DOI: 10.30727/0235-1188-2025-68-1-116-132
Оригинальная исследовательская статья
Original research article

Рождение субъектности у искусственного интеллекта: фантастика или реальная угроза?

А.А. Грибков
Научно-производственный комплекс «Технологический центр», Москва, Россия

Аннотация

В статье рассматривается тенденция все более широкого использования систем искусственного интеллекта, являющаяся следствием перехода человеческой цивилизации на новую стадию развития – стадию цивилизации когнитивных технологий, соответствующую удовлетворению потребностей посредством замещения людей искусственными когнитивными системами в интеллектуальном управлении машинами. Исследуется возможность и опасности рождения субъектности у искусственного интеллекта. Рассматриваются понятия сознания, самосознания и субъектности, констатируется ошибочность их отождествления. Предлагается авторское определение понятия субъектности, выражаемое через понятия сознания и самосознания. Выявляется связь между субъектностью и наличием потребностей. В контексте оценки возникающих угроз для человечества со стороны альтернативного интеллекта выделяется случай одновременного обладания потребностями и высоким уровнем разума. Опираясь на результаты предыдущих исследований, автор утверждает возможность достижения разумности искусственным интеллектом, в том числе обретения им способности решать творческие задачи. Средством для этого может стать общая теория систем, в том числе описанный автором метод мультисистемной интеграции знаний. Показано, что потребности искусственный интеллект самостоятельно не может получить, а человеку наделять его ими нет практической необходимости: для решения интеллектуальных задач субъектность не требуется. Кроме того, наделенный субъектностью искусственный интеллект при определенных условиях может представлять для человечества реальную угрозу. У него будут потребности и желания, которые неизбежно вступят в противоречие с потребностями и желаниями людей. В противостоянии с субъектным искусственным интеллектом, превосходящим человека по интеллекту, человечество может проиграть. Поэтому искусственный интеллект никогда не должен получить субъектность.

Ключевые слова: философия искусственного интеллекта, теория систем, сознание, самосознание, когнитивная система, потребности, разум, творчество, этика искусственного интеллекта.

Грибков Андрей Армович – доктор технических наук, ведущий научный сотрудник НПК «Технологический центр».

andarmo@yandex.ru

https://orcid.org/0000-0002-9734-105X

Для цитирования: *Грибков А.А.* Рождение субъектности у искусственного интеллекта: фантастика или реальная угроза? // Философские науки. 2025. Т. 68. № 1. С. 116–132. DOI: 10.30727/0235-1188-2025-68-1-116-132

The Emergence of Agency in Artificial Intelligence: Science Fiction or a Real Threat?

A.A. Gribkov
Scientific-Manufacturing Complex "Technological Centre,"
Moscow, Russia

Abstract

The article examines the trend of the increasingly widespread use of artificial intelligence (AI) systems, which is a consequence of human civilization's transition to a new stage of development: the civilization of cognitive technologies. This stage is characterized by the satisfaction of needs through the replacement of humans with artificial cognitive systems for the intelligent control of machinery. The paper investigates the possibility and dangers of the emergence of agency in artificial intelligence. The concepts of consciousness, self-consciousness, and agency (subjectivity) are analyzed, and it is argued that their conflation is erroneous. The author proposes an original definition of agency, formulated in terms of consciousness and self-consciousness. A link is established between agency and the possession of needs. In the context of assessing emerging threats to humanity from an alternative intelligence, the case of an entity simultaneously possessing both needs and a high level of intelligence is highlighted as a primary concern. Drawing upon previous research, the author asserts that it is possible for AI to achieve intelligence comparable to human reason, including the ability to solve creative problems. General system theory, particularly the author's proposed method of multisystem knowledge integration, is presented as a potential means to this end. It is shown that artificial intelligence cannot acquire needs autonomously, and there is no practical

necessity for humans to endow it with them, as agency is not required for solving intellectual problems. Moreover, an AI imbued with agency could, under specific circumstances, represent a tangible threat to humanity. Its needs and desires would inevitably enter into conflict with those of people. In a confrontation with an agentive AI of superior intellect, humanity would likely be vanquished. Therefore, the author concludes, artificial intelligence must never be allowed to attain agency.

Keywords: philosophy of artificial intelligence, systems theory, consciousness, self-consciousness, cognitive system, needs, intelligence, creativity, AI ethics.

Andrey A. Gribkov – D.Sc. in Technology, Leading Research Fellow, Scientific-Manufacturing Complex "Technological Centre."
andarmo@yandex.ru
https://orcid.org/0000-0002-9734-105X

For citation: Gribkov A.A. (2025) The Emergence of Agency in Artificial Intelligence: Science Fiction or a Real Threat? *Russian Journal of Philosophical Sciences = Filosofskie nauki.* Vol. 68, no. 1, pp. 116–132.
DOI: 10.30727/0235-1188-2025-68-1-116-132

Введение

Одним из ключевых факторов технологического развития, определяющих глобальную конкурентоспособность стран, в последние годы стали системы искусственного интеллекта. Ведущие мировые медиа- и финансовые корпорации, IT-компании, консалтинговые, высокотехнологичные промышленные предприятия все активнее используют системы искусственного интеллекта и, более того, выстраивают на их основе свою деятельность. Необходимо признать, что технологии искусственного интеллекта будут развиваться и в дальнейшем, их роль будет возрастать. Вариант технологического развития, при котором человечество откажется от использования искусственного интеллекта, не представляется возможным.

Искусственный интеллект становится неотъемлемой частью человеческой цивилизации. И это – естественное следствие ее развития, главным технологическим содержанием которого является замещение человека машинами при выполнении все более широкого круга задач, от физического труда по созданию материальных благ за счет построения соответствующих машин, управления последними по заданным алгоритмам до интеллектуального управления этими машинами.

Стадию развития цивилизации, соответствующую удовлетворению потребностей посредством замещения людей искусственными когнитивными системами в интеллектуальном управлении машинами, будем называть цивилизацией когнитивных технологий [Грибков 2024а]. Когнитивные системы – существенно более обширная группа систем по сравнению с системами искусственного интеллекта. Когнитивная система с точки зрения философии – это многоуровневая система, осуществляющая функции распознавания и запоминания информации, принятия решений, хранения, объяснения, понимания и производства новых знаний [Меркулов 2004]. Формирование цивилизации когнитивных технологий сопряжено с возникновением новых рисков, связанных с переопределением места человека в новой цивилизационной архитектуре, в том числе в определении неотчуждаемой функции человека в рамках цивилизации, заключающейся в генерации потребностей. Важной частной реализацией указанных рисков цивилизации когнитивных технологий служит возможность выхода искусственного интеллекта из-под контроля человека.

Популярной темой фантастического кино является бунт машин, восстающих против человечества, подчиняющих, порабощающих или даже уничтожающих людей. Почему возник страх перед «умными» машинами? Возможно, как это часто происходит, общественное сознание аккумулирует основанное на объективных предпосылках предчувствие существующей в действительности угрозы?

22 марта 2023 года было опубликовано открытое письмо, призывающее приостановить на полгода обучение систем искусственного интеллекта, более мощных чем GPT-4, пока не будут созданы, внедрены и проверены независимыми экспертами общие протоколы безопасности. В нем утверждается: «Мощные системы искусственного интеллекта следует разрабатывать только тогда, когда мы уверены, что их эффекты будут положительными, а риски – управляемыми»[1]. Письмо подписали Илон Маск, Стив Возняк и к настоящему времени еще более 30 тыс. человек. К сожалению, необходимые работы по обеспечению безопасности до настоящего времени в полной мере не проведены. Преобладание коммерциализации в развитии технологий искусственного интеллекта стало причиной ухода из американской компании OpenAI, одной из ведущих в мире в этой области, ряда ключевых специалистов в области суперинтеллекта (среди них – сооснователь OpenAI Илья Суцкевер,

[1] Pause Giant AI Experiments: An Open Letter // Future of Life Institute. 2023. March 22. – URL: https://futureoflife.org/open-letter/pause-giant-ai-experiments/.

Ян Лейке, Майлз Брандейдж, Сучир Баладжи и др.) и безопасности искусственного интеллекта (среди них – Леопольд Ашенбреннер, Лилиан Венг и др.).

Генезис и последствия субъектности

Одним из первых вопросов, на которые следует ответить для понимания степени и реальности угрозы бунта машин, является вопрос о причинах такого бунта. Условно их можно разделить на две группы: либо машины устраивают бунт в интересах людей, либо делают это для себя.

В некоторых фильмах обыгрывается идея о том, что разумные машины заботятся о людях. Люди разрушают экологию планеты, ведут бесконечные войны и т.д., поэтому машинам нужно взять контроль над человечеством, чтобы спасти его от самоуничтожения. Приведенный «условно хороший» вариант в действительности является крайне опасным, с возможными катастрофическими для человечества последствиями, которые требуют отдельного исследования. В настоящей статьи не станем рассматривать данный вариант.

Второй вариант: машины устраивают бунт для себя. И, чтобы он произошел, машины должны обладать правом принятия решений. Очевиден тот факт, что добровольно люди такого права и такой возможности машинам не предоставят. Следовательно, бунт возможен лишь в случае, если последние возьмут это право самостоятельно, по собственной инициативе. Это станет возможным, если машины (речь идет о системах искусственного интеллекта) будут обладать субъектностью.

Что означает термин «субъектность»? В философских и психологических статьях и книгах вкладывается принципиально различный смысл в понятия «сознание», «самосознание» и «субъектность». Иногда их даже отождествляют [Гегель 1977, 40–41; Кант 1994, 85, 127]. Отождествление обусловлено тем, что до недавнего времени единственным сознанием, с которым имело дело наука, было сознание человека, обладающего самосознанием и субъектностью. Развитие искусственных когнитивных систем, в частности систем искусственного интеллекта, требует универсализации понятий, нового, более точного определения терминов «сознание», «самосознание» и «субъектность», не обязательно связанного с человеком. Возможным подходом к такому определению служит информационная концепция сознания [Прыгин 2017; Степанский 2006; Дубровский 2009; Цветков 2021], которая наиболее полно и точно подходит для описания сознания как человеческого, так и

искусственного. Согласно авторской интерпретации информационной концепции, сознание – информационная среда, в которой реализуется расширенная модель реальности [Грибков, Зеленский 2023б]. В свою очередь, информационная среда – это система, образованная из информационных объектов, представляющих собой отражения (фиксированные или обновляемые) свойств реальных объектов [Грибков, Зеленский 2023б].

Самосознание – феномен сознания, центральным атрибутом которого является определение границ сознания в виде локализации носителем сознания своего положения в многомерном пространстве состояний исходных реальных объектов, а также их отражений в виде информационных объектов. Локализация носителя сознания обеспечивается наличием обратных связей. Через них сознание «определяет» границы между собой (носителем сознания) и окружающим миром. Тем самым человек осознает себя: осматривается, фиксирует свое положение относительно окружающих предметов, ощупывает их и себя и т.д. Двигаясь, испытывая зрительные, слуховые, тактильные, иные (не обязательно физические) ощущения, человек отслеживает обратные связи и через них осознает себя. Если изменения, демонстрирующие связность информационных объектов в рамках сознания, инициируются внешними по отношению к носителю сознания факторами, то самосознание у него будет, а субъектность не сформируется. Если инициатором аналогичных изменений выступает носитель сознания, то это предполагает наличие у него субъектности. Понятнее всего категорию субъектности можно объяснить через близкое к субъектности понятие «самость», которое определяется как способность субъекта выйти из-под контроля внешних причин и создать воспроизводящийся порядок жизни, детерминируемый изнутри [Гиренок 2010]. Интерпретируя данное описание для случая субъектности, можно сформулировать следующее определение: субъектность – это способность быть субъектом активности, в том числе в отношении объектов познания. Субъект принимает решение относительно того, что нужно реализовать ту или иную активность.

Движущей силой активности человека являются его потребности. Цивилизация – форма группового существования людей, обеспечивающая посредством социальных механизмов удовлетворение их биологических, социальных и интеллектуальных (духовных) потребностей [Грибков 2024а]. Таким образом, цивилизация служит цели удовлетворения человеческих потребностей. Вместе с тем желание их удовлетворить является основным стимулом активности.

Субъектность – это не свойство, исключительно присущее человеку. Субъектностью, в частности, обладают все животные и даже насекомые. Вопрос о субъектности растений представляется спорным: с одной стороны, они проявляют активность в удовлетворении своих потребностей и способны реагировать на внешние импульсы, с другой – организм растений, рассматриваемый как информационная среда (аналог сознания), характеризуется высокой децентрализацией и быстрым затуханием передаваемых электрических импульсов [Toyota et al. 2018], что не способствует формированию самосознания. Самосознание выступает необходимым условием формирования субъектности. Наряду с человеком, самосознанием обладают многие когнитивные системы, в том числе не имеющие широких возможностей при решении интеллектуальных или вычислительных задач. Обретение самосознания – это не слишком сложная задача. Самосознанием, например, обладают животные, в том числе насекомые, способные реагировать на внешние воздействия, отличая при этом себя от окружающей среды. Самосознанием (в принятой нами интерпретации) наделена несложная искусственная когнитивная система, например система автопарковки автомобиля, способная идентифицироваться (отделять свой автомобиль от остальных), определять собственное положение в пространстве относительно других автомобилей и места парковки.

Является ли наличие субъектности бинарной характеристикой? Безусловно, да. Частичной субъектности быть не может. Это, однако, не оспаривает существования автономных когнитивных систем, не имеющих субъектности, но реализующих самоуправление согласно заложенной программе. Внешне такие системы похожи на наделенные субъектностью. В чем же заключается отличие заложенной извне программы от заложенных в человека рефлексов, инстинктов и желаний? Отличие незначительное, и оно обусловлено инициатором активности: внутренними процессами в системе, инициирующими внешнюю активность, или заложенной извне программой, задающей активность в интересах внешних акторов.

Отдельного рассмотрения требует случай, при котором функция программы заключается в формировании в системе внутренних процессов, позволяющих ей реализовывать активность и существовать без внешнего управления. Система с такой программой может обладать субъектностью даже в случае, если ее активность служит интересам внешних акторов (вложившим в нее эту программу). Аналогично и активность дрессированного животного служит интересам его хозяина, а активность отдель-

ных групповых сообществ (в некоторых случаях) – целям их лидеров или организаторов. При этом субъектность индивидов не оспаривается.

В контексте вопроса о возможности частичной субъектности интерес представляют сложные системы, включающие в себя человека и искусственные когнитивные системы. Такие интегральные субъекты познания в целом обладают субъектностью, и существенную часть работы они выполняют без участия человека. Несубъектный искусственный интеллект [Грибков 2024в], даже не обладая субъектностью, способен в рамках поставленной человеком задачи формулировать и решать требуемые для ее решения частные подзадачи. Отсутствие субъектности накладывает лишь ограничение на инициализацию активности (познавательной и созидательной) целью удовлетворения потребностей, генерируемых человеком-оператором.

Креативный искусственный интеллект

Наличия субъектности недостаточно для того, чтобы представлять угрозу для человечества. Для отдельного человека собака может представлять угрозу, например может покусать его. Но для человечества в целом, конечно, ни собаки, ни иные животные не представляют угрозы. Чтобы для человечества возникла угроза, наряду с субъектностью, необходим высокий уровень разума. Именно разума, а не только рассудка. Это означает, что должна быть способность решения интеллектуальных задач, т.е. задач, для которых нет готовых решений [Рапацевич 1995, 39–40].

Способен ли искусственный интеллект достичь уровня человеческого разума, т.е. стать креативным [Грибков, Зеленский 2023а]? В настоящее время на поставленный вопрос предлагают два варианта ответа, и оба, по мнению автора данной статьи, ошибочны.

Согласно первому варианту ответа, искусственный интеллект, подобный человеческому, невозможен, поскольку человеческое сознание содержит в себе значительный, существенный, определяющий фактор иррациональности, трансцендентности. И, если его нет, то невозможны ни интуиция, ни прозрение, ни творчество, ни иные подобные феномены.

Опровержение представления об иррациональности мироздания – задача, решение которой осложнено тем, что принятие иррациональности открывает возможность отказа от объяснений причин и механизмов реализации объектов и процессов в мире. Тем не менее достоверное опровержение иррациональности видится возможным. В рамках разработанной автором статьи эмпирико-метафизической общей теории систем [Грибков 2024г] дано развернутое обоснование детерминиро-

ванности мироздания, основанное на исчисляемости материального бытия, верифицируемой его изоморфизмом. В парадигме эмпирико-метафизической общей теории систем интеллектуальная деятельность и творчество получают детерминированный характер.

Творческое мышление, как и рассудочное, имеет в своей основе определенные правила, логику. Решения, ответы, которые находятся в результате творчества, могут быть детерминированы. Методы и способы получения или нахождения творческих решений существенно отличаются от используемых при решении задач, для которых уже предусмотрены готовые решения, готовые ответы. Ключевую роль в решении творческих задач играют методы, которые формализуемы в рамках общей теории систем [Берталанфи 1969; Уемов 1978; Богданов 1989]. Эта теории исходит из изоморфизма мира, т.е. подобия форм и законов, которые лежат в основе объектов различной сложности в различных предметных областях. Иными словами, если человек творит, он не придумывает то, чего в мире не существует. Он смотрит вокруг, обобщает свой опыт из множества систем, в которые интегрирован, находит подобные формы, законы, отношения и использует их для решения творческих задач. Поскольку, как известно, не содержится в уме человека то, что прежде не было бы представлено в ощущении.

Описанное явление получило название мультисистемной интеграции знаний [Грибков 2024б]. Человек в своем сознании объединяет в целое знания из различных систем, в которые он интегрирован, частью которых он выступает. И на основании этого выстраивает логику организации и понимание мира. В рамках этой логики он находит ответы на вопросы и решает задачи, которые перед ним стоят.

Итак, был изложен первый вариант ответа. Существует и второй вариант ответа, согласно которому искусственный интеллект станет подобным человеческому, если достигнет высокого уровня сложности (например, в нем будет большое количество нейронов). Эта идея многократно отражена в фантастической литературе: если сложность компьютера достигает определенного уровня, он обретает субъектность, «оживает», становится подобным человеку. Способен ли искусственный интеллект стать сопоставимым человеку по вычислительной мощности? Безусловно, способен, и это уже произошло. Но это не означает, что искусственный интеллект стал подобным человеческому. Он не подобен человеческому интеллекту, поскольку у него отсутствует субъектность. Каковы потребности искусственного интеллекта? Имеются в виду потребности не формальные, которые может описать человек

(например, «компьютеру нужно электричество», «компьютеру нужен температурный режим» и т.д.; это внешнее описание), а с позиции самой компьютерной, вычислительной системы. Осознаваемые потребности у искусственного интеллекта отсутствуют, поскольку у него нет желаний, в том числе желания жить, он не живет, он неживой. Искусственному интеллекту можно задавать вопросы, и он будет предлагать правильные ответы. Но это – лишь правильные ответы, а не правда, не выражение его подлинных желаний и потребностей. Он изображает жизнь, но не является живым. Он не является живым, т.к. у него нет потребностей, нет желаний и нет субъектности.

Отсутствуют основания утверждать, что способность искусственного интеллекта решать сложные, в том числе творческие задачи, имеет объективные ограничения, которые не позволят ему достичь (и превзойти!) человека. Ограничения искусственного интеллекта, отделяющие его от естественного, человеческого интеллекта, – это ограничения иного рода, и они связаны с отсутствием у него субъектности. У биологических объектов (животных или человека) субъектность – необходимое для выживания свойство, к тому же не связанное с уровнем интеллекта. У искусственных объектов субъектность, как и остальные свойства, возникнет, если ее искусственно вложить (напрямую либо посредством ряда эволюционных механизмов, порождающих субъектность в процессе реализации). Случайно или в процессе масштабирования системы (например, увеличения количества элементов искусственной нейронной сети) субъектность не появится. В наиболее больших современных моделях искусственного интеллекта (например, GPT-4) количество используемых параметров оценивается в несколько триллионов, что сравнимо с количеством синапсов в мозге животных. Например, в мозге мыши несколько сотен миллиардов синапсов при 71 млн нейронов [Herculano-Houzel 2009]

Если бы для обретения субъектности требовалось лишь большое число нейронов, она давно бы проявилась у мощных нейронных сетей.

Может быть, субъектность присуща только живым системам? Отношения категорий субъектности и жизни понятны: что обладает субъектностью, то является живым. Однако обратное утверждение становится неверным: простейшие микроорганизмы и, возможно, растения не обладают субъектностью, однако являются живыми. В целом разделение объектов на живые и неживые сопряжено со значительными сложностями. Это разделение очевидно, если исходить из формальных признаков: живой объект – это объект, сформировавшийся в процессе биологической эволюции. Однако такое утверждение ничего не объ-

ясняет, а лишь констатирует способ получения объекта, который не определяет его конечных свойств.

Дополнительным фактором, и его следует учитывать при установлении различий живых и неживых систем, служит способ реализации их устойчивости. Для неживой природы характерна устойчивость, достигаемая за счет достижения баланса, равновесия сил, процессов и т.д. В живой природе главной формой устойчивости является неравновесная устойчивость [Бауэр 1935]. Принцип устойчивого неравновесия получает развитие в концепции динамической кинетической стабильности (ДКС): «…концепция ДКС совершенно отлична от обычного вида стабильности в природе – термодинамической стабильности…» [Pross, Pascal 2013], «…для наблюдения специфического поведения реплицирующихся систем необходимо постоянно поддерживать далекие от равновесия условия…» [Pascal, Pross, Sutherland 2013]. Приход живой системы в равновесие означает остановку в ней процессов, т.е. смерть.

К сожалению, неравновесная устойчивость не может быть рассмотрена как маркер жизни. Сложные, динамические неживые системы [Гленсдорф, Пригожин 1973], в частности системы креативного искусственного интеллекта, тоже будут характеризоваться такой формой устойчивости. Проведенные исследования показали перспективность использования для искусственных когнитивных систем механизмов неравновесной устойчивости [Грибков, Зеленский 2024]. Это обусловлено спецификой когнитивных систем, для которых основой является процесс мышления, неравновесный по своей природе.

Необходимо констатировать, что принципиальных структурных или организационных различий между искусственными и естественными системами (в том числе когнитивными) не существует. Это, однако, не означает, что человечество должно попустительствовать распространению различных альтернативных человеку когнитивных систем, пусть даже они и объективно не хуже нас. Человечество как биологический вид и как социальная общность вправе отстаивать свои интересы. В этом и заключается подлинный смысл гуманизма, в центре которого человек и только человек.

Субъектный искусственный интеллект

Можно ли наделить искусственный интеллект субъектностью? Да, можно. В наиболее простом случае это требует выполнения трех условий. Во-первых, наделение искусственного интеллекта механизмами преобразования потребностей (на начальном этапе формальных) в

активность по их удовлетворению. Во-вторых, «свобода воли», технически реализуемая в виде флуктуаций параметров информационных объектов в сознании искусственного интеллекта, масштаб которых достаточен для инициации последовательностей (цепочек) преобразований и взаимодействий этих объектов, но недостаточен для дестабилизации сознания, при котором спонтанные процессы преобладают над регулярными. В-третьих, спонтанно инициируемые или регулярные процессы, претерпеваемые информационными объектами в сознании, способствующие удовлетворению потребностей или улучшению (упрощению, гармонизации, вариативности и т.д.) их удовлетворения, должны закрепляться как позитивные посредством реакции снижения текущей активности или группы активностей (расслабления, удовлетворения). В результате искусственный интеллект приобретает свойства, функционально соответствующие желаниям и эмоциям, – стремление к определенным формам активности и удовлетворение при их реализации. В зависимости от полноты удовлетворения этих условий или их расширения, проявления субъектности могут быть более или менее выраженными, напоминающими субъектность животных и человека.

Возможно ли непреднамеренное удовлетворение таких условий в результате развития систем искусственного интеллекта? Вероятно, невозможно (особенно после того, как они нами указаны). Сформулированные условия формирования субъектности у систем искусственного интеллекта представляют собой максимально упрощенное определение генезиса субъектности животных и человека. Для эволюционно формирующихся систем эти условия естественны, для искусственно создаваемых – могут быть реализованы исключительно по аналогии с живыми системами.

Следует ли наделять искусственный интеллект субъектностью? Искусственный интеллект нам нужен для того, чтобы решать определенные сложные задачи, с которыми мы не справляемся. В идеальной ситуации мы хотим, чтобы искусственный интеллект помогал нам в решении интеллектуальных задач. Как мы утверждали ранее, интеллектуальные, творческие задачи формализуемы, и по мере развития общей теории систем можно «научить» искусственный интеллект решать творческие задачи. Для этого ему не требуется субъектность. Таким образом, с точки зрения прагматизма нет необходимости наделять искусственный интеллект субъектностью.

Вместе с тем, если какой-то безответственный ученый все-таки это сделает (а риск того, что подобное случится, велик), то наделенный субъектностью искусственный интеллект при определенных условиях может представлять для человечества реальную угрозу. У него будут

свои потребности и желания, которые, что очевидно, не будут соответствовать потребностям и желаниям людей. Если произойдет столкновение потребностей и желаний человека и искусственного интеллекта, при котором вторая из сторон будет превосходить по вычислительной, да и в целом по интеллектуальной мощности, человечество может проиграть.

Несубъектный искусственный интеллект, независимо от его интеллектуальной силы, является лишь инструментом (мощной вычислительной машиной), служащим интересам человека. Этические проблемы отношений человека и искусственного интеллекта в случае несубъектного искусственного интеллекта не возникают.

К числу проблем, связанных с появлением искусственного интеллекта, обладающего субъектностью, не относятся этические. Проведенные ранее автором исследования в области генезиса этических представлений показали, что область применения этики – это область общественных отношений, которая ограничена человечеством [Грибков 2023]. Убивая животных и употребляя в пищу их мясо, мы не совершаем зла, поскольку это не социальный акт, а биологический. Согласно этой же логике, любое злодеяние в отношении индивида, относящегося к человеческому виду, может при определенных условиях квалифицироваться как зло. Искусственный интеллект, в том числе наделенный субъектностью, не является частью человечества, а значит, этические нормы к нему неприменимы.

Определение характера отношений человека с другими носителями сознания, в том числе обладающими субъектностью, например животными или инопланетянами, если человечество когда-нибудь вступит с ними в контакт, – важный вопрос, требующий решения. Он, однако, находится за пределами этики. В одних случаях (при отношении к животным и растениям) – это вопрос сохранения биоразнообразия и в целом биологической среды, в которой существует общество. В других (в гипотетической ситуации с инопланетянами) – вопрос распространения сферы общественных отношений и на субъекты альтернативного происхождения в той мере, в которой это служит интересам (пользе) человечества.

Тема контакта человечества с инопланетной цивилизацией, превосходящей человечество (с уступающей встреча маловероятна ввиду того, что космические технологии человечества пока развиты недостаточно), обыграна во многих фантастических произведениях (например, в романе А. Кларка «Конец детства»). Итог этого контакта в большинстве случаев представляется трагическим: стремительное или

постепенное угасание человеческой цивилизации вследствие остановки научно-технического прогресса, культурного вытеснения и подавления. Подобный сценарий, скорее всего, ожидает человечество и в случае распространения сферы общественных отношений на субъектный искусственный интеллект.

Необходимо осознавать невозможность гарантированного подчинения интеллектуально более сильного субъекта менее сильному, т.е. человеку. Следовательно, искусственный интеллект как субъект не должен появиться. Это может быть обеспечено не учеными или наукой в целом, а государством, при условии прямого запрета, преследования всех, кто попытается такого рода интеллект создать. Наделение искусственного интеллекта субъектностью – опасная игра, которую человечеству не стоит начинать.

Заключение

Резюмируя изложенное, можно констатировать следующее:

1. Технологии искусственного интеллекта развиваются, и в дальнейшем их роль будет только повышаться. Без обладания этими технологиями невозможно обеспечить глобальную конкурентоспособность компании или страны.

2. Искусственный интеллект может выйти из-под контроля человека только в случае наделения его субъектностью – способностью быть субъектом активности, в том числе в отношении объектов познания.

3. Для решения интеллектуальных, в том числе творческих задач, наделять искусственный интеллект субъектностью не требуется. Творческие задачи могут быть решены с использованием подходов общей теории систем.

4. Разделение объектов (в том числе когнитивных систем) на живые и неживые выполнимо лишь формально, по механизмам формирования, но не по организационным или структурным свойствам.

5. Создание субъектного искусственного интеллекта технически видится возможным, однако не требуется для выполнения им своего функционального назначения. При этом оно порождает существенные угрозы для человечества ввиду несоответствия и неизбежного столкновения потребностей и желаний человека и субъектного искусственного интеллекта.

ЦИТИРУЕМАЯ ЛИТЕРАТУРА

Бауэр 1935 – *Бауэр Э.С.* Теоретическая биология. – М.; Л.: ВИЭМ, 1935.

Берталанфи 1969 – *Берталанфи Л. фон.* Общая теория систем: критический обзор // Исследования по общей теории систем: сб. переводов / общ. ред. и вступ. ст. В.И. Садовского, Э.Г. Юдина. – М.: Прогресс, 1969. С. 23–82.

Богданов 1989 – *Богданов А.А.* Тектология. Всеобщая организационная наука: в 2 кн. – М.: Экономика, 1989.

Гегель 1977 – *Гегель Г.В.Ф.* Энциклопедия философских наук: в 3 т. Т. 3. – М.: Мысль, 1977.

Гиренок 2010 – *Гиренок Ф.И.* Самость // Энциклопедия фонда знаний «Ломоносов». 2010. 3 декабря. – URL: http://www.lomonosov-fund.ru/enc/ru/encyclopedia: 0127733.

Гленсдорф, Пригожин 1973 – *Гленсдорф П., Пригожин И.* Термодинамическая теория структуры, устойчивости и флуктуаций / пер. с англ. Н.В. Вдовиченко, В.А. Онищука; под ред. Ю.А. Чизмаджева. – М.: Мир, 1973.

Грибков 2023 – *Грибков А.А.* Генезис и место в системе знаний представлений о добре и зле // Общество: философия, история, культура. 2023. № 8. С. 15–22.

Грибков 2024а – *Грибков А.А.* Человек в цивилизации когнитивных технологий // Философия и культура. 2024. № 1. С. 22–33.

Грибков 2024б – *Грибков А.А.* Творчество как имплементация представления о целостности мира // Философская мысль. 2024. № 3. С. 44–53.

Грибков 2024в – *Грибков А.А.* Несубъектный искусственный интеллект в системе субъект-объектных отношений // Философская мысль. 2024. № 5. С. 11–21.

Грибков 2024г – *Грибков А.А.* Эмпирико-метафизическая общая теория систем. – М.: Издательский дом Академии естествознания, 2024.

Грибков, Зеленский 2023а – *Грибков А.А., Зеленский А.А.* Общая теория систем и креативный искусственный интеллект // Философия и культура. 2023. № 11. С. 32–44.

Грибков, Зеленский 2023б – *Грибков А.А., Зеленский А.А.* Определение сознания, самосознания и субъектности в рамках информационной концепции // Философия и культура. 2023. № 12. С. 1–14.

Грибков, Зеленский 2024 – *Грибков А.А., Зеленский А.А.* Синергетика искусственных когнитивных систем с неравновесной устойчивостью // Философия и культура. 2024. № 6. С. 93–103.

Дубровский 2009 – *Дубровский Д.И.* Проблема сознания: опыт обзора основных вопросов и теоретических трудностей // Проблема сознания в философии и науке: коллективная монография / под ред. Д.И. Дубровского. – М.: Канон+, 2009. С. 5–52.

Меркулов 2004 – *Меркулов И.П.* Когнитивная система // Философия: энциклопедический словарь / под ред. А.А. Ивина. – М.: Гардарики, 2004.

Кант 1994 – *Кант И.* Собрание соч.: в 8 т. Т. 3. Критика чистого разума. – М.: Чоро, 1994.

Прыгин 2017 – *Прыгин Г.С.* Феномен сознания: является ли информационная концепция сознания прорывом в его понимании // Вестник Удмуртского университета. Серия: Философия. Психология. Педагогика. 2017. Т. 27. Вып. 4. С. 456–463.

Рапацевич 1995 – *Рапацевич Е.С.* Словарь-справочник по научно-техническому творчеству. – Минск: Этоним, 1995.

Степанский 2006 – *Степанский В.И.* Психоинформация. Теория. Эксперимен. – М.: Московский психолого-социальный институт, 2006.

Цветков 2021 – *Цветков В.Я.* Информационная синергетика // Образовательные ресурсы и технологии. 2021. № 2. С. 72–78.

Уемов 1978 – *Уемов А.И.* Системный подход и общая теория систем. – М.: Мысль, 1978.

Toyota et al. 2018 – *Toyota M., Spencer D., Sawai-Toyota S., Jiaqi W., Zhang T., Koo A.J., Howe G.A., Gilroy S.* Glutamate Triggers Long-Distance, Calcium-Based Plant Defense Signaling // Science. 2018. Vol. 361. No. 6407. P. 1112–1115.

Herculano-Houzel 2009 – *Herculano-Houzel S.* The Human Brain in Numbers: A Linearly Scaled-Up Primate Brain // Frontiers in Human Neuroscience. 2009. Vol. 3. Article 31.

Pross, Pascal 2013 – *Pross A., Pascal R.* The Origin of Life: What We know, What We Can Know and What We Will Never Know // Open Biology. 2013. Vol. 3. No. 3. Article 120190.

Pascal, Pross, Sutherland 2013 – *Pascal R., Pross A., Sutherland J.D.* Towards an Evolutionary Theory of the Origin of Life Based on Kinetics and Thermodynamics // Open Biology. 2013. Vol. 3. No. 11. Article 130156.

REFERENCES

Bauer E.S. (1935) *Theoretical Biology.* Moscow: All-Union Institute of Experimental Medicine (in Russian).

Bertalanffy L. von (1969) General System Theory: A Critical Review. In: Sadovskiy V.I. & Yudin E.G. (Eds.) *Studies in General Systems Theory: A Collection of Translations* (pp. 23–82). Moscow: Progress (Russian translation).

Bogdanov A.A. (1989) *Tektology. The Universal Organizational Science* (Vols. 1–2). Moscow: Ekonomika (in Russian).

Dubrovsky D.I. (2009) The Problem of Consciousness: An Attempt to Review the Main Issues and Theoretical Difficulties. In: Dubrovsky D.I. (Ed.) *The Problem of Consciousness in Philosophy and Science* (pp. 5–52). Moscow: Kanon+ (in Russian).

Girenok F.I. (2010) Selfhood. In: *Encyclopedia of "Lomonosov" Knowledge Foundation.* Retrieved from http://www.lomonosov-fund.ru/enc/ru/encyclopedia:0127733 (in Russian).

Glansdorff P. & Prigogine I. (1973) *Thermodynamic Theory of Structure, Stability, and Fluctuations* (N.V. Vdovichenko & V.A. Onishchuk, Trans.). Moscow: Mir (Russian translation).

Gribkov A.A. (2023) Genesis and Place in the System of Knowledge of Ideas about Good and Evil. *Obshchestvo: filosofiya, istoriya, kul'tura.* No. 8, pp. 15–22 (in Russian).

Gribkov A.A. (2024a) Man in the Civilization of Cognitive Technologies. *Filosofiya i kul'tura.* No. 1, pp. 22–33 (in Russian).

Gribkov A.A. (2024b) Creativity as an Implementation of the Idea of the Integrity of the World. *Filosofskaya mysl'.* No. 3, pp. 44–53 (in Russian).

Gribkov A.A. (2024c) Non-Subjective Artificial Intelligence in the System of Subject-Object Relations. *Filosofskaya mysl'.* No. 5, pp. 11–21 (in Russian).

Gribkov A.A. (2024d) *Empirico-Metaphysical General System Theory.* Moscow: Publishing House of the Academy of Natural Science (in Russian).

Gribkov A.A. & Zelenskiy A.A. (2023a) General System Theory and Creative Artificial Intelligence. *Filosofiya i kul'tura.* No. 11, pp. 32–44 (in Russian).

Gribkov A.A. & Zelenskiy A.A. (2023b) Definition of Consciousness, Self-Consciousness, and Subjectivity within the Information Concept. *Filosofiya i kul'tura.* No. 12, pp. 1–14 (in Russian).

Gribkov A.A. & Zelenskiy A.A. (2024) Synergetics of Artificial Cognitive Systems with Non-Equilibrium Stability. *Filosofiya i kul'tura.* No. 6, pp. 93–103 (in Russian).

Hegel G.W.F. (1977) *Encyclopaedia of the Philosophical Sciences in 3 Vols.* (Vol. 3). Moscow: Mysl' (Russian translation).

Herculano-Houzel S. (2009) The Human Brain in Numbers: A Linearly Scaled-Up Primate Brain. *Frontiers in Human Neuroscience.* Vol. 3, article 31.

Kant I. (1994) *Collected Works in 8 Vols. Vol. 3: Critique of Pure Reason.* Moscow: Choro (Russian translation).

Merkulov I.P. (2004) Cognitive System. In: Ivin A.A. (Ed.) *Philosophy: An Encyclopedic Dictionary.* Moscow: Gardariki (in Russian).

Pascal R., Pross A., & Sutherland J.D. (2013) Towards an Evolutionary Theory of the Origin of Life Based on Kinetics and Thermodynamics. *Open Biology.* Vol. 3, no. 11, article 130156.

Pross A. & Pascal R. (2013) The Origin of Life: What We Know, What We Can Know and What We Will Never Know. *Open Biology.* Vol. 3, no. 3, article 120190.

Prygin G.S. (2017) The Phenomenon of Consciousness: Is the Information Concept of Consciousness a Breakthrough in its Understanding? *Bulletin of Udmurt University. Series Philosophy. Psychology. Pedagogy.* Vol. 27, no. 4, pp. 456–463 (in Russian).

Rapatsevich E.S. (1995) *Dictionary-Reference on Scientific and Technical Creativity.* Minsk: Etonim (in Russian).

Stepanskiy V.I. (2006) *Psycho-Information. Theory. Experiment.* Moscow: Moscow Psychological-Social Institute (in Russian).

Toyota M., Spencer D., Sawai-Toyota S., Jiaqi W., Zhang T., Koo A.J., Howe G.A., & Gilroy S. (2018) Glutamate Triggers Long-Distance, Calcium-Based Plant Defense Signaling. *Science.* Vol. 361, no. 6407, pp. 1112–1115.

Tsvetkov V.Ya. (2021) *Information Synergetics. Obrazovatel'nye resursy i tekhnologii.* No. 2, pp. 72–78 (in Russian).

Uemov A.I. (1978) *The Systems Approach and General System Theory.* Moscow: Mysl' (in Russian).