудь продукт питания. В этом смысле к инновационным можно отнести многие задачи по повышению эффективности бизнес-процессов, которые решаются с использованием цифровых технологий и искусственного интеллекта.

В Руководстве Осло большое внимание уделяется использованию цифровых технологий для сбора статистических данных об инновациях (глава 9), внедрению автоматизированных методов сбора данных при опросе респондентов об инновационности (главы 9, 10), использованию инструментов семантического анализа и визуализации собранной информации (глава 11). Однако для цели данного исследования интерес представляют: обсуждение роли информации, представления такой роли с точки зрения инновационности одновременно продукта и процесса (глава 3), потенциальная инновационная роль управления данными наряду с разработкой программного обеспечения (глава 4), данные с точки зрения информационного контента и возможности использования для инновационного развития организации (глава 5). Именно эти моменты целесообразно обсудить подробнее.

Как уже говорилось, в целях измерения инновации делятся в зависимости от объекта: инновации в продуктах и инновации в бизнес-процессах. «Инновации в продуктах должны обеспечивать значительное улучшение одной или нескольких характеристик или технических характеристик. Это включает в себя добавление новых функций или улучшение существующих функций или пользовательских утилит» (раздел 3.25). Таким образом, улучшение цифрового продукта или продукта, у которого есть цифровая составляющая (например, автомобиль с элементами автоматического ассистирования), является инновацией. Применительно к продуктам различают инновации в товарах и инновации в услугах. Обсуждение того, как соотносятся инновационность и доходность (раздел 3.28), показывает, что нецелесообразно ставить инновационность в зависимость от дохода, поскольку, например, часто цифровые инновационные продукты (товары или услуги) могут быть предоставлены потребителю бесплатно, доход предпринимателю приносит реклама, сопровождающая продажу

таких продуктов. По всей видимости, при оценке инновационности продуктов было бы правильно говорить об их ценности для потребителей. В стандарте говорится о том, что ценность является неявной целью инноваций (раздел 2.22), однако при этом делается замечание, что в имеющихся системах сбора статистической информации нет единого показателя экономической или социальной ценности. Необходимо учитывать, что использование ценности, возникающей благодаря технологическим изменениям, в значительной степени обусловлено человеческим капиталом, инвестициями в исследования и разработки, финансовыми ресурсами, эффективным управлением [Castaño-Martínez 2020]. При этом необходимо учитывать траектории инноваций с точки зрения процесса создания стоимости. Так, например, в работе [Shimelis, Eshetie 2021] предлагается выделять элементы, которые характеризуют инновационную цепочку создания стоимости: поиск, отбор и распространение знаний на уровне фирмы, бизнес-модель, реализация.

В цифровую эпоху продукт создается при активном участии пользователя, например заказ через Интернет сувенирной продукции с пользовательскими логотипами или личных фотоальбомов, перевод денег через личный кабинет на сайте банка и т.п. По всей видимости, к инновационным стоит отнести создание новых услуг, основанных на информационном самообслуживании. В Руководстве предлагается отдельно выделить наукоемкие продукты (knowledge-capturing products), получившие особое распространение благодаря развитию информационных технологий и снижению стоимости цифровых носителей. В качестве примера приведем сетевые информационные базы данных. Сама информация оказывается одновременно и товаром, и предметом услуги: «…продукты для сбора знаний похожи на товар, если потребители могут делиться или продавать их другим после покупки, но они похожи на услугу, если права потребителя ограничены лицензией, которая ограничивает обмен или продажу» (раздел 3.32).

В четвертом издании Руководства Осло цифровые технологии выделены в отдельную группу, в третьем они составляли одну группу с учетными функциями. В разделе 4.2 определены различные виды инновационной деятельности: научно-исследовательские и опытно-конструкторские работы, инженерная и проектная деятельность, маркетинг и продвижение бренда, работа с интеллектуальной собственностью, обучение

сотрудников, приобретение или аренда материальных активов, управление инновациями, отдельно указана деятельность по разработке программного обеспечения (ПО) и баз данных. В состав последней включены: собственная разработка и покупка ПО, как системного, так и прикладного (включая ПО, встроенное в продукты или оборудование); приобретение или внутренняя разработка компьютерных баз данных, включая сбор и анализ данных; мероприятия по модернизации или расширению функций информационных систем (ИС). Инновации в бизнес-процессах затрагивают производство продуктов, дистрибуцию и логистику, маркетинг и продажи, информационно-коммуникационные технологии, управление и администрирование, разработку новых продуктов и бизнес-процессов.

В работе [Huesig, Endres 2018] изучаются факторы, влияющие на внедрение конкретных программных средств для поддержки методов управления инновациями и их специфической функциональности. В статье [Wiesböck, Hess 2019] предлагается концепция, согласно которой реализация и внедрение цифровых инноваций в организации происходит по трем концентрическим кольцам: разработка, основанная на технологиях, и различные категории внедрения цифровых инноваций в основу (первое вольцо), средства, способствующие цифровым инновациям (второе кольцо), и управление цифровыми инновациями (третье кольцо).

Разработка ПО, ведущая к незначительным изменениям, а также «приобретение и анализ баз данных для бухгалтерского учета и других обычных бизнес-функций» (раздел 4.27), не являются инновационными. И наоборот, разработка ПО, которая ведет к появлению новых или улучшению имеющихся бизнес-процессов, работа с базами данных, используемых для анализа данных о свойствах материалов или предпочтениях клиентов, являются инновационными. Поскольку разработка ПО и баз данных включает в себя покупку таких продуктов, можно сказать, что практически вся деятельность ИТ службы, связанная с развитием информационной системы (в отличие от деятельности по поддержке ИС), является инновационной с точки зрения методологии оценки инноваций ОЭСР.

Особенности оценки инноваций на основе цифровых технологий

Среди технологических возможностей для инноваций указаны экспертиза, проектирование и использование цифровых

технологий и анализа данных (раздел 5.5). Выделение цифровых технологий и анализа данных в отдельный тип технологических возможностей связано с их универсальностью. В недалеком прошлом к технологиям, которые дают прорывные возможности, относились «биотехнологии, передовые методы производства, нанотехнологии, ИКТ и их приложения. Более поздними областями интереса стали квантовые вычисления, искусственный интеллект и робототехника, а также приложения на базе Интернета, такие как облачные сервисы и аналитика больших данных» (раздел 5.80).

Цифровые технологии определяются как «электронные инструменты, системы, устройства и ресурсы, которые генерируют, хранят, обрабатывают, обмениваются или используют цифровые данные». В руководстве проводится различие между понятиями:

• *digitisation* (оцифровка) – преобразование аналоговой информации (видео, изображение, текст, звук и т.п.) в цифровой формат (двоичные биты);

• *digitalisation* (цифровизация) – применение цифровых технологий на уровнях организации, отрасли, государства и на межгосударственном уровне.

Именно цифровизация имеет прямое отношение к инновациям. Более того, с точки зрения методологов ОЭСР в области инноваций, сегодня внедрение цифровых технологий, электронных средств связи, а также инструментов анализа данных, включая искусственный интеллект (ИИ), сегодня создает благоприятные возможности для инновационного развития предприятий. Поэтому при оценке инновационности учитывается, насколько в организации развиты ИТ, имеется ли соответствующее подразделение, бюджет, стратегия и т.д.

Для использования цифровых технологий в инновациях у персонала должны быть цифровые навыки, что также говорит и о социогуманитарной сущности самих цифровых технологий. Цифровые компетенции, по всей видимости, должны быть включены в критерии оценки инноваций, использующих цифровые технологии и ИИ. На первый взгляд, цифровые технологии заменяют человеческий труд. Фактически же, заменяя рутинную деятельность человека, цифровые технологии и ИИ создают новые возможности для разработки новых продуктов, внедрения новых процессов, а следовательно, увеличивают потребность в

человеческом труде. Именно поэтому сегодня в ИТ отрасли во всех странах мира наблюдается кадровый голод. «Общей особенностью цифровых технологий является их способность соединять различные виды деятельности и бизнес-функции, формируя интегрированную систему со структурированным обменом данными между различными функциями и подразделениями» (раздел 5.105). Таким образом, оценивая инновации, основанные на цифровых технологиях, целесообразно оценивать, насколько инновационной оказалась интеграция различных видов деятельности и функций.

Возможно, наиболее значимо то, что цифровые технологии «позволяют фирмам генерировать и хранить огромные объемы данных (часто в режиме реального времени) по целому ряду бизнес-операций, как внутри фирмы, так и связанных с поставщиками и пользователями» (раздел 5.106). Благодаря использованию аналитических инструментов (включая ИИ) и накапливаемым данным можно выстраивать новые стратегии, новые бизнес-модели и сервисы, которые сами по себе являются инновациями. К инновациям, основанным на цифровых технологиях, целесообразно относить инновации, которые не только напрямую используют цифровые инструменты, но и возникли в результате использования аналитических инструментов работы с данными. Для оценки инновационности организации Руководство Осло прямо рекомендует опрашивать сотрудников, используются ли на их предприятиях «методы и инструменты анализа данных, как внутри компании, так и за счет приобретения услуг анализа данных извне: систем управления базами данных, инструментов интеллектуального анализа данных, машинного обучения, моделирования данных, прогнозной аналитики, анализа поведения пользователей и анализа данных в реальном времени». В документе приводится ссылка на исследование [The OECD Model… 2015], которое показывает значительную роль инноваций, основанных на цифровых технологиях, благодаря чему методологи (раздел 5.107) при оценке предлагают не делать различий между инновациями, если они «содержат или были разработаны с использованием цифровых технологий».

Именно в силу важности цифровых технологий для инновационного развития организаций в Руководстве Осло предлагается особо оценивать общую цифровую компетенцию, «которая отражает способность фирмы извлекать выгоду из цифровизации и

решать связанные с этим проблемы. Некоторые соответствующие аспекты цифровой компетентности включают показатели:

• цифровой интеграции различных бизнес-функций;

• доступа и способности использовать аналитику данных для проектирования, разработки, коммерциализации и улучшения продуктов, включая данные о пользователях продуктов организации и использовании ими таких продуктов;

• доступа к сетям и использования соответствующих решений и архитектур (аппаратного и программного обеспечения);

• эффективное управление рисками конфиденциальности и кибербезопасности;

• внедрение соответствующих бизнес-моделей для цифровых сред, таких как электронная коммерция, цифровые платформы и т.д.» (раздел 5.108).

Интересно, что в Руководстве Осло говорится также об использовании цифровых платформ. «Цифровые платформы являются отличительной чертой цифровой эпохи. Платформы объединяют производителей и пользователей на различных этапах цепочки создания стоимости. Они часто образуют экосистему, в которой разрабатываются и продаются новые продукты, а также генерируются и обмениваются данными» (разделе 5.110). По мнению методологов оценки инноваций, использование цифровых платформ свидетельствует о потенциале организации в использовании цифровых технологий. Кроме того, использование цифровых платформ, объединяющих поставщиков, производителей и потребителей, генерирует данные, которые также могут способствовать росту инноваций. Ссылаясь на исследование об использовании цифровых платформ [Evans, Gawer 2016], авторы Руководства Осло делают вывод, что «эти платформы обеспечивают благодатную почву для развития и распространения инноваций» (раздел 7.53), и, следовательно, для оценки инноваций целесообразно выяснить, насколько в организации и у ее партнеров используются цифровые платформы. Возможно, что и для оценки инноваций, использующих цифровые технологии, стоит учитывать их участие в работе цифровых платформ. Так, например, в статье [Paredes-Frigolett, Pyka 2022] представлена глобальная модель цифровых платформ в стейкхолдинговом капитализме. Описывается роль, которую большие данные играют в формировании новых расширенных инновационных экосистем цифровых платформ.

Стандарты ISO в области инноваций

Стандарты *ISO* в области управления и оценки инноваций разрабатываются техническим комитетом *ISO/TC* 279 «Стандартизация терминологии, инструментов, методов и взаимодействия между соответствующими сторонами для обеспечения инноваций». Его работу поддерживает французская организация по стандартизации *Association Française de Normalisation* (*AFNOR*). Именно *AFNOR* выступила с предложением создания серии стандартов *ISO*, посвященных инновациям. Основу новых документов должны составить на основе европейских стандартов серии *CEN/TS* 16555 [Хохлявин, Кудрявцева 2013], которые, в свою очередь, обобщили опыт стандартизации менеджмента инноваций в Великобритании, Ирландии, Испании и Португалии.

Стандарт *ISO* 56002 посвящен общим вопросам организации инновационной деятельности на предприятиях. В нем сформулированы принципы инновационного менеджмента: реализация ценности, лидерство, нацеленное в будущее, стратегическая направленность инноваций, инновационная культура, использование уникальной и достоверной информации, управление неопределенностью, адаптируемость и системный подход к инновациям [Tidd 2021]. Дана общая схема системы инновационного менеджмента (*Innovation Management System, IMS*), куда входит среда предприятия, лидерство, инструменты инновационного развития. Как и в большинстве стандартов *ISO*, гармонизированных со стандартом ISO 9001, основным форматом развития является цикл Деминга (*Plan – Do – Check – Act, PDCA*). С точки зрения анализа критериев оценки инноваций интересна формулировка требований к целям инновационного развития, которые должны распространяться и на инновации, использующие цифровые технологии. В стандарте отмечено семь требований к целям инновационного развития: соответствие инновационной политике организации, наличие на всех уровнях и для всех функций организации, измеримость и верифицируемость, соответствие всем применимым требованиям, возможность непрерывного контроля, доступность для передачи и понятность, возможность обновления при необходимости быть обновляемым.

Стандарт *ISO* 56004 (ГОСТ 59062) посвящен оценке инновационного менеджмента (инструменты и метрики будут описаны в стандарте ISO 56008, который пока разрабатывается). В нем даны определения и основные подходы к оценке инновационного

менеджмента (*Innovation Management Assessment, IMA*) на основе Руководства Осло. Для проведения *IMA* в стандарте *ISO* 56004 предлагается определить следующие показатели:

• цель *IMA* (для выявления соответствия плану, для совершенствования *IMS* и/или для повышения возможностей инновационного менеджмента);

• охват *IMA* (одно, несколько или все подразделения организации);

• оцениваемые объекты (один или все);

• тип экспертизы (внутренняя или внешняя);

• тип сбора данных (по документам, интервью или опрос);

• средства сбора данных (вручную или автоматически);

• типы данных (качественные, количественные);

• инструменты анализа (вручную, частично или полностью автоматизированные);

• тип сравнения и ссылок (до и после, корреляционный анализ, бенчмаркинг);

• тип результатов (сильные/слабые стороны, недостатки, рекомендации);

• формат вывода результатов и др.

Для оценки инноваций, использующих цифровые технологий, также целесообразно определить цель и охват, типы и инструменты сбора данных, типы сравнения и выводы по результатам. Учитывая универсальность цифровых технологий и доступность для интеграции различных областей деятельности, считаем такую многомерную оценку безусловно целесообразной.

Заключение

Исследование оснований критериев оценки инноваций в области оценки инновационного менеджмента, использующего ИИ и ИТ, позволяет сделать следующие выводы:

• К инновациям, использующим цифровые технологии и ИИ, целесообразно отнести не только инновации, задействующие цифровые инструменты, но и инновации, которые были получены с использованием цифровых технологий и ИИ. При этом допустимо установить критерии оценки, учитывающие различие между такими инновациями.

• Поскольку цифровые технологии позволяют интегрировать различные функции и отрасли, целесообразно учесть масштаб и глубину такой интеграции. Например, инновация в области дис-

танционного определения заражения коронавирусом на основе опроса может одновременно использовать технологии обработки больших данных, безопасный доступ к медицинским данным, психологические особенности дистанционного интервьюирования и т.п.

• Инновации тесно связаны с использованием цифровых технологий, а значит, критерии оценки инноваций должны учитывать, насколько эффективно использует ИТ не только сама инновация, но и насколько цифровые технологии используются при доставке продуктов инновации клиентам, в работе с инновацией в организации и т.д.

• К критериям оценки качества инноваций целесообразно отнести то, насколько она эффективно генерирует качественные данные, в какой мере порождает новые инновации.

• Критерием инновационности также может служить то, используется ли в инновации (для создания или для предоставления продукта или услуги) платформа, где взаимодействуют поставщики, производители и потребители товаров и услуг.

• При разработке критериев оценки инновационности необходимо уделить особое внимание понятию ценности. Часто ценность сводят к коммерческому успеху или к повышению интеллектуальной стоимости, однако это понятие гораздо шире, и должно также иметь социо-гуманитарное измерение.

ЦИТИРУЕМАЯ ЛИТЕРАТУРА

Хохлявин, Кудрявцева 2013 – *Хохлявин С., Кудрявцева Ю.* Европейский стандарт для управления инновационной деятельностью как предтеча международного // Стандарты и качество. 2013. № 11. С. 46–48.

Castaño-Martínez 2020 – *Castaño-Martínez M.S.* Innovation, Value Creation, and Entrepreneurship by Opportunity // *Galindo-Martín M.A., Mendez-Picazo M.T., Castaño-Martínez M.S.* Analyzing the Relationship Between Innovation, Value Creation, and Entrepreneurship. – Hershey, PA: IGI Global, 2020. P. 43–63.

Evangelista, Sirilli 1995 – *Evangelista R., Sirilli G.* Measuring Innovation in Services // Research Evaluation. 1995. Vol. 5. No. 3. P. 207–215.

Evans, Gawer 2016 – *Evans P.C., Gawer A.* The rise of the platform enterprise: A Global Survey // The Center of Global Enterprise. 2016. – URL: https://www.thecge.net/app/uploads/2016/01/PDF-WEB-Platform-Survey_01_12.pdf

Gault 2020 – *Gault F.* Measuring Innovation Everywhere: The Challenge of Better Policy, Learning, Evaluation and Monitoring. – Northampton, MA: Edward Elgar Publishing, 2020.

Huesig, Endres 2018 – *Huesig S., Endres H.* Exploring the Digital Innovation Process: The Role of Functionality for the Adoption of Innovation Management Software by Innovation Managers // European Journal of Innovation Management. 2018. Vol. 22. No. 2. P. 302–314.

OECD, Eurostat 2018 – *OECD, Eurostat.* Oslo Manual 2018: Guidelines for Collecting, Reporting and Using Data on Innovation / 4th ed. – Paris: OECD Publishing; Luxembourg: Eurostat, 2018.

Paredes-Frigolett, Pyka 2022 – *Paredes-Frigolett H., Pyka A.* The Global Stakeholder Capitalism Model of Digital Platforms and Its Implications for Strategy and Innovation from a Schumpeterian Perspective // Journal of Evolutionary Economics. 2022. Vol. 32. No. 2. P. 463–500.

Shimelis, Eshetie 2021 – *Shimelis T., Eshetie B.* Innovation Value Chain: A Systematic and Narrative Review // International Journal of Quality and Innovation. 2021. Vol. 6. No. 1. P. 91–114.

The OECD Model… 2015 – The OECD Model Survey on ICT Usage by Businesses: 2nd revision // Working Party on Measurement and Analysis of the Digital Economy. OECD. 2015. – URL: https://www.oecd.org/sti/ieconomy/ICT-Model-Survey-Usage-Businesses.pdf

Tidd 2021 – *Tidd J.* A Review and Critical Assessment of the ISO56002 Innovation Management Systems Standard: Evidence and Limitations // International Journal of Innovation Management. 2021. Vol. 25. No. 1. P. 1–17.

Wiesböck, Hess 2019 – *Wiesböck F., Hess T.* Digital Innovations // Electronic Markets. Vol. 30. No. 1. P. 75–86.

REFERENCES

Castaño-Martínez M.S. (2020) Innovation, Value Creation, and Entrepreneurship by Opportunity. In: Galindo-Martín M.A., Mendez-Picazo M.T., Castaño-Martínez M.S. *Analyzing the Relationship Between Innovation, Value Creation, and Entrepreneurship* (pp. 43–63). Hershey, PA: IGI Global.

Evangelista R. & Sirilli G. (1995) Measuring Innovation in Services. *Research Evaluation.* Vol. 5, no. 3, pp. 207–215.

Evans P.C. & Gawer A. (2016) The rise of the platform enterprise: A Global Survey. *The Center of Global Enterprise.* Retrieved from https://www.thecge.net/app/uploads/2016/01/PDF-WEB-Platform-Survey_01_12.pdf

Gault F. (2020) *Measuring Innovation Everywhere: The Challenge of Better Policy, Learning, Evaluation and Monitoring.* Northampton, MA: Edward Elgar Publishing.

Huesig S. & Endres H. (2018) Exploring the Digital Innovation Process: The Role of Functionality for the Adoption of Innovation Management Software by Innovation Managers. *European Journal of Innovation Management.* Vol. 22, no. 2, pp. 302–314.

Khokhlyavin S. & Kudryavtseva Yu. (2013) The European Standard for Innovation Management as the Forerunner of the International. *Standardy i kachestvo.* No. 11, pp. 46–48 (in Russian).

OECD (2015) *The OECD Model Survey on ICT Usage by Businesses: 2nd revision*. Retrieved from https://www.oecd.org/sti/ieconomy/ICT-Model-Survey-Usage-Businesses.pdf

OECD, Eurostat (2018) *Oslo Manual 2018: Guidelines for Collecting, Reporting and Using Data on Innovation* (4th ed.). Paris: OECD Publishing; Luxembourg: Eurostat.

Paredes-Frigolett H. & Pyka A. (2022) The Global Stakeholder Capitalism Model of Digital Platforms and Its Implications for Strategy and Innovation from a Schumpeterian Perspective. *Journal of Evolutionary Economics.* Vol. 32, no. 2, pp. 463–500.

Shimelis T. & Eshetie B. (2021) Innovation Value Chain: A Systematic and Narrative Review. *International Journal of Quality and Innovation.* Vol. 6, no. 1, pp. 91–114.

Tidd J. (2021) A Review and Critical Assessment of the ISO56002 Innovation Management Systems Standard: Evidence and Limitations. *International Journal of Innovation Management.* Vol. 25, no. 1, pp. 1–17.

Wiesböck F. & Hess T. (2019) Digital Innovations. *Electronic Markets.* Vol. 30, no. 1, pp. 75–86.

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 25.06.2022 г. Формат 60x90/16. Печать цифровая.
Бумага офсетная № 1. Печ. л. 10,0. Тираж 1000 экз. Заказ
Отпечатано в типографии Издательского дома «Гуманитарий».
E-mail: humanist@academyrh.info