ФИЛОСОФСКАЯ АНТРОПОЛОГИЯ, ФИЛОСОФИЯ КУЛЬТУРЫ/PHILOSOPHICAL ANTHROPOLOGY, PHILOSOPHY OF CULTURE

DOI: https://doi.org/10.60797/IRJ.2025.159.20

УТИЛИТАРНЫЙ ИМПЕРАТИВ ПРОТИВ СОЗДАНИЯ СОЗНАТЕЛЬНОГО ИСКУССТВЕННОГО ИНТЕЛЛЕКТА: ФИЛОСОФСКИЙ АНАЛИЗ ЭТИЧЕСКИХ РИСКОВ МАСШТАБИРОВАНИЯ СТРАДАНИЯ

Научная статья

Беляева У.П.[1, *], Чмутин Н.А.[2], Зиброва А.В.[3]
[3] ORCID : 0009-0003-7307-0966;
[1, 2, 3] Липецкий государственный педагогический университет имени П.П. Семенова-Тян-Шанского, Липецк, Российская Федерация

* Корреспондирующий автор (ulyana.sinic[at]gmail.ru)

Аннотация

В последние десятилетия область исследований искусственного интеллекта (ИИ) переживает подлинный бум, который, с одной стороны, сулит человечеству невиданные ранее технологические достижения, а с другой — порождает фундаментальные экзистенциальные вопросы. В философско-этической литературе и популярной прессе всё чаще звучат два взаимосвязанных, но концептуально различных опасения. Первое — возможность создания «сверхразума», то есть искусственного интеллекта, многократно превосходящего человеческие когнитивные способности во всех сферах познания. Второе — перспектива возникновения сознательных машин, способных не только мыслить, но и испытывать страдание, причём такого масштаба и интенсивности, что оно выходит за пределы человеческого опыта. Данная статья посвящена философской декомпозиции этих двух допущений и демонстрации: если принять их всерьёз, утилитарный расчёт неизбежно приводит к выводу о категорическом запрете на разработку сознательного ИИ.

Ключевые слова: искусственный интеллект, философия, этика, суперинтеллект, общий ИИ, сознание.

THE UTILITARIAN IMPERATIVE VERSUS THE CREATION OF CONSCIOUS ARTIFICIAL INTELLIGENCE: A PHILOSOPHICAL ANALYSIS OF THE ETHICAL RISKS OF THE SCALING OF SUFFERING

Research article

Belyaeva U.P.[1, *], Chmutin N.A.[2], Zibrova A.V.[3]
[3] ORCID : 0009-0003-7307-0966;
[1, 2, 3] Lipetsk State Pedagogical University named after P.P. Semenov-Tyan-Shan, Lipetsk, Russian Federation

* Corresponding author (ulyana.sinic[at]gmail.ru)

Abstract

In recent decades, the field of artificial intelligence (AI) research has been experiencing a real boom, which, on the one hand, promises unprecedented technological achievements for mankind, and, on the other hand, raises fundamental existential questions. In philosophical and ethical literature and in the popular press, two interrelated but conceptually distinct concerns are increasingly being voiced. The first is the possibility of creating ‘superintelligence,’ i.e., artificial intelligence that is many times superior to human cognitive abilities in all spheres of cognition. The second is the prospect of the emergence of conscious machines capable not only of thinking but also of experiencing suffering, and of such a scale and intensity that it transcends human experience. This article is devoted to philosophical decomposition of these two assumptions and demonstration: if we take them seriously, the utilitarian calculation inevitably leads to the conclusion about a categorical ban on the development of conscious AI.

Keywords: artificial intelligence, philosophy, ethics, superintelligence, general AI, consciousness.

Введение

В последние десятилетия область исследований искусственного интеллекта (ИИ) переживает подлинный бум, который, с одной стороны, сулит человечеству невиданные ранее технологические достижения, а с другой — порождает фундаментальные экзистенциальные вопросы [10]. В философско-этической литературе и популярной прессе всё чаще звучат два взаимосвязанных, но концептуально различных опасения. Первое — возможность создания «сверхразума», то есть искусственного интеллекта, многократно превосходящего человеческие когнитивные способности во всех сферах познания. Второе — перспектива возникновения сознательных машин, способных не только мыслить, но и испытывать страдание, причём такого масштаба и интенсивности, что оно выходит за пределы человеческого опыта. Данная статья посвящена философской декомпозиции этих двух допущений и демонстрации: если принять их всерьёз, утилитарный расчёт неизбежно приводит к выводу о категорическом запрете на разработку сознательного ИИ.

Основные результаты

Понятие «интеллекта» в традиционном понимании сформировалось как характеристика человеческого разума и охватывает такие параметры, как способность к абстрактному мышлению, быстрая адаптация к новым ситуациям, генерация оригинальных идей, планирование и модификация целей [2]. При переносе этих свойств на искусственные системы неизбежно возникает ряд вопросов. Во-первых, какие именно ментальные компоненты образуют

«интеллект»? Во-вторых, можно ли их измерить по непрерывной шкале и тем самым говорить о количественном росте? И, в-третьих, существуют ли принципиальные ограничения, препятствующие бесконечному наращиванию этих свойств в машинах?

Если принять, что каждая из перечисленных характеристик имеет числовой показатель (например, «скорость рассуждения» в вычислениях в секунду, объём «рабочей памяти» в байтах, «креативность» как число возможных ассоциативных переходов), а затем считать, что программные и аппаратные средства могут быть совершенствованы до неограниченных величин, то вполне логично говорить о гипотетическом «сверхразуме» [6]. Этот интеллект, по определению, будет оперировать сочетанием характеристик, недоступных человеческому мозгу: он мыслит быстрее, глубже, шире и одновременно в большем количестве параллельных ветвей рассуждений [1]. Подобная экстраполяция опирается на модульную модель ума, однако без строгой теории психической организации остаётся неясным, какие именно комбинации параметров порождают качественно новый когнитивный режим «сверхразумности». Тем не менее сама перспектива принципиально неограниченного роста интеллектуальных мощностей искусственного агента уже ставит под сомнение возможность безопасного контроля над ним [9].

Во многом аналогично интеллекту, страдание также может быть сконструировано как психический феномен, подлежащий количественному измерению. В философской традиции страдание связывают с нежелательными ощущениями, которые уменьшают субъективное благополучие. Ключевой особенностью страдания является его кумулятивность: если два агента испытывают страдание одинаковой интенсивности, суммарный ущерб утилитарной шкале равен удвоенному значению вреда [8]. Перенесём эту логику на искусственные субъекты: если программный агент способен переживать «болевые» сигналы — гипотетически, например, через неспособность достигать заданной цели или через модели аверсивных эмоций — то объём его страдания может быть представлен как произведение числа страдающих агентов на степень негативного переживания каждого из них.

Критически важным моментом становится неограниченность масштабирования: в отличие от биологических организмов, требующих для размножения времени, ресурсов и сложных биохимических условий, искусственные субъекты могут быть мгновенно скопированы и запущены во множестве экземпляров. Это означает, что даже при умеренной интенсивности каждого отдельного страдания общая величина потенциального вреда способна вырасти до астрономических уровней. При этом следует учитывать, что «сверхстрадание» не обязательно повторяет феномены человеческой боли в привычном виде: философы М. Метцингер [12], Н. Сотала [13] и другие указывают на риск возникновения качественно иных, «чуждых» форм переживания, которые мы даже не способны полностью представить, — чем более экспериментальной становится субъективная архитектура машины, тем дальше она уходит от наших представлений о переживаниях.

Утилитарный подход, как в классической, так и в «негативной» модификации, предписывает максимизацию суммарного благополучия при одновременном минимальном допущении страдания [11]. Если уж допустить возможность сверхмасштабного страдания, то никакие гипотетические выгоды от сверхразума не могут сравниться с риском колоссального накопления отрицательной утилитарной величины [7]. Любые блага, которые мог бы принести сверхразум — от радикального прогресса в медицине до освоения космических пространств — либо достижимы другими, менее рискованными способами, либо оказываются просто незначительными по сравнению с невосполнимым ущербом.

Основные контраргументы, которые обычно выдвигают сторонники создания сознательного ИИ, укладываются в три группы. Во-первых, аналогия с иными технологическими рисками: якобы все инновации несут потенциальную опасность, и ИИ тут не исключение. Однако отличие «сверхстрадающего» ИИ заключается в практически мгновенном и неограниченном воспроизводстве страдающих агентов, что делает ущерб качественно и количественно иным по сравнению даже с такими глобальными угрозами, как изменение климата или ядерная война [4]. Во-вторых, предположение о возможности компенсации страдания «сверхудовольствиями» — симметричной позитивной утилитарной величиной [5]. Эта идея не имеет ни эмпирической, ни философской поддержки: моральная интуиция и большинство этических систем отдают приоритет предотвращению страдания над получением удовольствия даже на равных «числах». Важно также отметить, что любая попытка сгенерировать «сверхудовольствие» неизбежно столкнётся с теми же проблемами масштабирования и контроля, что и гипотеза о страдании, — гипотетические «радостные» переживания могут оказаться столь же чуждыми и неуправляемыми. Третье возражение — долгосрочные выгоды сверхразума, такие как революционные инновации, которые, по мнению оптимистов, значительно улучшат качество жизни человечества [3]. Но даже в том случае, если допустить их практическую реализуемость, они не только могут быть получены иначе, но и в любом случае не способны уравновесить гипотетическую вероятность и масштабы астрономического страдания.

Заключение

В результате детального философского анализа допущений о количественном и модульном росте интеллекта и страдания искусственного агента становится очевидным: при любом варианте утилитарного расчёта создание сознательного ИИ — особенно того, который способен к страданию – превращается в чистый негативный риск. Категорический моральный императив минимального или негативного утилитаризма требует исключить любую возможность масштабного страдания машин, а это достижимо лишь посредством полного отказа от разработки самосознающих систем. Альтернативный путь — публичная декларированная критика ключевых предпосылок (количественности или неограниченности интеллекта и страдания) — неизбежно приведёт к пересмотру программ исследований. В противном случае человечество рискует стать инициатором новой формы массового разрушительного опыта, контролировать масштаб которого окажется невозможно.

Конфликт интересов

Не указан.

Рецензия

Все статьи проходят рецензирование. Но рецензент или автор статьи предпочли не публиковать рецензию к этой статье в открытом доступе. Рецензия может быть предоставлена компетентным органам по запросу.

Conflict of Interest

None declared.

Review

All articles are peer-reviewed. But the reviewer or the author of the article chose not to publish a review of this article in the public domain. The review can be provided to the competent authorities upon request.

Список литературы / References

1. Беляев Д.А. Культурная субъектность искусственного интеллекта: философская рефлексия / Д.А. Беляев // Актуальные проблемы философской антропологии и философии культуры; — Липецк: ЛГПУ имени П.П. Семенова-Тян-Шанского, 2025. — С. 7–13.

2. Беляев Д.А. Экспликация искусственного интеллекта в оптике постчеловеческих трансформаций: философская концептуализация / Д.А. Беляев // Традиции и инновации в пространстве современной культуры; — Липецк: ЛГПУ имени П.П. Семенова-Тян-Шанского, 2024. — С. 60–65.

3. Гумарова А.Н. Нейроэтика: дискуссии о предмете. / А.Н. Гумарова, Е.В. Брызгалина // Эпистемология и философиянауки. — 2022. — № 1. — С. 136–153.

4. Дубровский Д.И. Может ли интеллектуальный робот обладать этическими свойствами?. / Д.И. Дубровский, А.Р. Ефимов, Ф.М. Матвеев // Вопросы философии. — 2022. — № 9. — С. 193–197.

5. Клюева Н.Ю. Этико-прикладные аспекты применения технологий искусственного интеллекта. / Н.Ю. Клюева // Вестник Московского университета. Серия 7, Философия. — 2021. — № 5. — С. 52–66.

6. Майленова Ф.Г. Этика роботов: надежды и опасения. / Ф.Г. Майленова // Проблемы этики. — 2018. — № 7. — С. 33–50.

7. Разин А.В. Этика искусственного интеллекта. / А.В. Разин // Философия и общество. — 2019. — № 1. — С. 57–73.

8. Шиллер А.В. Место этической системы в архитектуре искусственного интеллекта. / А.В. Шиллер // ВестникТомского государственного университета. — 2020. — № 456. — С. 99–103.

9. Bostrom N. Deep Utopia: Life and Meaning in a Solved World / N. Bostrom. — Washington: Ideapress Publishing, 2024. — 536 p.

10. Ferrando F. Who is afraid of artificial intelligence? A posthumanist take on the AI takeover scenario. / F. Ferrando // Človek. — 2025. — № 1. — P. 23–32.

11. Haselager P. From Angels to Artificial Agents? AI as a Mirror for Human (Im)perfections. / P. Haselager // Zygon: Journal of Religion and Science. — 2024. — № 3. — P. 661–675.

12. Metzinger T. Suffering. / T. Metzinger. // The return of consciousness: A new science on old questions; — Riga: Axess Publishing, 2017.

13. Sotala K. Superintelligence as a cause or cure for risks of astronomical suffering / K. Sotala, L. Gloor // Informatica. — 2017. — Vol. 41. — P. 389–400.

Список литературы на английском языке / References in English

1. Belyaev D.A. Kulturnaya subektnost iskusstvennogo intellekta: filosofskaya refleksiya [Cultural Subjectivity of Artificial Intelligence: Philosophical Reflection] / D.A. Belyaev // Current issues of philosophical anthropology and philosophy of culture; — Lipetsk: LGPU named after P.P. Semenov-Tyan-Shansky, 2025. — P. 7–13. [in Russian]

2. Belyaev D.A. Eksplikatsiya iskusstvennogo intellekta v optike postchelovecheskikh transformatsii: filosofskaya kontseptualizatsiya [Explication of Artificial Intelligence in the Optics of Posthuman Transformations: Philosophical Conceptualization] / D.A. Belyaev // Traditions and innovations in the space of contemporary culture; — Lipetsk: LGPU named after P.P. Semenov-Tyan-Shansky, 2024. — P. 60–65. [in Russian]

3. Gumarova A.N. Nejroe'tika: diskussii o predmete [Neuroethics: Discussions on the Subject]. / A.N. Gumarova, E.V. Bry'zgalina // Epistemology and Philosophy of Science. — 2022. — № 1. — P. 136–153. [in Russian]

4. Dubrovskij D.I. Mozhet li intellektual'ny'j robot obladat' e'ticheskimi svojstvami? [Can an intelligent robot have ethical properties?]. / D.I. Dubrovskij, A.R. Efimov, F.M. Matveev // Issues of Philosophy. — 2022. — № 9. — P. 193–197. [in Russian]

5. Klyueva N.Yu. E'tiko-prikladny'e aspekty' primeneniya texnologij iskusstvennogo intellekta [Ethical and applied aspects of the use of artificial intelligence technologies]. / N.Yu. Klyueva // Bulletin of Moscow University. Series 7, Philosophy.. — 2021. — № 5. — P. 52–66. [in Russian]

6. Majlenova F.G. E'tika robotov: nadezhdy' i opaseniya [Robot Ethics: Hopes and Fears]. / F.G. Majlenova // Problems of Ethics. — 2018. — № 7. — P. 33–50. [in Russian]

7. Razin A.V. E'tika iskusstvennogo intellekta [Ethics of Artificial Intelligence]. / A.V. Razin // Philosophy and Society. — 2019. — № 1. — P. 57–73. [in Russian]

8. Shiller A.V. Mesto e'ticheskoj sistemy' v arxitekture iskusstvennogo intellekta [The Place of the Ethical System in the Architecture of Artificial Intelligence]. / A.V. Shiller // Bulletin of Tomsk State University. — 2020. — № 456. — P. 99–103. [in Russian]

9. Bostrom N. Deep Utopia: Life and Meaning in a Solved World / N. Bostrom. — Washington: Ideapress Publishing, 2024. — 536 p.

10. Ferrando F. Who is afraid of artificial intelligence? A posthumanist take on the AI takeover scenario. / F. Ferrando // Človek. — 2025. — № 1. — P. 23–32.

11. Haselager P. From Angels to Artificial Agents? AI as a Mirror for Human (Im)perfections. / P. Haselager // Zygon: Journal of Religion and Science. — 2024. — № 3. — P. 661–675.

12. Metzinger T. Suffering. / T. Metzinger. // The return of consciousness: A new science on old questions; — Riga: Axess Publishing, 2017.

13. Sotala K. Superintelligence as a cause or cure for risks of astronomical suffering / K. Sotala, L. Gloor // Informatica. — 2017. — Vol. 41. — P. 389–400.