УДК 1(091):0048:17

ЦИФРОВЫЕ ТЕХНОЛОГИИ: РАЗМЫШЛЕНИЯ О РАЗЛИЧИИ МЕЖДУ ИНСТРУМЕНТАЛЬНОЙ РАЦИОНАЛЬНОСТЬЮ И ПРАКТИЧЕСКИМ РАЗУМОМ

Л. Нагль[1]

Находятся ли компьютеры на пути к обретению «суперинтеллекта»? Способно ли механическое выполнение программ искусственного интеллекта полностью имитировать размышление и принятие решений в том виде, в каком это происходит у человека? При ближайшем рассмотрении эти вопросы кажутся необоснованными, так как алгоритмы (или, в кантовской терминологии, «императивы умения», реализуемые техническими методами) имеют, по сути, гетерономные характеристики. Так называемая автономность искусственного интеллекта – это автоматизм исполнения на основе полученных от датчиков данных. В сравнении с потенциалом человеческого «практического разума» к этическому суждению такой автоматизм оказывается ограниченным в существенных аспектах, несмотря на способность цифровых технологий к вероятностной адаптации к новым данным при помощи «машинного обучения». Это утверждение подробно разбирается в свете идеи «цифрового гуманизма», предложенной Ю. Нидой-Рюмелином и Н. Вайденфельд. Данная концепция признает возможную полезность алгоритмов как «инструментов», при этом подчеркивая (и тем самым отвергая крайние проявления связываемых с искусственным интеллектом утопических и антиутопических идей «постгуманизма») существенную разницу между человеческим действием и его (частичной) имитацией искусственным интеллектом. С одной стороны, «цифровой гуманизм» Ниды-Рюмелина и Вайденфельд базируется на кантианском представлении об автономности человека как самоопределяющегося существа. С другой – обосновываемое ими понятие «структурной рациональности» весьма проблематично. Делается вывод, что концепция «цифрового гуманизма» может быть усо-

[1] Институт философии, Венский университет. Австрия, 1010, Вена, Универзитетсштрассе, д. 7.
Поступила в редакцию: 04.10.2021 г.
doi: 10.5922/0207-6918-2022-1-3

DIGITAL TECHNOLOGY: REFLECTIONS ON THE DIFFERENCE BETWEEN INSTRUMENTAL RATIONALITY AND PRACTICAL REASON

L. Nagl[1]

Are computers on the way to acquiring "superintelligence"? Can human deliberation and decision-making be fully simulated by the mechanical execution of AI programmes? On close examination these expectations turn out not to be well-founded, since algorithms (or, in Kantian terms, "imperatives of skill" that are implemented by technological means) do, ultimately, have "heteronomous" characteristics. So-called AI-"autonomy" is a sensor-directed performance automatism, which – compared with the potential for ethical judgment in human "practical reason" – proves to be limited in significant ways (even if, in so-called "machine learning", digital technologies are able to probabilistically adapt to new data). This is shown in some detail with reference to the idea of a "digital humanism", which was introduced by Julian Nida-Rümelin and Nathalie Weidenfeld, who argue that algorithms (possibly) are useful "tools", but emphasise – thus rejecting excessive "post-humanist" (Utopian or dystopian) ideas about AI – that there exists a crucial difference between human action and its (partial) AI-simulation. While Nida-Rümelin/Weidenfeld´s "digital humanism" is, on the one hand, inspired by Kant's conception of human autonomous self-determination, the concept of "structural rationality" that they advocate is, on the other hand, quite problematic. "Digital humanism", however, can be improved

[1] Institute of Philosophy, University of Vienna. Universitätsstraße 7, Vienna, 1010, Austria.
Received: 04.10.2021.
doi: 10.5922/0207-6918-2022-1-3

вершенствована с опорой на предложенный Б. Херман анализ «морального суждения» и рефлексию А. Вуда о «человеческом достоинстве».

Ключевые слова: *искусственный интеллект, рациональность, практический разум, способность суждения, цифровой гуманизм, Ю. Нида-Рюмелин, Н. Вайденфельд, Б. Херман, А. Вуд*

Введение

Используемые машинами алгоритмы имеют множество полезных применений. Они механическим образом выполняют задачи *подсегмента* праксиса, а именно «императивов умения» («инструментальной» рациональности в терминах М. Хоркхаймера и Т. Адорно). Тем не менее этот подсегмент, согласно кантовскому многогранному анализу практического разума, не является *праксисом в полном смысле слова*. Следовательно, чтобы подчеркнуть операционную природу алгоритмических процессов и провести различие между ними и человеческим *логосом*, искусственный интеллект (ИИ) стоит определять в терминах рациональности, а не разума. Алгоритмическая рациональность неотделима от ее инструментальной природы, в то время как человеческий разум ("Vernunft im eigentlichen Sinn"[2]) выражается в «человеческом суждении, рассудке и суждении» (Mersch, 2019, S. 853).

В настоящей статье пять разделов. В первом кратко рассматриваются «постгуманистические» фантазии, появление которых сопутствовало развитию ИИ (например, громкие заявления о *грядущем «суперинтеллекте»* самоуправляемых компьютеров и возникшие в пику им апокалиптические пророчества). Второй раздел посвящен постаналитическому и герменевтическому анализу достоинств и ограничений «цифровой рациональности». В третьем разделе разбирается предложенная Юлианом Нидой-Рюмелином и Натали Вайденфельд идея «цифрового гуманизма», во многом восходящая к кантовской концепции практического

[2] «Разум в собственном смысле» (нем.).

as I argue – with reference to Barbara Herman's analysis of "moral judgment" and to Allen Wood's reflections on "human dignity".

Keywords: *Artificial Intelligence, rationality, practical reason, power of judgment, digital humanism, Julian Nida-Rümelin, Nathalie Weidenfeld, Barbara Herman, Allen Wood*

Introduction

Algorithms instantiated by machines prove useful in many regards. They mechanically execute a *sub-segment* of praxis: "imperatives of skill" (or "instrumental" rationality in Horkheimer and Adorno's terms). This sub-segment is, however, – as Kant shows in his complex analyses of practical reason – not at all *praxis in its full sense*. It is thus of the utmost importance to define AI in terms of "rationality" rather than "reason", in order to make visible the operative character of algorithmic processes and to distinguish them from human *logos*. While algorithmic rationality is chained to its instrumental character, human reason ("*Vernunft im eigentlichen Sinn*") is manifest in "human judgment, understanding, and justification" (Mersch, 2019, p. 853).

My essay has *five* parts: after a brief look, in part *one*, at some "post-humanistic" fantasies which accompany the rise of AI (e.g. excessive claims about that *the upcoming "super-intelligence"* of self-steering computers, and their apocalyptic inversions), part *two* focuses on post-analytic and hermeneutic analyses of the merits and the limits of "digital rationality". Part *three* presents Julian Nida-Rümelin and Nathalie Weidenfeld's idea of a "Digital Humanism", which is inspired (though only in some regards) by Kant's conception of practical reason, arguing that algorithms, as ensembles of tools, are in permanent need of human, socio-political control. Part *four* asserts that

разума. Нида-Рюмелин и Вайденфельд подчеркивают, что алгоритмы, будучи совокупностями инструментов, постоянно нуждаются в человеческом, общественно-политическом контроле. В четвертом разделе утверждается, что основные этические идеи «цифрового гуманизма» находят подкрепление в кантианском анализе в «Практике морального суждения» Барбары Херман (Herman, 1993) и в рефлексии Аллена Вуда по поводу «Человечества как цели самой по себе» (Wood, 1998). В пятой, заключительной части делаются выводы. Утверждается, что разнообразные неверные прочтения «автономии» и «подлинности» могут обсуждаться в нашу «цифровую эпоху» (как это недавно отметили К. Беккер и С. Зойберт) в рамках вдохновленной Хоркхаймером и Адорно критической теории, переносящей в современный контекст ключевые элементы кантовской концепции практического разума.

1. За пределами «постгуманистических» фантазий и апокалиптических прогнозов: трезвый философский взгляд на искусственный интеллект

Тщательный анализ реальности и истинного потенциала цифровых технологий не может опираться ни на *утопическую журналистику*, ни на ее *антиутопический аналог*. Постгуманисты преувеличивают возможности искусственного интеллекта, фантазируя о *грядущем* суперинтеллекте саморегулируемых машин, который приведет к технологической сингулярности (Kurzweil, 2005). Приверженцы таких взглядов воспроизводят довольно старомодный и мейнстримный «аналитический» догматический сциентизм, далекий от вдумчивого философского анализа сложного, социально структурированного, недетерминистического горизонта действий, в котором существуют инновации в сфере ИИ (см.: Habermas, 2019, S. 593, Anm. 2). *Обратная сторона* постгуманистических утопий — безудержные *антиутопические*

the leading ethical ideas of a "Digital Humanism" can be further solidified with reference to Barbara Herman's Kant-based analysis of *The Practice of Moral Judgment*, as well as to Allen Wood's reflections on *Humanity as End in itself*. Part *five,* the *coda and conclusion* of the presentation, asserts that the manifold misreadings of "autonomy" and "authenticity" in our "digital age" can be discussed (as Becker and Seubert have recently pointed out) in a (Horkheimer- and Adorno-inspired) Critical Theory that re-articulates in contemporary contexts core motifs of Kant's concept of practical reason.

1. Beyond "Post-Humanistic" Fantasies and Apocalyptic Prognoses: A Sober, Philosophical Look at Artificial Intelligence

For a careful analysis of the reality as well as the real potentials of digital technologies, neither today's *Utopian journalism* nor its *dystopian counterpart* are of much help. Post-humanists exaggerate the possibilities of Artificial Intelligence by fantasising about *the upcoming* "super-intelligence" of self-regulating computers that will lead to their "singularity" (Kurzweil, 2005). They thus re-enact an (actually quite old-fashioned, mainstream "analytical") dogmatic scientism that avoids any careful, philosophical analysis of the complex, socially structured, non-deterministic action horizons in which AI innovations are embedded (see Habermas, 2019, p. 593n2). The *negative inversion* of these post-humanistic "super-intelligence" Utopias are spelled out in equally excessive *dystopian* tales that predict the advent of an unavoidably apocalyptic development of AI (Bostrom, 2014). Both fantasies have their basis in an *abstract generalisation* of "instrumen-

фантазии, предрекающие неминуемое апокалиптическое развитие ИИ (Bostrom, 2014). Оба нарратива укоренены в *абстрактном обобщении* «инструментальной рациональности», а именно в искаженном представлении о праксисе, уделяющем чрезмерное внимание причинно-следственному отношению между средствами и целью, то есть в представлении, подлежащем, как это будет показано ниже, критике в свете кантовской многосоставной концепции «практического разума».

Трезвая оценка достоинств и структурных ограничений технологий ИИ, в отличие от утопических и антиутопических нарративов об ИИ, во-первых, должна отражать неоспоримые преимущества цифровизации (то есть увеличивающуюся *скорость*, экспоненциально возрастающий *охват* и *потенциал к адаптации* основанных на алгоритмах программ, во многих отношениях превосходящих предшествующие ИИ человеческие технологии). Во-вторых, необходимо эффективное *общественно-политическое регулирование* новых цифровых «инструментов». Австриец Зепп Хохрайтер, профессор машинного обучения, высказался за подобный трезвый взгляд на ИИ в недавнем интервью, озаглавленном «Алгоритм может научиться чему угодно — хорошему и плохому»: «Я не страшусь того, что мы придем к вселенной "Терминатора" или "Матрицы". Меня больше волнует, что технологиями ИИ будут злоупотреблять крупнейшие корпорации, правительства и организованная преступность, чтобы незаметно манипулировать людьми, управлять их действиями» (Hochreiter, 2020, p. 374—375).

Согласно Хохрайтеру, *технологическое ядро* управляемых алгоритмами программ ИИ (то есть инструментальная рациональность средств и целей, управляющая функционированием ИИ и автоисправлением) следует рассматривать как то, чем оно на самом деле и является: не как *высшую точку* развития разума, через которую алгоритмы и их самосовершенствующаяся «петля обучения» приведут нас к (постгуманитал rationality": in a distorted image of praxis which over-emphasises the causal relationship between means and end, i. e. in an image that can be criticised, as will be shown in the following, with reference to Kant's complex concept of "practical reason".

A sober assessment of the merits, as well as the structural limits, of AI-based technologies has — unlike the Utopian and the dystopian AI narratives — first to reflect on the indubitable benefits of digitalisation (i. e. on the gains in *speed*, as well as on the exponentially increasing *reach* and the *adaptation-potentials* of algorithm-guided programmes, which in many regards quite stunningly outperform pre-AI human technologies). After this, however, it has secondly to focus carefully on the need for an efficient *socio-political regulation* of these new digital "tools". The Austrian Sepp Hochreiter, a professor of Machine Leaning, supported this sober view of AI recently in his interview "The algorithm can learn anything — good things as well as bad things" in the following words: "I am not afraid that we'll move in the direction of a 'Terminator' or 'Matrix' scenario," Hochreiter (2020, pp. 374-375) said: "I am much more worried that [AI] technologies will be abused by major corporations, governments, or criminal organisations to manipulate the population, steer them in a particular direction, without us ever knowing or noticing."

In Hochreiter's considerations the *technological core* of algorithm-directed AI programmes (i. e. the instrumental means-end rationality which organises AI-directed performances and auto-corrections) is seen as what it actually is: *not as the culmination point* of reason where algorithms lead us (in a kind of self-improving "learning loop") towards a (post-human) "super-intelligence", but as a mere *sub-species* of *ratio*, i. e. as "purposive rationality" (in Max Weber's, and Horkheimer

стическому) суперинтеллекту, но как *подвид рацио*, то есть «целевую рациональность» (в понимании М. Вебера, а также Хоркхаймера и Адорно), или, используя кантовскую терминологию, (автоматическую, механическую) реализацию гипотетических императивов умения (AA 04, S. 415; Кант, 1997, с. 125). Эти императивы, ориентированные на отношения средств и целей, систематически игнорируют, согласно Канту, (ключевой *этический*) вопрос о том, являются ли их цели добром или злом. Что касается ориентации на цели, проистекающей из императива умения, можно привести известные слова Канта из «Основоположения к метафизике нравов»: «...добра ли цель, — об этом здесь совсем нет вопроса, но только о том, что необходимо делать, чтобы ее достигнуть» (AA 04, S. 415; Кант, 1997, с. 125). Логика этих процессов фокусируется исключительно на причинно-следственной связи между средствами и целями и, следовательно, избегает обращения к основному моральному вопросу о том, добра или зла избранная цель. Кант приводит известный пример (ограниченного характера) инструментальной рациональности: «Предписания врачу, чтобы он мог основательным образом вылечить своего пациента, и составителю ядов, чтобы вернее его убить, равноценны постольку, поскольку каждое служит для совершенного исполнения своего намерения» (AA 04, S. 415; Кант, 1997, с. 125)[3]. Целевой рациональный *подсегмент* человеческого *рацио* (наших *технических знаний*, которые можно расширять и совершенствовать, о средствах, необходимых для достижения старых или новых поставленных целей) остается (не только в бытовой, но и в алгоритмически запрограммированной машинно-генерируемой форме) постоянно, но не всегда в явном виде, встроенным в *многосоставную*, разнообразную и этически нагруженную среду *практического разума*. Дело не

[3] Кантовский пример находит развитие у Хохрайтера: «Химическим препаратом можно отравить, а можно и вылечить. Решающий фактор — человек, и человек способен использовать системы ИИ как во зло, так и во благо» (Hochreiter, 2020, p. 370).

and Adorno's sense); or, to use Kant's terminology, as an (automated, mechanical) realisation of hypothetical "imperatives of skill" (*GMS*, AA 04, p. 415; Kant, 1996, p. 68). Those imperatives, since they are exclusively focused on means-end relations, systematically exclude, as Kant points out, the (core *ethical*) question whether the ends that they implement are good or evil. In a goal-orientation that follows the logic of an imperative of skill, Kant famously argues in his *Groundwork of the Metaphysics of Morals*, "[w]hether the end is […] good is not at all the question here, but only what one must do in order to attain it" (*GMS*, AA 04, p. 415; Kant, 1996, p. 68). The logic of these processes is focused exclusively on the causal connection between means and end, and hence avoids the core moral question which concerns the goodness or evilness of the desired end. Kant's famous example of this (limited character) of instrumental rationality reads as follows: "The precepts for a physician to make his man healthy in a well-grounded way, and for a poisoner to be sure of killing him, are of equal worth insofar as each serves perfectly to bring about its purpose" (*GMS*, AA 04, p. 415; Kant, 1996, p. 68).[2] The purposive-rational *sub-segment* of human *ratio* (our expandable and improvable *technical knowledge* regarding the means necessary to reach old or new goals that we desire) remains (not only in its everyday mode, but also in its algorithmically programmed, machine-generated forms) at all times (albeit not explicitly) embedded *in a complex*, richly textured and ethically charged environment of *practical rea-*

[2] Kant's example is taken up by Hochreiter (2020, p. 370) who writes: "We can use chemicals to poison someone or to cure someone's illness. The human is the decisive factor, the human can use an AI-system for good or something malicious."

только в том, что цифровые технологии, в сущности, происходят из практических изысканий человечества, но и в том, что использование таких технологий постоянно требует моральной оценки полученных с их помощью результатов. Технологии ИИ не работают без человеческого контроля. Если нашей целью является нередуктивная оценка преимуществ и недостатков текущего глобального процесса цифровизации, абсолютно необходимым становится философское исследование человеческих возможностей в области «практического разума», то есть нашей способности размышлять о моральном и правовом обосновании программ, автоматически (и, естественно, не «автономно» в строгом кантовском смысле) выполняемых алгоритмами.

2. «Слабый» и «сильный» ИИ Дж. Сёрла и некоторые (пост)аналитические и герменевтические соображения о различии между «цифровой рациональностью» и человеческим праксисом

В эссе «Сознание, мозг и программы» американский философ Дж. Сёрл проводит различие между «сильным» и «слабым» пониманием ИИ. Согласно сильному ИИ, утверждает Сёрл, «подходящим образом запрограммированный компьютер буквальным образом обладает когнитивными состояниями» (Сёрл, 1998, с. 376). Полагая, что представление о сильном ИИ ошибочно, Сёрл показывает, что компьютеры лишь «моделируют» некую сущность, но не становятся ею (Там же, с. 396). Поэтому он склоняется к «осторожной» интерпретации ИИ, или «слабому ИИ». Основная ценность алгоритмов заключается в том, что они являются «очень мощным инструментом» (Там же, с. 376), позволяющим создавать копии некоторых *аспектов* человеческого разума, но не *всего* человеческого разума, как утверждает сильная интерпретация ИИ.

son. Not only do all digital technologies ultimately have their origin in human practical explorations; our use of these technologies invites at all times a moral assessment of the results which they produce. AI-directed technologies do not run without supervision. If a sober and non-reductive evaluation of the benefits, as well as the losses, that are produced by today's globalised digitalisation processes is seriously on the agenda, the human capacity for "practical reason" has to be philosophically explored, i. e. our ability to reflect on the moral, as well as legal, justifiability of the programmes that algorithms automatically (but certainly nowhere, in Kant's well defined sense, "autonomously") execute.

2. Searle's Distinction between "Weak" and "Strong" AI, and some (Post-)Analytic and Hermeneutic Explorations of the Difference between "Digital Rationality" and Human Praxis

In his essay "Minds, Brains, and Programs", the American philosopher John Searle suggests that we are well advised to distinguish between "strong" and "weak" interpretations of Artificial Intelligence. "According to strong AI", Searle (1987, p. 18) writes, "appropriately programmed computers literally have cognitive states". "Strong AI must be false" (*ibid*.), Searle (1987, p. 37) shows: what computers produce are "simulations", but not the entity they simulate. He thus opts for a "cautious version" of Artificial Intelligence, for "weak AI". The principal value of algorithms is that they are "very powerful tools" (*ibid*., p. 18) which allow us to duplicate *aspects* of human intelligence but not, as strong AI falsely claims, human intelligence *altogether*.

В русле предложенной Сёрлом критики «сильного ИИ» (пост)аналитические философы, такие как Хилари Патнэм, и последователи герменевтической традиции, такие как Хьюберт Дрейфус и Чарльз Тейлор, в последние десятилетия выделили качества *человеческого практического разума*, не поддающиеся полному копированию компьютерными программами.

2.1. В эссе «Проект искусственного интеллекта» Патнэм, логик из Гарварда, одним из первых рассмотрел теории, сводящие «познание к вычислению или происходящим в мозгу процессам» (Putnam, 1992, p. 18). Любые попытки сделать универсальным свойственный физике причинно-следственный способ объяснения и тем самым представить человеческую мысль как автоматизированный алгоритмический процесс (эти попытки свойственны сегодняшним *неверным*, то есть «сильным», интерпретациям ИИ) связаны на уровне глубинных структур с предположением, что точные науки дают желаемое исчерпывающее описание истинного устройства вселенной. В упомянутом тексте 1992 г. Патнэм подвергает критике позитивистски обоснованное понимание сильного ИИ (то есть представление о сознании как о «некоей "вычислительной машине"» (Ibid., p. 3). Эта идея, пишет Патнэм, восходит ко временам «зарождения научного мировоззрения, XVII—XVIII вв.», в частности к Гоббсу, считавшему, что «мышление — это манипуляция знаками в соответствии с правилами, похожими на правила вычисления», и материалистской концепции человека, предложенной Ламетри в его работе «Человек-машина» (Ibid.). Сегодня, и по мнению Патнэма «весьма опрометчиво», «концепция машины Тьюринга» воспринимается как «способ придать точность этой материалистической идее» (Ibid., p. 4). Но и само словосочетание «искусственный интеллект» вводит в заблуждение: «Вообще, ИИ даже не пытается моделировать интеллект… на самом деле он занят написанием хитроумных программ

Along this line of Searle's critique of "strong AI", some of those qualities of *human practical reason* which cannot be sufficiently duplicated by computer programmes were pointed out during the last decades by (post-)analytic philosophers like Hilary Putnam, and by hermeneutics-inspired thinkers like Hubert Dreyfus and Charles Taylor.

2.1. In his essay "The Project of Artificial Intelligence", the Harvard logician Putnam (1992, p. 18) early on took issue with theories that reduce "cognition either to computations or to brain processes". All attempts to universalise the causal explanation mode of physics and thus to interpret human thought as an automated algorithmic process (attempts which are a central characteristic of today's *false*, i. e. "strong" [mis-]readings of AI) are tied in a structurally deep mode to the assumption that the exact sciences [offer us] the long-sought description of the true and ultimate furniture of the universe. In his 1992 text Putnam takes issue with this central "positivism-informed" assumption of strong AI (the assumption that the mind is "a sort of 'reckoning machine'" (*ibid*., p. 3). This idea, Putnam points out, goes back to the "birth of the scientific world view in the seventeenth and eighteenth centuries" — back to Hobbes, who claimed that "thinking is a manipulation of signs according to rules (analogous to calculating rules)", as well as to La Mettrie's materialist conception of man in *L'Homme-Machine* (*ibid*.). Today — although as Putnam remarks "hardly well thought out" — "the notion of a Turing machine" is seen "as a way of making this materialist idea precise" (*ibid*., p. 4). But Artificial Intelligence is actually "a misnomer", Putnam writes: "AI does not "really try to simulate intelligence at all […], its real activity is just writing clever programmes for a variety of tasks" (*ibid*., p. 13). To write

для решения разных задач» (Ibid., p. 13). Конечно, продолжает Патнэм, написание таких программ «важное и полезное дело, хотя и звучит менее увлекательно, чем "моделирование человеческого интеллекта" или "создание искусственного интеллекта"» (Ibid.).

«Цифровая реализация» *ratio* средствами программ ИИ не исчерпывает (практический) разум в полном смысле этого термина, так как, согласно Канту, ни царство человеческой мысли, ни царство человеческого праксиса (и его горизонт надежд) не могут быть достоверно воспроизведены. ИИ в сильном его понимании — сциентистская фантазия, получившая широкую поддержку благодаря, как утверждает Патнэм, нашей склонности ошибочно считать, что «лучшая метафизика — это физика» (Ibid., p. 2). Несостоятельность такого редукционизма можно продемонстрировать с помощью, во-первых, понятия языковой игры, введенного Витгенштейном в «Философских исследованиях» (Putnam, 1981, p. 17—20), и, во-вторых, постаналитического прочтения (некоторых элементов) кантовской концепции *Vernunft*. В своих поздних работах Виттгенштейн убедительно показал, что *использование человеком знаков* заложено в разнообразии *языковых игр*, из которых слагается наш мир. Таким образом, алгоритмические процессы (в той мере, в какой они поддаются интерпретации, то есть являются для нас значимыми) неизбежно находятся в нашем *многосоставном обыденном мире*. Этот мир представляет собой (латентный) фон, который *всегда подразумевается* в нашем восприятии опосредованных алгоритмами процессов: это происходит даже при моделировании (и, *следовательно*, частичном изменении) реального опыта в *виртуальной реальности*, например при имитации полета или в компьютерных играх, которые запускаются и останавливаются игроками — *невиртуальными*, *реальными субъектами-людьми*, а также проявляется *в способе нашего восприятия (реальной реальности)* фильмов, воспевающих

such digital programmes, Putnam continues, is certainly "an important and useful activity, although it does not sound as exciting as 'simulating human intelligence' or 'producing artificial intelligence'" (*ibid*.).

The "digitally executed" *ratio* that AI programmes execute is not at all (practical) reason in its full sense, since in it (in Kant's words) neither the realm of human thought nor that of human *praxis* (and its horizon of hope) can be fully "reproduced". Strong AI is a "scientistic" fantasy, which, as Putnam writes, gets massive support from a tendency in our culture to assume, falsely, "that the best metaphysics is physics" (*ibid*., p. 2). That such a reductionism is seriously flawed can be demonstrated, as Putnam points out, with reference, first, to the concept of the language game in Wittgenstein's *Philosophical Investigations* (Putnam, 1981, pp. 17-20) and, second, via a post-analytical re-reading of (elements of) Kant's conception of *Vernunft*. Late Wittgenstein convincingly argued that *human sign-use* is embedded in a plurality of *language games* that constitute a shared world. All algorithmic processes are thus (insofar as they are interpretable, i. e. meaningful, for us) ultimately and unavoidably situated in our *complex everyday world*. This shared world is the (latent) background that is *always already presupposed* in our receptions of algorithm-mediated processes: even in the simulation (and partial alteration, that is) of reality-experiences in *Virtual Reality* — in flight simulation, for instance; or in video games which are started and stopped *by non-virtual, real human agents* who play them; or in *our (real-reality) reception mode* of VR-glorifying films like *The Matrix*[3]. For a careful exploration of these *preconditions* in which all digital processes are embedded,

[3] See Dreyfus and Taylor (2015, p. 100), where the authors convincingly argue that "we can only conceive [of the scenario of this film] within a framework which structurally offers the possibility of 'waking up' to the real world."

виртуальную реальность, таких, например, как «Матрица»[4]. Вдумчивое исследование подобных *предпосылок*, в рамках которых происходят все цифровые процессы, требует глубокого философского осмысления человеческого праксиса и использования знаков, то есть, как пишет Патнэм, рефлексии, основанной на ключевых кантовских концепциях и многогранной, ориентированной на действие концепции языка позднего Витгенштейна. Размышления об «условиях возможности» ИИ, согласно Патнэму, «тесно связаны с тем, что Кант называл "трансцендентальным исследованием"» (Ibid., p. 16).

2.2. Схожим образом критиковали утрированные и зачастую неверные интерпретации ИИ калифорнийский философ Х. Дрейфус[5] и его канадский коллега Ч. Тейлор. Их совместная работа «Возвращение реализма», опубликованная в 2015 г., критикует текущие попытки *обобщить* инструменталистские интерпретации *праксиса*. Дрейфус и Тейлор тем не менее *основывают свои критические аргументы* не на частично кантианском переосмыслении поздней витгенштейновской концепции «использования языка», как это делает Патнэм, а на ключевых феноменологических представлениях о человеческом существовании как способе бытия-в-мире (*In-der-Welt-sein*). Дрейфус и Тейлор придерживаются — вопреки утверждениям, из которых складывается «сильная» интерпретация ИИ, — *неколебимого плюралистического реализма*[6], ставящего под вопрос «цаan in-depth philosophical reflection on human praxis and sign-use is needed: i. e., as Putnam insists, a reflection that is firstly well advised to learn from the *multifaceted, action-related language concept* of late Wittgenstein, and, secondly, from central thoughts developed by Kant. Reflections on the "conditions of possibility" of AI, Putnam writes, have "a close relation to what Kant called a 'transcendental' investigation" (*ibid*., p. 16).

2.2. In a similar vein, exaggerated (mis-) readings of AI were criticised by the Californian philosopher Hubert Dreyfus[4] and his Canadian colleague Charles Taylor. In their joint book, *Retrieving Realism*, published in 2015, the two philosophers take issue with today's attempts to *generalise* an instrumentalist interpretation of *praxis*. Dreyfus and Taylor, however, do not *develop their critical arguments* like Putnam in a (partially Kantian-inspired) re-interpretation of late Wittgenstein's conception of "language use", but with reference to core phenomenological thoughts that concern our human existence as a mode of *In-der-Welt-sein* (Being-in-the-World). Dreyfus and Taylor defend — against the claims that are raised by "strong" interpretations of AI — a "*pluralist robust realism*"[5] which questions "the vogue in recent decades for accounts of thinking based on the idea that the brain operates in some respects like a computer" (Dreyfus and Taylor, 2015, p. 15). In such accounts, Dreyfus and Taylor argue, *complex* "intuitions that humans have as embodied, social, and cultural agents" are

[4] См. работу Дрейфуса и Тейлора, где авторы убедительно демонстрируют, что «мы можем помыслить [сценарий этого фильма] только в рамках, подразумевающих структурную возможность "пробуждения" в настоящем мире» (Dreyfus, Taylor, 2015, p. 100).

[5] Ранний критический анализ ИИ можно найти в (Dreyfus, Dreyfus, 1986).

[6] «Неколебимый плюралистический реализм» избегает «*редуктивного реализма*, утверждающего, что *наука* объясняет любые способы бытия», и критически относится к «*научному реализму*, согласно которому есть только один способ, которым вселенная может быть разбита на классы так, чтобы каждый использующий подобную терминологию называл именно то, к чему отсылают нас термины естественных классов» (Dreyfus, Taylor, 2015, p. 160).

[4] For an early critical analysis of AI see H. Dreyfus and S. Dreyfus (1986).

[5] This "pluralist robust realism" avoids "a *reductive realism*, which holds that *science* explains all modes of being", and it thus criticises "a *scientific realism*, which holds that there is only one way the universe is carved up into kinds so that every user of such terms must be referring to what our natural-kind terms refer to" (Dreyfus and Taylor, 2015, p. 160).

рящую в последние десятилетия моду описывать мышление, опираясь на представление о том, что функционирование мозга в некотором роде схоже с функционированием компьютера» (Dreyfus, Taylor, 2015, p. 15). Дрейфус и Тейлор утверждают, что в таких описаниях *сложные* «интуиции, которыми обладают люди как воплощенные, социальные и культурные субъекты», привычно исключаются — например, когда человек понимает (в рамках текущего социального взаимодействия), что его собеседник «злится на него» или «что атмосфера на вечеринке внезапно стала напряженной» (Ibid., p. 15—16). Человеческий рассудок чрезвычайно сложен. Поэтому, как показывают приведенные выше примеры, он не сводится к сегментам, имитируемым цифровой рациональностью. «Человеческая языковая способность во всей ее полноте», к детальному анализу которой Тейлор приступает в своей книге «Языковое животное» (Taylor, 2016), располагает *множеством способов рассуждения*, не поддающимся алгоритмическому воспроизведению: будничные моральные размышления, подлинное художественное творчество, религиозные интерпретации человеческого бытия, а также философские попытки систематического объяснения человеческого праксиса и мысли[7].

При всех своих достоинствах предложенные Патнэмом, Дрейфусом и Тейлором постаналитическая и герменевтическая критики (основных компонентов) нынешней риторики «сильного» ИИ — *лишь первый шаг в нужном направлении*. Это подготовка почвы для *перезапуска нередуктивного дискурса* о структуре *практического разума*, то есть дискурса, который выступает в защиту тезиса об алгоритмах как частных случаях *подвида только рацио, выполняющих запрограммированные функции*, которые в конечном счете не являются самодостаточными, но всегда требуют оценки со стороны человеческого практического разума.

[7] См. об этом: (Nagl, 2018).

routinely excluded — for instance, that we are able to know (in an ongoing social interaction) whether my communication partner is "mad at me", or that we can realise "that the atmosphere of the party has suddenly become tense", etc. (*ibid.*, pp. 15-16). Human understanding has an extremely rich texture, and it is thus, as these examples show, not at all limited to those of its segments that digital rationality is able to simulate. The "full shape of human linguistic capacities", which Taylor starts to explore carefully in his book *The Language Animal* (Taylor, 2016), encompasses *many modes of reasoning* that cannot be algorithmically re-enacted: everyday moral reflections as well as genuine artistic creativity, religious interpretations of human existence as well as philosophical attempts to systematically explicate human *praxis* and thought.[6]

Meritorious as Putnam's and Dreyfus and Taylor's post-analytic and hermeneutic critiques of (core components of) today's "strong" AI rhetoric are, they are *just a first step in the right direction*. They prepare the ground for a *re-opening of a non-reductive discourse* on the structure of *practical reason*; a discourse which defends the thesis that algorithms are examples of a *sub-species* of *ratio only*; *that they execute programmed skills* which, ultimately, are not at all self-sufficient, but always require evaluation by human practical reason.

3. "Digital Humanism"

AI is nowhere able to fully "regulate itself"; quite on the contrary, it is in dire need of socio-political control. The increasing awareness of the downside of *legally uncontrolled* AI was expressed, in 2017, by Tim Berners-Lee,

[6] See in this context Nagl (2018).

3. «Цифровой гуманизм»

Ни в одном аспекте ИИ не способен к «саморегулированию». Он отчаянно нуждается в общественно-политическом контроле. Об опасности отсутствия *законодательного регулирования* ИИ говорил в 2017 г. создатель Интернета Тим Бернерс-Ли. В частности, он заявил, что «система дает сбой»[8]. Опасения Бернерса-Ли вызвали широкий отклик — в частности, обсуждалась идея «цифрового гуманизма»[9] как продукта «просвещения и гуманизма», опирающегося на кантовское понятие практического разума (или на элементы этого понятия).

Ниже будет обоснован следующий тезис: философия Канта может быть чрезвычайно полезна в попытках продолжить и расширить постаналитическую герменевтическую критику ИИ, предложенную Патнэмом, Дрейфусом и Тейлором. Для этого необходимо *полное (то есть ориентированное на мораль) понимание* того, почему «инструментальная рациональность» (основное средство технологических инноваций) — не высшая точка развития *Vernunft*, а всего лишь *под*вид рацио.

«Алгоритмы» (или, в кантовских терминах, «императивы умения», генерируемые технологическими средствами) не изобретают сами себя, сами себя не запускают и не ставят перед собой цели, являющиеся результатом рефлексии. Проще говоря, имитируемый алгоритмами «интеллект» направлен на другого, «гетерономен». Он не является продуктом «я», имеющего независимую волю. Так называемая «автономия» ИИ — не что иное, как автоматизм выполнения задач на основе полученных от датчиков данных (при этом он сильно ограничен, несмотря на способность ИИ вероятност-

[8] См.: *Tim Berners-Lee* on the future of the web: 'The system is failing'. URL: https://www.theguardian.com/technology/2017/nov/15/tim-berners-lee-world-wide-web-net-neutrality (дата обращения: 29.08.2021).

[9] «Цифровой гуманизм» Ниды-Рюмелина и Вайденфельд (Nida-Rümelin, Weidenfeld, 2018) — это взвешенная интерпретация ИИ, подчеркивающая необходимость согласования технологий с человеческими ценностями.

the inventor of the World Wide Web, in his famous statement that "the system is failing".[7] Berners-Lee's dismay provoked many reactions: it brought, *inter alia*, the idea of a "Digital Humanism" onto the scene[8], for "enlightenment and humanism", based on (elements of) Kant's concept of practical reason.

The central thesis of the following parts of this essay is that Kant's philosophy can be extremely helpful in the attempt to continue and expand Putnam's and Dreyfus and Taylor's post-analytical and hermeneutical critiques of AI by looking for a *full (i.e., morality-focused) understanding* of why "instrumental rationality" (the key medium of technological innovation) is not the culmination point of *Vernunft*, but a mere *sub*-species of *ratio*.

"Algorithms", (or, in Kantian terms: "imperatives of skill" produced by technological means) neither invent themselves nor start themselves; nor do they set their own goals as a result of reflection. To put it simply: algorithm-simulated "intelligence" is other-directed, "heteronomous"; it is not at all the product of a "self" with an independent will. Its so-called "autonomy" is nothing but a sensor-guided performance automatism (which remains severely limited, even if AI programmes in so-called "machine-learning" can probabilistically adapt to new data). As a technical system, AI is a human product that, like all human products, implements ideas formulated by its author; thus it is not at all able to produce something that fully represents its inventor's capacity for the development of (other, further)

[7] See "Tim Berners-Lee on the Future of the Web": https://www.theguardian.com/technology/2017/nov/15/tim-berners-lee-world-wide-web-net-neutrality (Accessed 29.08.2021).

[8] Nida-Rümelin's and Weidenfeld's *Digitaler Humanismus* (2018) is a sober re-reading of AI which focuses on the need to shape technologies in accordance with human values.

но адаптироваться к новым данным при помощи «машинного обучения»). Как техническая система ИИ — это продукт рук человеческих, который, как и все такие продукты, воплощает идеи своих авторов. Следовательно, он не способен произвести нечто, что бы полностью имитировало способности, позволяющие его создателям прийти к новым изобретениям. (Таким образом, крайне маловероятно, что компьютеры когда-либо смогут полностью имитировать человеческий интеллект[10].)

ИИ можно представить как постоянно увеличивающуюся совокупность потенциально полезных цифровых «инструментов». Эти инструменты тем не менее необходимо систематически контролировать так, чтобы они не могли причинить вред тем, кто попадает под их воздействие. ИИ может нанести ущерб как *невольно* (как, например, в случае программ ИИ, используемых рекрутинговыми агентствами и оценивающих при помощи непрозрачных критериев то, насколько безработный клиент «впишется» в рынок труда, см.: (Klingel, Kraft, Zweig, 2020))[11], так и *намеренно* (самый ужасный пример — автоматизированные атаки военных дронов, управляемых ИИ).

Изначально на цифровизацию возлагались большие надежды. Ожидалось, что она даст миру большую свободу общения. Но по мере внедрения инноваций стало ясно, что эти на-

inventions. (This fact makes it most unlikely that computers will ever be able to simulate human intelligence fully.[9])

AI is best understood as an (ever increasing) assemblage of (possibly useful) digital "tools". These tools need, however, to be carefully controlled in order not to harm those who are affected by them. AI-caused damage can occur either *involuntarily* (as the result, for instance, of AI-guided programmes used by employment agencies that evaluate by non-transparent criteria the future "integration potential" of their unemployed clients in the labour market (see Klingel, Kraft and Zweig, 2020)[10] or *voluntarily* (as is, most abhorrently, the case with automated assaults performed by AI-guided war drones).

Digitalisation started as a big promise — the promise to enhance freedom of communication on a global basis. The introduction of this innovation soon made it clear, however, that these hopes were exaggerated. In AI-driven social media, one-sided, affectively tuned-up communication bubbles that insistently re-affirm narrow, closed world-views started to destroy more and more the open, consensus-oriented

[10] Сара Шпикерман, возглавляющая Институт информационных систем и общества Венского университета экономики и бизнеса, выделяет пять причин, по которым «системы ИИ никогда не будут похожи на человека: 1) у систем ИИ мало информации, схожей с той, которой оперируют люди; 2) у систем ИИ не бывает человекоподобных реакций; 3) системы ИИ не могут думать как человек; 4) у систем ИИ нет мотивации, подобной человеческой; 5) системы ИИ не обладают автономией, подобной человеческой». Пункты 1 и 3 находят широкую поддержку в критической рефлексии Патнэма, Дрейфуса и Тейлора. То, что системы ИИ не «действуют» ни в одном из тех смыслов, в которых мы понимаем человеческое действие (пункты 4 и 5 у Шпикерман), легло в основу идеи «цифрового гуманизма», которая обоснованно указывает на то, что ИИ — это всего лишь «инструмент», который не может «сам себя направлять», но создается и контролируется человеческим субъектом (Spiekermann, 2020, p. 90).

[11] См. критическое исследование Кэтрин Форрест (Forrest, 2021).

[9] Sarah Spiekermann, Head of the Institute of Information Systems and Society at Vienna's University of Economics and Business, points out that there are at least five reasons why "AI systems will never become human-like: 1. AI systems have little human-like information; 2. AI systems cannot react in a human-like manner; 3. AI systems cannot think in a human-like manner; 4. AI systems have no human-like motivation; 5. AI systems have no human-like autonomy." Reasons 1 to 3 find ample support in the critical reflections on AI that Putnam, Dreyfus and Taylor have voiced. That AI systems do not "act" in any sense that is fully identical with human action (Spiekermann's reasons 4 and 5) is a core theme of the idea of a "digital humanism", which, for very good reasons, insists that AI is ultimately nothing but a "tool" that does not "direct itself" but is created and must be controlled by human agents (Spiekermann, 2020, p. 90).

[10] See in this context also the critical study by Katherine B. Forrest (2021).

дежды были во многом напрасны. В социальных медиа, управляемых ИИ, односторонние, аффективно настроенные коммуникационные пузыри, в которых пользователи постоянно находят подтверждение своему узкому, ограниченному взгляду на жизнь, начали разрушать открытый, ориентированный на консенсус дискурс, без которого невозможна плодотворная публичная дискуссия (см.: Marantz, 2019)[12].

Именно эти побочные продукты «цифровой революции» заставили Бернерса-Ли высказаться по поводу рисков провала цифровизации[13]. В 2017 г. он убедился, что вопреки его изначальным ожиданиям цифровизация не привела к созданию открытой платформы, позволяющей обмениваться информацией и развивать сотрудничество, невзирая на границы, а также использовать иные новые возможности. Дело не только в том, пишет он, что Интернет используется для распространения дезинформации и пропаганды, а вниманием людей управляет тонко настроенный искусственный интеллект, научившийся отвлекать пользователей от чего угодно. Что важнее, в Интернете все заметнее становятся действия цифровых регуляторов, чьи алгоритмы стали оружием в руках искусных манипуляторов.

Заявление Бернерса-Ли о том, что «система дает сбой», существенно повлияло на текущую дискуссию о цифровой трансформации и этике. Нида-Рюмелин и Вайденфельд критикуют «сильную» интерпретацию ИИ («идеологию Силиконовой долины»), подчеркивая, что только гуманизм, способный вобрать в себя ИИ как (контролируемый) «инструмент», может быть

[12] Ч. Тейлор, П. Нанц и М. Бобьен-Тейлор пишут в недавней работе, что, хотя, с одной стороны, «цифровизация дает гражданам простой и широкий доступ к информации», с другой стороны, «социальные сети, которые по большей части анонимны», еще дальше отдаляют «граждан от сферы политического. Позволяя найти людей с той же точкой зрения в "эхо-камерах", в которых альтернативное мнение оказывается под запретом, такой вид медиапотребления препятствует коллективному обучению и вдумчивому размышлению. Более того, он создает благодатную почву для электронного популизма» (Taylor, Nanz, Beaubien-Taylor, 2020, p. 3).

[13] См. сноску 8 выше.

discourse which is essential for any flourishing public debate (see Marantz, 2019).[11]

Undesirable (by-)products of the "digital revolution" like these lead Berners-Lee to his famous warning that the digitalisation process is threatened with failure.[12] In 2017 he saw his early expectations that digitalisation would produce an open platform that allows anyone to share information, access opportunities and collaborate across geographical boundaries challenged on several fronts. It was not only, as he writes, that the internet was being used to spread misinformation and propaganda, and people's attention was being held by very finely-tuned AI that worked out how to distract them; the net was being used more and more by increasingly powerful digital gatekeepers whose algorithms were weaponised by master manipulators.

Berners-Lee's awareness that "the system is failing" significantly influenced the recent debate on digital transformation and ethics. Nida-Rümelin and Weidenfeld criticised "strong AI" — the "Silicon Valley ideology" — pointing out that only a humanism which is able to integrate AI as a (controllable) "tool" is a plausible retort to ideologically motivated misreadings of Artificial Intelligence: "The only plausible contrast programme to strong AI and its implicit mechanistical thinking is a Digital Humanism — a humanism which neither doubts nor endangers human authorship, but expands

[11] Charles Taylor, Patrizia Nanz and Madeleine Beaubien-Taylor pointed out recently that while, on the one hand, "digitalization provides citizens with easy and broad access to information", the "largely anonymous social networks" distance, on the other hand, "citizens from the political sphere. With its focus on finding sympathetic others within 'echo chambers' that reject dissenting opinion, this form of media consumption acts as a barrier to collective learning and meaningful deliberation. Instead, it provides fertile ground for electronic populism" (Taylor, Nanz and Beaubien-Taylor, 2020, p. 3).

[12] See footnote 7, above.

разумным ответом на идеологически мотивированные неверные интерпретации искусственного интеллекта: «Единственная разумная программная альтернатива сильному прочтению ИИ и его имплицитному механистическому мышлению — это цифровой гуманизм, то есть гуманизм, не сомневающийся в «человеческом творчестве и не ставящий его под угрозу, а лишь расширяющий его через внедрение цифровых технологий» (Nida-Rümelin, Weidenfeld, 2018, S. 60). Последовательно оспаривая новый механистический редукционизм, подпитываемый неколебимой верой в нейронауки, Нида-Рюмелин и Вайденфельд подчеркивают (порой используя почти кантианскую терминологию)[14] различие между алгоритмами и человеческим действием (Ibid., S. 61). Системы ИИ не имеют ни «восприятия», ни «эмоций». Они неспособны «принимать решения», даже если были созданы с целью «имитировать» человеческое поведение (Ibid., S. 205). Но, говоря о (предполагаемых) возможностях ИИ, идеологи Силиконовой долины продолжают безответственно использовать сильные метафоры[15].

[14] Концепция «практического разума», предложенная Нидой-Рюмелином, уделяет особое внимание тому, что он называет «ограниченной» ("bounded") или структурной рациональностью, и при этом критикует то, что ее автор описывает как кантовский деонтологический рационализм. См.: (Nida-Rümelin, 2019, p. IX—X). С одной стороны, Нида-Рюмелин (вместе с Кантом) утверждает справедливость для многих форм человеческого действия того факта, что «действие не выбирается на основе соответствия текущему желанию человека. Именно это его свойство кажется наиболее близким кантианскому представлению о том, что разумный выбор действия автономен, а не гетерономен, то есть не задан сиюминутными склонностями». Принимая формалистское и, вероятно, неверное прочтение кантовской концепции морального суждения в ее полноте (у такого прочтения имеются далеко идущие последствия), Нида-Рюмелин, с другой стороны, отмежевывает свою концепцию практического разума от того, что он считает кантовским проектом (см. раздел 4 данной статьи).

[15] Это, по мнению Ниды-Рюмелина и Вайденфельд, переводит ИИ-лобби в гротескный режим самообмана: «Нам следует остерегаться самообмана, — пишут они, — когда мы, сначала разрабатывая цифровые машины, имитирующие эмоции, познания и решения, затем с удивлением обнаруживаем, что эти машины производят впечатление, что у них есть эмоции и они способны познавать и принимать решения» (Nida-Rümelin, Weidenfeld, 2018, S. 205).

it by means of digital techniques" (Nida-Rümelin and Weidenfeld, 2018, p. 60; tr. L. N.). In strict opposition to the new mechanistic reductionisms that today, in AI discourse, are further nourished by the uncritical boom of neuroscience, Nida-Rümelin and Weidenfeld insist (often in almost Kantian terms[13]) on the difference between algorithms and human action (*ibid*., p. 61). AI systems do neither "perceive", nor do they have "emotions", nor are they able to "decide" — even if some of them are designed in a manner which (more or less) "simulates" human behavior (*ibid*., p. 205). However, in connection with the (purported) abilities of AI, the Silicon Valley ideologues use *metaphors* in an extremely careless manner.[14]

In opposition to the feuilleton-related fantasies of the AI-lobby, "digital humanists" want to promote the real potentials, as well as to analyse the real dangers of AI in a sober manner: "Digital Humanism does not dream of a totally new form of human existence like the transhumanists; it remains skeptical vis-à-vis Utopian

[13] Nida-Rümelin's concept of "practical reason", is — as focused on (what he calls) "bounded" or "structural" rationality — critical of (what he claims to be) Kant's deontological "rationalism". See Nida-Rümelin (2019, pp. IX-X). Nida-Rümelin, on the one hand, points out (with Kant) that in many forms of human action, "the action is not chosen because it corresponds to a present desire of the person. It is this feature which one might find most closely related to the Kantian conception: the reasonable choice of actions is autonomous and not heteronomous, i. e., determined by momentary inclinations." But in a formalistic [mis-]reading of Kant's full conception of moral judgment which has far-reaching consequences, Nida-Rümelin, on the other hand, distances his own conception of practical reason from (what he sees as) Kant's project (see segment four of this essay, below).

[14] This, Nida-Rümelin and Weidenfeld argue, leads the AI-lobby to a grotesque mode of self-deceit: "We should avoid deceiving ourselves", they write, "by first constructing digital machines which are able to simulate emotions, knowledge and decisions, and then to note in surprise that these machines give the impression that they have emotions and are able to know and to decide" (Nida-Rümelin and Weidenfeld, 2018, p. 205, tr. L. N.).

В противовес фантазиям ИИ-лобби, напоминающим бульварный роман, «цифровые гуманисты» говорят о реальном потенциале технологий и трезво оценивают опасности, связанные с ИИ: «Цифровой гуманизм не мечтает о некоей новой форме человеческого существа, как это делают трансгуманисты; он скептически относится к утопическим ожиданиям, но с оптимизмом смотрит на способность человека творчески использовать цифровые возможности» (Ibid., S. 207).

Эта программа убедила многих. В апреле 2019 г. факультет компьютерных наук Венского технологического университета провел Первый международный семинар «Цифровой гуманизм»[16]. В Венском манифесте, опубликованном по итогам мероприятия, были изложены следующие идеи: «Термин "цифровой гуманизм" намеренно отсылает к концепциям гуманизма и Просвещения, согласно которым человек, их центральный элемент, сам отвечает за свое мышление (ср.: Nida-Rümelin, Weidenfeld, 2018). Под этим термином мы понимаем подход, описывающий, анализирующий и прежде всего пытающийся повлиять на сложное взаимодействие технологий и человечества в интересах общества, полностью соответствующего требованиям всеобщих прав человека. Мы можем, имеем право и должны использовать мыслительные способности, будучи творцами своей собственной жизни; личная автономия и свобода принятия решений — предпосылки открытого и демократического общества» (Werthner, 2020, p. 350). Как показывает это краткое

[16] На этом мероприятии причины сегодняшнего неопределенного статуса ИИ были сформулированы в виде шести кратких характеристик негативных тенденций цифровизации: 1) «тенденция растущей концентрации в Интернете»; 2) «IT как решающий фактор в экономических, политических и военных конфликтах»; 3) «влияние, оказываемое на политику эхо-камерами и фейками»; 4) «серьезные проблемы в области безопасности и защиты данных»; 5) «растущая автоматизация работы»; 6) «вопрос, связанный с разработками в области ИИ, относительно того, в какой степени людей можно/стоит заменить машинами в процессах принятия решений» (Werthner, 2020, p. 349—350).

expectations, but is optimistic with regard to human ability to shape positively the potential offered by digitalisation" (*ibid*., p. 207; tr. L. N.).

This programme was seen by many as quite convincing. In April 2019 a First International Workshop "Digital Humanism" was organized by the Faculty of Computer Science at the University of Technology in Vienna.[15] In the Vienna Manifesto, which was published on that occasion, the following ideas were formulated: "The term 'Digital Humanism' is intentionally chosen to refer to the concepts of Humanism and Enlightenment, according to which the human is responsible for his or her thinking and is at the focus (*cf*. Nida-Rümelin and Weidenfeld, 2018). We understand this term as an approach which describes, analyzes, and more than anything tries to influence the complex interaction of technology and humanity — in the interest of a society which complies fully with the requirements of universal human rights. We have the freedom, the right, and the responsibility to make use of our own thought power, we are the authors of our own lives; personal autonomy and freedom to make decisions are the prerequisites for an open, democratic society" (Werthner, 2020, p. 350). As this short summary of the core motifs of Digital Humanism shows, the idea of an ethical regulation of AI which this project advocates, is — in substantial respects — based on Kantian thought. In the context in which these leading ideas are expressed in the Vienna Manifesto they are not philosoph-

[15] In this workshop the reason for today's precarious status of AI was summarised in the following six short characterisations of negative developments in the digitalisation process: 1) "Concentration tendencies in the Web"; 2) "IT as a decisive factor in economic, political and military conflicts"; 3) "Political influencing by echo chambers and fake news"; 4) "Drastic problems in the area of data protections and security"; 5) "Increasing automation of work"; 6) "The question associated with developments in the field of Artificial Intelligence as to the extent to which humans can/should be replaced by machines in decision-making processes" (Werthner, 2020, pp. 349-350).

кое изложение основных положений цифрового гуманизма, идея этического регулирования ИИ, лежащая в основе данного проекта, позаимствовала многие существенные идеи из кантианской мысли, хотя в контексте Венского манифеста эти идеи не подвергаются философскому анализу, а лишь упоминаются. Кантианские идеи цифрового гуманизма не могут получить силу аргумента без тщательной их интерпретации. Но прежде чем перейти к философскому анализу в четвертом разделе данной статьи, необходимо сделать небольшое предварительное замечание об ИИ и этике. Как я уже отмечал в своем выступлении в 2020 г. (см.: Nagl, 2022), для регулирования и контроля ИИ необходимы два вида этического дискурса. Это, во-первых, дискурс о внутренних «правилах», которыми руководствуются алгоритмы, и, во-вторых, рефлексия о внешних критериях (юридических и общественно-политических), определяющих публичное регулирование ИИ, то есть серьезная дискуссия об этических принципах человеческого действия, в которых кантовская многосоставная концепция этики может сыграть существенную роль.

Что касается первого дискурса, очевидно, что в серьезных попытках исследовать «внутренние правила» «роботов» не место аналитическим уловкам[17]. Компьютерные программы, вычисляющие трансиндивидуалистическую «общую пользу» и, таким образом, действую-

[17] Здесь можно привести пример, взятый из дискуссии о «беспилотных автомобилях». Так как на дороге беспилотные машины могут столкнуться с неожиданными ситуациями, «проблема вагонетки» («необходимый ингредиент курса этики для бакалавров кафедр аналитической философии», как называет ее М. Митчелл), становится «основным тезисом в дискуссиях об этических аспектах ИИ» (Mitchell, 2019, p. 127—128). Редуктивным этическим дискурсам нет места в дискуссиях, посвященных, например, тому, стоит ли беспилотным автомобилям давать разрешение переезжать пожилых, чтобы спасти молодых, когда столкновение неизбежно. (Постгуманисты заявляют даже, что «роботы» [= (квази-)разумные машины] должны обладать теми же правами, что и люди. Ср. критическую работу Ниды-Рюмелина и Вайденфельд «Роботы — новые [цифровые] рабы» (Nida-Rümelin, Weidenfeld, 2018, S. 23—31.)

ically analysed, however, but just mentioned casually. In order to gain their full argumentative strength, the Kantian motifs of Digital Humanism will have to be elucidated in a more careful way. Before we proceed, in part four of this presentation, to this philosophical analysis, a short preliminary remark on AI and ethics is necessary. As I have pointed out in a presentation in 2020 (*cf.* Nagl, 2022), two modes of ethical discourse are required for the regulation and control of AI. These are, first, a discourse about the internal "rules" that guide algorithm-directed operations and, second, extensive reflections on the external — juridical and socio-political — criteria governing the public regulation of AI, i. e. an in-depth debate on the ethical principles of human action in which Kant's complex conception of ethics can play a substantial role.

With regard to the first discourse it is obvious that, in serious attempts to examine closely the "inbuilt rules" that direct "bots", analytic shortcuts will be totally inadequate.[16] Computer programmes that merely calculate a trans-individualistic "general use" and thus operate along the lines of a utilitarian "consequentialism", are quite obviously incompatible with *the basic right of every human individual to life*, a right which is encoded in the constitution of all modern states. Nida-Rümelin and Weiden-

[16] The following example taken from the debate about "self-driving cars" will illustrate this. In view of the unexpected traffic situations that automatic cars could face, the so called "trolley problem" (a "staple of undergraduate ethics courses in Analytic Philosophy Departments", as Melanie Mitchell writes) has become "a central talking point in discussing AI ethics" (Mitchell, 2019, pp. 127-128). Reductionist ethical discourses tend to terminate in debates that centre, for instance, on the question whether automatic cars are to be allowed, in the situation of an unavoidable collision, to run over senior citizens in order to protect young lives. (Post-humanists even suggest that "bots" [= (quasi-)"intelligent" machines] are entitled to "human rights". *Cf.* in this context Nida-Rümelin's and Weidenfeld's (2018, pp. 23-31) critique "Robots as new [digital] slaves".

щие по принципу утилитарного «консеквенциализма», совершенно очевидным образом несовместимы с *фундаментальным правом человека на жизнь*, заложенным в конституциях всех современных государств. Нида-Рюмелин и Вайденфельд сформулировали эту восходящую к Канту максиму в «Цифровом гуманизме» следующим образом: «Нарушение фундаментальных прав не может быть компенсировано выгодами, полученными третьими лицами, какими бы значительными эти выгоды ни были. К человеку нельзя относиться только как к средству. Люди не "оптимизируются". В чрезвычайных ситуациях мы действуем в соответствии с моральной интуицией, а не оптимизирующими вычислениями» (Nida-Rümelin, Weidenfeld, 2018, S. 96—97).

Необходим ясный дискурс о внешних (юридических и общественно-политических) критериях, регулирующих публичное и частное использование ИИ, то есть серьезная философия дискуссия об этических основаниях человеческого действия (принципах, которые, согласно Канту, не могут быть полностью смоделированы путем «гетерономного» механического выполнения запрограммированных «норм», даже если эти нормы, как в случае продвинутого ИИ, способны вероятностно «переадаптироваться»).

4. Некоторые соображения о том, как ключевые этические идеи «цифрового гуманизма» могут быть подкреплены прямым обращением к идеям Канта

Кантовская многосоставная теория «моральных суждений» может оказаться полезной в поиске общеприемлемых критериев человеческой оценки программ ИИ и контроля над ними. «Цифровой гуманизм» опирается во многих существенных аспектах на преимущественно популярные изложения кантовского этического дискурса. Этические концепции, к которым обращаются авторы Венского манифеста (2018, pp. 96-97; tr. L. N.) expressed this Kant-inspired maxim in *Digitaler Humanismus* correctly as follows: "The violation of basic rights cannot be offset by the advantages gained by third parties, however great they may be. No human being ought to be treated merely as a means. Humans do not 'optimise'. In emergency situations we act in accordance with moral intuition and not an optimising calculation."

What is called for, is an explicit discourse about the external — juridical and socio-political — criteria governing the public and private use of AI: a philosophical, in-depth debate, that is, on the ethical principles of human action (on principles that, in Kant's words, cannot be fully simulated by "heteronomous" mechanical executions of programmed "norms", even if these norms, as in advanced AI, are able probabilistically to "re-adapt").

4. Some Suggestions how the Leading Ethical Ideas of "Digital Humanism" can be Further Solidified with Explicit References to Kant

Kant's complex theory of "moral judgment" can be of great help in the search for generally acceptable criteria for the human control and evaluation of AI-guided programmes. "Digital Humanism" is influenced — in substantial respects — by popularised versions of Kant's discourse on ethics. The ethical concepts used by the authors of the *Vienna Manifesto* remain, however, rather vague. Part of the reason for this may be that Nida-Rümelin — who coined the term "digital humanism" — advocates a concept of ethics which is based on Kant in some respects, but which is *ultimately non-Kantian*.

феста, тем не менее не совсем очевидны. Вероятно, это можно объяснить тем, что Нида-Рюмелин, автор термина «цифровой гуманизм», продвигает такую концепцию этики, которая хотя и восходит к Канту в некоторых аспектах, является *в конечном счете некантианской*.

4.1. Критические размышления о критике Канта Нидой-Рюмелином

Анализ «практического разума», проведенный Нидой-Рюмелином, уделяет особое внимание тому, что тот называет «структурной рациональностью». В этом анализе наиболее очевидно влияние Канта, но при этом Нида-Рюмелин, *в сущности, критикует (по его собственному определению) кантовский деонтологический «рационализм»*. В главе «Структуры субъектности» книги «Структурная рациональность и другие эссе о практическом разуме» Нида-Рюмелин указывает (вслед за Кантом), что для человеческого действия характерна способность агента к автономному самоопределению, способ действия, при котором «выбор действия не основывается на текущем желании человека» (Nida-Rümelin, 2019, p. 18). Именно эту особенность человеческого действия, пишет он, «можно считать наиболее близкой к кантовской концепции: разумный выбор действий является автономным, а не гетерономным, то есть не определяется сиюминутными склонностями» (Ibid.). Однако и после этого кантианского заявления Нида-Рюмелин придерживается преимущественно формалистической интерпретации этики Канта, которая исключает кантовскую многосоставную концепцию морального суждения. Таким образом он проводит различие между своей собственной концепцией «структурной рациональности» и тем, что он видит как небезупречное основание этики Канта: «Данное сходство [с Кантом] не должно приводить к нормативному априоризму», — пишет он. Кант в прочтении Ниды-Рюмелина ошибочно привержен помещенному в *форма-*

4.1. Critical Reflections on Nida-Rümelin's Critique of Kant

Nida-Rümelin's analyses of "practical reason" are focused on (what he calls) "structural rationality". They are most obviously influenced by Kant, but *they ultimately terminate in a critique of (what he claims to be) Kant's deontological "rationalism"*. In his analysis of the "Structures of Agency" in *Structural Rationality and Other Essays on Practical Reason* Nida-Rümelin points out (along the lines of Kant) that human action is characterised by the ability of the agent to exercise autonomous self-determination; by a mode of action, where "the action is not chosen because it corresponds to a present desire of the person" (Nida-Rümelin, 2019, p. 18). It is this feature of human action, he writes, "which one might find most closely related to the Kantian conception: the reasonable choice of actions is autonomous and not heteronomous, i. e. determined by momentary inclinations" (*ibid*.). However, after this pro-Kantian statement, Nida-Rümelin uses a predominantly formalistic interpretation of Kant's ethics that excludes Kant's complex conception of moral judgment in order to distinguish his own conception of "structural rationality" from (what he sees as) Kant's flawed foundation of ethics: "This similarity [with Kant] should not lead to normative apriorism," he writes. Kant, in Nida-Rümelin's reading, falsely advocates, via his *formalistically-dimensioned* "categorical imperative", a situation-*insensitive* apriorism, in which the actual concreteness of moral deliberation and of moral judgment is nowhere validly reconstructed. Thus Nida-Rümelin concludes: "[T]he practical philosophy of Kant and of many of his successors is connected to an untenable form of rationalism" (*ibid*.).

листическое измерение «категорическому императиву», независимому от ситуации априоризму, в котором реальная конкретность морального размышления и морального суждения не может быть достоверно реконструирована. Нида-Рюмелин приходит к следующему выводу: «Практическая философия Канта и многих его преемников неразрывно связана с несостоятельной формой рационализма» (Ibid.).

Однако, как будет показано ниже, данная критика опирается на неверное прочтение процесса морального размышления и принятия решений, которое, согласно теории самого Канта, никогда не следует абстрактной идее «формального» подведения «случая» под «правило».

Нида-Рюмелин, симпатизируя некоторым элементам этики Канта, не учитывает, что кантовский анализ автономного действия имеет две составляющие. Во-первых, кантовский анализ (в своем формальном описании) опирается на идею о том, что человеческая способность действовать как самоопределяющееся существо делает каждого человеческого субъекта «целью самой по себе». В этом *формальном* принципе заложено различие между «человеческими личностями» и другими сущностями, не способными действовать, опираясь на практический разум (в контексте наших размышлений здесь можно привести пример управляемых алгоритмами «роботов»). Во-вторых, этот формальный принцип, согласно Канту, должен быть ситуативно конкретизирован в любом из своих воплощений. В нашем «практическом познании», пишет Кант в «Основоположении к метафизике нравов», «чистая часть» этики, то есть «законы *a priori*», требует, кроме прочего, «изощренной опытом способности суждения для того, чтобы отчасти различить, в каких случаях они находят свое применение, отчасти открыть им доступ к воле человека и убедительность для исполнения» (AA 04, S. 389; Кант, 1997, с. 47—49). Эти два процесса (которые не были проанализированы Нидой-Рюмелином в его критике Канта) тщательно исследовали американские кантоведы Б. Херман и А. Вуд.

However, as I will try to show in the following, this critique rests on a (mis-)reading of the process of moral deliberation and decision-making which, according to Kant's own theory, never merely follows the abstract idea of a "formal" subsumption of a "case" under a "rule".

Nida-Rümelin, while sympathising with elements of Kant's ethics, does not bear in mind that Kant's analysis of autonomous action has two constitutive elements: it firstly (and formally described) rests on the idea that, since every human being has the ability to act in a self-determined manner, every human subject is an "end in itself". This *formal* principle marks the essential difference between "human persons" and other entities that are not able to act from practical reason (such as, for instance, in the context of our reflections, algorithm-directed "bots"). This formal principle must, however, secondly — as Kant is well aware — in any of its concrete implementations be situationally specified. Among our "practical cognitions", Kant writes in *Groundwork of the Metaphysics of Morals*, the "pure part" of ethics — "the laws *a priori*" — requires in addition "a judgment sharpened by experience, partly to distinguish in what cases they are applicable and partly to provide them with access to the will of the human being and efficacy for his fulfilment of them" (*GMS*, AA 04, p. 389; Kant, 1996, p. 45). These two processes (which are left unanalysed in Nida-Rümelin's reading and critique of Kant) were carefully analysed, however, by the American Kant scholars Barbara Herman and Allen Wood.

4.2. The Practice of Moral Judgment

In her book, *The Practice of Moral Judgment*, Barbara Herman (1993) shows that many of the

4.2. Практика морального суждения

В книге «Практика морального суждения» Барбара Херман показывает, что многие стандартные прочтения кантовской «процедуры КИ» (как ее называют аналитические философы) не позволяют провести всеобъемлющий анализ ее сложности, ошибочно сближая ее с *процедурой подведения*, в которой «общее правило» применяется к «конкретному случаю». В седьмой главе, озаглавленной «Моральное размышление и вывод обязанностей», Херман пишет: «Интерпретаторы и критики кантовской этики уделяют значительное внимание категорическому императиву... Идет нескончаемая дискуссия о том, работает ли КИ... Гораздо реже задается вопрос, какую роль КИ играет в моральном суждении? Предполагается, что ответ на него очевиден» (Herman, 1993, p. 132). Однако здесь Херман высказывает серьезные сомнения: «Я все больше убеждаюсь, что это не так». Она ставит под вопрос «общепринятое убеждение», которое можно найти в стандартных аналитических теориях и теориях рационального выбора, о том, что кантовская «процедура КИ» — это своего рода «алгоритм для морального размышления». Критикуя это «рационалистическое» заблуждение, Херман «очерчивает альтернативную роль КИ в рамках более сложной кантовской теории морального суждения» (Ibid.).

Херман утверждает, что в кантовской концепции практического разума каждое конкретное моральное размышление затрагивает *герменевтическую чувствительность субъекта к контексту предполагаемого поступка* (то есть человеческий потенциал, который, как было показано во втором разделе в контексте идей Дрейфуса и Тейлора, ни один компьютер не в состоянии полностью имитировать). Таким образом, моральное действие зависит от способности субъекта оценивать ситуацию, в которой совершается действие. Если упустить из виду этот *герменевтический* элемент кантовского анализа праксиса, этику Канта легко принять за абстрактную «формальную» процедуру[18].

[18] См. об этом: (Nagl, 1983, S. 41—42 (раздел 1.3.2)), а также (Nagl, 2022).

standard readings of (what analytic philosophers call) Kant's "CI-procedure" evade a full analysis of its complexity by wrongly approximating it to a *subsumption procedure* in which a "general rule" is applied to a "particular case". In chapter 7 of her book, headed "Moral Deliberation and the Derivation of Duties", she writes: "Interpreters and critics of Kant's ethics are heavily invested in the Categorical Imperative […]. There is an endless discussion about how or whether the CI works […]. A question that is much less frequently asked is: what role does the CI have in moral judgment? That is supposed to be obvious" (Herman, 1993, p. 132). Herman doubts this very much, however, and thus concludes: "I am increasingly sure it is not." She calls into question the "received view", expressed in standard analytic and rational choice theories, that Kant's "CI-procedure" provides a kind of "algorithm for moral deliberation". Criticising this as a "rationalistic" misunderstanding, Barbara Herman "sketch[es] an alternative role for the CI within a more complex Kantian theory of moral judgment" (*ibid*.).

Herman points out in her analyses that in Kant's concept of practical reason every concrete moral deliberation includes the agent's *hermeneutical sensitivity for the context in which her or his intended action is embedded* (a human potential, i. e. one which — as was argued with reference to Dreyfus and Taylor in part two of this paper — no computer is able to simulate fully). Moral action thus depends on the agent's capacity to assess the situation in which his action is performed. If this *hermeneutical* element of Kant's analysis of *praxis* is overlooked, his ethics is easily misunderstood as nothing but an abstract "formal" procedure.[17]

[17] See in this context Ludwig Nagl (1983, pp. 41-42 (sec. 1.3.2)) as well as Nagl (2022).

Херман подчеркивает (вслед за Кантом) существование «фоновых условий морального суждения, наличие которых и делает размышление возможным» (Ibid., p. 157). Она справедливо отмечает, что этика Канта как теория не призывает к механическому применению «всеобщих моральных правил». Такие правила, пишет Херман, «абстрактны, беспристрастны и безличны»: они «не задают единый стандарт действий и размышлений в кантовской этике» (Ibid., p. 43). Она продолжает: «Категорический императив — это подлежащий рассмотрению принцип более высокого порядка, а не абстрактное и всеобщее правило. Он не оперирует общими описаниями действий, под которые *подводятся* конкретные случаи, а дает процедуру, с помощью которой конкретное структурируется в моральном смысле» (Ibid., p. 44; курсив мой. — *Л. Н.).* Полностью проанализированная структура КИ включает в себя существенные элементы герменевтической оценки ситуации: «Поскольку КИ используется для оценки максим, а максимы — это субъективные принципы, на основании которых действуют субъекты, — пишет Херман, — выстраивающий максиму субъект должен включать в нее только те характеристики личности и обстоятельств, которые необходимы для описания *его* собственного действия» (Ibid., p. 44). Моральное действие, таким образом, не является механически осуществляемым «подведением». Как было показано выше, сам Кант в предисловии к «Основоположению к метафизике нравов» обращается именно к этой сложной структуре. Отметив, что «основу обязательности до́лжно искать не в природе человека или в обстоятельствах в мире, в какие он поставлен, но *a priori* исключительно в понятиях чистого разума», Кант поясняет, что моральные законы «требуют изощренной опытом *способности суждения* для того, чтобы отчасти различить, в каких случаях они находят свое применение» (AA 04, S. 389; Кант, 1997, с. 47; курсив мой. — *Л. Н.*) Этот важный аспект этики Канта упускает из виду как большинство аналитических прочтений «процедуры КИ», так и критика Канта, предложенная Нидой-Рюмелином.

Herman (1993, p. 157) explicitly emphasises (with Kant) that there are "background conditions of moral judgment that must be present for deliberation to be possible." She rightly points out that Kant's ethics is not a theory which calls for a mechanical application of "general moral rules". Such rules, as Herman writes, "are abstract, impartial, and impersonal": they "do not provide the normal standard of action and deliberation in Kantian ethics" (*ibid*., p. 43). And somewhat later: "The Categorical Imperative is a higher-order deliberative principle, not an abstract and general rule. Instead of including very general descriptions of actions under which the particular is to be *subsumed*, it provides a procedure for structuring the particular in a moral way" (*ibid*., p. 44; emphasis L. N.). If the CI procedure is fully analysed, it thus includes substantial elements of a hermeneutical assessment of the present situation: "Since the CI is used to assess maxims, and maxims are the subjective principles on which the agents in fact act", Herman points out, "when an agent constructs his maxim he is to include in it just that detail of person and circumstance necessary to describe *his* action" (*ibid*.). Moral action is thus not a "subsumption" that can be mechanically carried out. As has been mentioned, Kant himself in the "Preface" to his *Groundwork of the Metaphysics of Morals* deals with this complex structure when he writes — after pointing out that, in an analysis of moral action, "the ground of obligation must not be sought in the nature of the human being or in in the circumstances of the world in which he is placed, but a priori simply in concepts of pure reason" — that these laws "no doubt still require a *judgment* sharpened by experience, […] to distinguish in what cases they are applicable" (*GMS*, AA 04, p. 389; Kant, 1996, p. 45; emphasis L. N.) This important aspect of Kant's ethics is overlooked in most analytic readings of the "CI procedure", as well as in Nida-Rümelin's critique of Kant.

При правильном прочтении кантовские изыскания о *сложности* человеческих размышлений могут оказаться полезными для «цифрового гуманизма», который подчеркивает разницу между полноценным человеческим действием и его (частичной) цифровой симуляцией, пытается эту разницу отстоять и переосмыслить. Алгоритмы ИИ, по крайней мере в двух существенных аспектах, неспособны «автономно выполнять» разумные моральные действия. Во-первых, они неспособны *полностью* «оценить» сложную социальную фактуру «ситуации», в которой «применяются» их «программы» (применяются автоматически, но ни в коем случае не «автономно» в истинном смысле этого слова). Во-вторых, алгоритмы ИИ неспособны самостоятельно определить и тем более доказать *обязательность* выполняемых ими законов, так как являются гетерономными получателями «команд» (то есть выполняемых «правил», которые они «выполняют»). Интересно толкование последнего (и весьма существенного) различия между алгоритмической «инструментальной рациональностью» и человеческим практическим разумом, данное в предпринятом А. Вудом анализе кантовского тезиса о том, что моральный субъект является *Selbstzweck*, то есть «самоцелью».

4.3. Аллен Вуд о «человечестве как цели самой по себе»

Подробно анализируя вторую формулировку «морального закона» Канта из «Основоположения…», которая гласит: «Поступай так, чтобы ты никогда не относился к человечеству, как в твоем лице, так и в лице всякого другого, только как к средству, но всегда в то же время и как к цели» (AA 04, S. 429; Кант, 1997, с. 169), американский кантовед А. Вуд справедливо пишет, что «идея человеческого достоинства, лежащая в основе [этой формулы], — это... кантов-

Read correctly, Kant's explorations of the *complexity* of human deliberation can thus indeed prove very helpful for a "Digital Humanism" which is aware of, and tries to rethink and defend, the difference between full human action and its (partial) digital simulations. AI algorithms are, in at least two substantial respects, unable to "autonomously perform" intelligent moral action: they are unable, firstly, to "assess" *fully* the complex social texture of the "situation" in which they "apply" their "programmes" (automatically, but nowhere in a genuine sense "autonomously"); and they are, secondly, unable to find out for themselves, let alone affirm, the *binding quality* of the laws that they implement, since they are heteronomous recipients of "orders" (i. e. of "rules" which they "execute"). This substantial second difference between algorithmic "instrumental rationality" and human practical reason is explored, in interesting ways, in Allen Wood's analysis of Kant's thesis that the moral subject is *Selbstzweck*, an "end in itself".

4.3. Allen Wood on "Humanity as End in itself"

In his elaborate analysis of Kant's second formulation of the "moral law" in *Groundwork of the Metaphysics of Morals*, "So act that you use humanity, whether in your own person or in the person of any other, always at the same time as an end and never merely as a means" (*GMS*, AA 04, p. 429; Kant, 1996, p. 80), the American Kant scholar Allen Wood (1998, p. 165) rightly insists that "the idea of human dignity, which grounds [this formula] is […] the Kantian principle which arguably has the greatest resonance with our culture's moral consciousness." It is the obvious point of ref-

ский принцип, который, возможно, в наибольшей степени резонирует с моральным сознанием нашей культуры» (Wood, 1998, p. 165). Это соображение — очевидная отправная точка для попытки полной философской реконструкции сути автономного человеческого действия. Вуд утверждает, что «формула человечности как цель сама по себе» может «многое рассказать [философам] об этической теории Канта», поскольку она представляет этику (вопреки распространенному неверному толкованию) «как основанную на цели», «характер которой четко отличает теорию Канта от всех форм консеквенциализма» (Ibid., p. 166) и тем самым, можно было бы добавить, от *инструменталистского рацио* отношений «средство — цель», характеризующего алгоритмические программы.

Что же именно означает фраза «люди — это цели сами по себе»? Вуд указывает на особое свойство этой цели, рассматривая различные сформулированные Кантом концепции «ценности» и критикуя стандартные аналитические характеристики кантовской этики:

> Тем, кто считает, что этическая теория Канта является *«деонтологической»*, будучи теорией, рассматривающей моральные принципы *независимо от цели, достигаемой путем следования им*, будет любопытно узнать, что Кант явным образом отвергает подобную позицию[19], а также осознать, что именно это неприятие лежит в основе приведенного в «Основоположении…» аргумента, что рациональную волю можно мотивировать к подчинению категорическому императиву только *особым видом* цели (Ibid., p. 168).

Согласно Вуду, Кант проводит различие между человечеством как «целью самой по

[19] Ср. «Основоположение к метафизике нравов», где Кант пишет: «Но, положим, что существует нечто, *чье бытие само по себе* имеет абсолютную ценность, что, *как цель сама по себе*, могло бы быть основанием определенных законов; тогда в этой цели и только в ней одной могло бы заключаться основание возможного категорического императива, т. е. практического закона» (AA 04, S. 428; Кант, 1997, с. 165).

erence for the attempt to reconstruct fully, in philosophical terms, the core of autonomous human action. Wood (1998, p. 166) points out that the "formula of humanity as an end in itself" has "much to teach [philosophers] about Kantian ethical theory itself", since it presents ethics (in contrast to common misunderstandings of that theory) "as based on an end" — an end, however, "whose character clearly distinguishes Kant's theory from all forms of consequentialism" — and thereby, we might add, also from the *instrumentalist ratio* of purposive "means-end" relations that characterises algorithmic programmes.

What exactly does it mean, then, that "human beings are ends-in-themselves"? Wood elucidates the special quality of this "end" by discussing various conceptions of "value" that Kant introduces, and in doing so he criticises the standard analytic characterisation of Kant's ethics:

> To those who think of Kant's ethical theory as '*deontological*' in the sense that it is a theory which regards moral principles *independent[ly] of any end served in following them* it should be enlightening to find Kant explicitly rejecting any such position,[18] and to realize that it is this rejection which lies behind the *Groundwork*'s argument that a rational will can be motivated to obey the categorical imperative only by *a distinctive kind* of end (Wood, 1998, p. 168).

Wood points out that Kant distinguishes humanity as an "end in itself" from three

[18] See *Groundwork of the Metaphysics of Morals*, where Kant writes: "But suppose there were something *the existence of which in itself* has an absolute worth, something which *as an end in itself* could be a ground of determinate laws; then in it, and in it alone, would lie the ground of a possible categorical imperative, that is, of a practical law" (*GMS*, AA 04, p. 428; Kant, 1996, p. 78).

себе» и тремя другими понятиями «цели». Во-первых, он противопоставляет ее «относительным целям», которые «являются основанием только гипотетических императивов» (AA 04, S. 428; Кант, 1997, с. 165). Во-вторых, он противополагает ее как «самосущую цель» «подлежащей осуществлению цели» (AA 04, S. 437; Кант, 1997, с. 196). Можно привести следующий пример цели, не являющейся самосущей: «когда мы строим дом, наше действие направлено на то, чтобы этот дом возник, а ценность дома для нас — это причина, по которой мы его строим» (Wood, 1998, p. 169). В-третьих, Кант отличает «цель саму по себе» от «цели, имеющей лишь *относительную ценность*, или *цену*, ценность которой может быть соотнесена с ценностью чего-то другого и может быть рационально принесена в жертву ради получения чего-то другого эквивалентной или большей ценности» (Ibid., p. 170—171).

В отличие от этих *трех ограниченных* понятий «цели», «человечество как цель сама по себе» — это «цель, имеющая абсолютную ценность» или (в терминологии Канта) «достоинство», то есть ее ценность нельзя сравнить, обменять, компенсировать или заменить другой ценностью (AA 04, S. 434; Кант, 1997, с. 187). «Говоря, что [человечество] является целью самой по себе», Кант имеет в виду «способность [человеческих существ] ставить цели и выбирать средства для их достижения» (Wood, 1998, p. 197). Эта способность, очевидно, включает в себя (как подструктуру) «"техническую склонность", которую слагают все наши приобретенные навыки и способности к размышлению, используемые нами для произвольных целей» (Ibid., p. 172). Однако эта подструктура ни при каких условиях не является самодостаточной; она зиждется на «способности [человеческой личности] быть самозаконодательствующей» (Ibid.). Следовательно, ни одна структура ИИ, ни один имитирующий человека «бот» не обладает такой способностью в прямом смысле.

other notions of "end": He firstly contrasts it with a "relative end", which "could only ground hypothetical imperatives" (*GMS*, AA 04, p. 428; Kant, 1996, p. 78). He secondly opposes it — as a "self-sufficient end" (*GMS*, AA 04, p. 437; Kant, 1996, p. 86) — to an "end to be effected" (*ibid*.). An example of such a *non-self-sufficient* end would be: "when we build a house, our action is for the sake of bringing the house into being, and the house's value to us is the reason why we build it" (Wood, 1998, p. 169). And Kant thirdly distinguishes the "end in itself" from an "end with *only relative* worth, or *price*, whose value can be measured against the value of something else and can be rationally sacrificed to obtain something else of equivalent or greater worth" (*ibid*., pp. 170-171).

In contrast to these *three limited* notions of an "end", "humanity as end in itself" is an "*end with absolute worth*" or (as Kant also says) "*dignity*", something whose value cannot be compared to, traded off against, compensated for, or replaced by any other value (*GMS*, AA 04, p. 434; Kant, 1996, p. 84). What Kant refers to "when he says [humanity] is an end in itself", Wood (1998, p. 197) concludes "is the capacity [of human beings] to set ends and choose means to them." This capacity obviously includes (but *as a mere sub-structure*) "the 'technical predisposition', which includes all our learned skills and deliberative abilities used for arbitrary ends" (*ibid*., p. 172). This sub-structure is *nowhere self-sufficient*, however; it is based on the "capacity [of any human person] to be self-legislative" (*ibid*.). No AI structure, no person-simulating "bot" has in any real sense this ability, we might conclude.

5. Заключение и выводы: необходимость критической теории искусственного интеллекта: Беккер и Зойберт о (неверной) интерпретации «автономии» в «цифровую эпоху»

Самозаконодательство, или автономия, — ключевая идея кантовского морального мотива, восприимчивого к герменевтике. Правильно понятая (и изложенная в полном объеме) кантовская концепция практического разума может стать философским основанием «цифрового гуманизма», отвергающего постгуманистическую фантазию о желательности «технологической сингулярности», в направлении которой и должен развиваться процесс цифровизации.

Тем не менее считается, что идее основанного на этике общественно-политического контроля над ИИ (критический проект «цифрового гуманизма») угрожают симулякры автономии, продвигаемые и навязываемые с помощью цифровых технологий. В недавнем эссе «Самоуничтожение автономии» Карлос Беккер и Сандра Зойберт отмечают, что современная общественно-политическая ситуация осложняется по всему миру тем, что в нашу цифровую эпоху даже возможные *локусы* ответственной рефлексии («автономия», «индивидуальность», «подлинность») становятся мишенью и подрываются культурной индустрией, через цифровых «инфлюенсеров» и поддерживаемых ИИ политических акторов массово распространяющей *симулякры* «подлинной, индивидуализированной свободы» и продвигающей усиливающие нарциссизм заменители автономии, которые разрушают то, что они якобы дают: критическую рефлексию и самоопределение. Поэтому, как пишут Беккер и Зойберт, полузабытая ранняя классика критической теории, например «Диалектика Просвещения» Хоркхаймера и Адорно, приобретает сегодня неожиданную актуальность. Проница-

5. Coda and Conclusion: The Need for a Critical Theory of Artificial Intelligence: Becker and Seubert on the (Mis-)Reading of "Autonomy" in a "Digital Age"

Self-legislation, "autonomy", is thus the core idea of Kant's hermeneutics-sensitive moral reason. Rightly understood (and fully spelled out), Kant's concept of practical reason is able to provide the philosophical basis for a "digital humanism" which rejects the post-humanistic fantasy that AI "singularity" is the desirable direction in which the process of digitalisation should develop.

However, the idea of an ethics-based, socio-political control of AI — the critical project of a "digital humanism" — is today, as many claim, endangered, by digitally promoted and enforced *simulacra* of autonomy. In a recent essay headed "The Self-Endangerment of Autonomy" Carlos Becker and Sandra Seubert point out that contemporary socio-political situations are complicated, worldwide, by the fact that in our digital age even the possible *loci* of responsible reflection — "autonomy", "individuality" and "authenticity" — are targeted and subverted by a cultural industry that, via digital "influencers" and AI-supported political *acteurs*, mass-distributes *simulacra* of "authentic, individualised freedom", and promotes narcissism-enhancing substitutes for autonomy which destroy what they claim to offer: critical reflection and self-determination. Hence, Becker and Seubert write, a much neglected early classic of Critical Theory, Horkheimer and Adorno's *Dialectic of Enlightenment*, has today gained unexpected relevance. The sharp-sighted reflections on the economy-driven production of "(quasi)-subjective needs" in the chapter "The Culture Industry: Enlighten-

тельная рефлексия о создании экономическими средствами «(квази)субъективных потребностей», изложенная Хоркхаймером и Адорно в главе «Культуриндустрия. Просвещение как обман масс» упомянутой книги (Хоркхаймер, Адорно, 1997, с. 149—209), способна внести существенный вклад в анализ социальных патологий, сопровождающих все более активную «коммодификацию и овеществление частной жизни» (Becker, Seubert, 2020, S. 230—231).

Важность подобного анализа современных цифровых форм (квази)«подлинности» подчеркивает и специалист по медиа Андреас Зудман, один из редакторов новой серии выходящих в издательстве *transcript* книг по ИИ. В частности, от отмечает, что критика «со времен Канта означала постановку под сомнение явлений в части их функционирования» (Sudmann, 2019, p. 16—17). Сегодня претендующий на глубину анализ общественно-политического влияния ИИ должен не только оценивать достоинства искусственного интеллекта, но и детально исследовать «создание иллюзий и ложного сознания» оцифрованной коммуникацией. Такой анализ, пишет Зудман, «[в доцифровые времена] был целью Адорно и Хоркхаймера, например в их критическом анализе, представленном в "Диалектике Просвещения"» (Ibid., p. 17). Следовательно, «цифровой гуманизм», с философской позиции исследующий достоинства и ограничения ИИ, должен, настаивая на значительности роли, которую *практический разум (в кантовском смысле)* продолжает играть в цифровую эпоху, вобрать в себя стимул, заключенный в критической теории Хоркхаймера и Адорно.

Список литературы

Кант И. Основоположение к метафизике нравов // Соч. на нем. и рус. яз. М. : Московский философский фонд, 1997. Т. 3. С. 39—275.

Сёрл Дж. Сознание, мозг и программы // Аналитическая философия: становление и развитие / под ред. А. Ф. Грязнова. М. : Дом интеллектуальной книги, 1998. С. 376—400.

ment as Mass Deception" in Horkheimer and Adorno's book (2016, pp. 120-167) are able to provide a significant contribution to an analysis of the social pathologies that accompany today's ever-increasing "commodification and reification of privacy" (Becker and Seubert, 2020, pp. 230-231).

The importance of these analyses of today's digitally produced modes of (quasi-)"authenticity" has recently also been emphasised by the media specialist Andreas Sudmann (2019, pp. 16-17), one of the editors of the new *transcript* book series on *AI*, who points out that critique "since Kant has meant questioning phenomena with regard to their functioning". Today, any in-depth analysis of the socio-political influence of AI has not only to assess the merits of Artificial Intelligence, but also to include a careful analysis of the "production of illusion and false consciousness" by digitised communication — an analysis to which, as Sudmann points out, "Adorno and Horkheimer [in pre-digital times] felt deeply committed [to] in their critical analysis of the *Dialectic of Enlightenment*" (*ibid.*, p. 17) A "Digital Humanism" that philosophically explores the merits and limits of AI has thus to integrate the stimulus provided by Horkheimer and Adorno's Critical Theory into its philosophical defence of the core role that *practical reason (in Kant's sense)* continues to play in our digital age.

References

Becker, C. and Seubert, S., 2020. Die Selbstgefährdung der Autonomie. Eckpunkte einer Kritischen Theorie der Privatheit im digitalen Zeitalter. In: P. Kruse and S. Müller-Mall, eds. 2020. *Digitale Transformation der Öffentlichkeit*. Weilerswist: Velbrück Wissenschaft, pp. 229-261.

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

Хоркхаймер М., Адорно Т. Диалектика Просвещения. Философские фрагменты / пер. с нем. М. Кузнецова. М. ; СПб. : Медиум ; Ювента, 1997.

Becker C., Seubert S. Die Selbstgefährdung der Autonomie. Eckpunkte einer Kritischen Theorie der Privatheit im digitalen Zeitalter // Digitale Transformation der Öffentlichkeit / hrsg. von P. Kruse, S. Müller-Mall. Weilerswist : Velbrück Wissenschaft, 2020. S. 229—261.

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

Dreyfus H., Dreyfus S. Coping with Change: Why Computers Can't and Humans Can // Wo steht die Analytische Philosophie heute? / hrsg. von L. Nagl, R. Heinrich. Wien : Oldenbourg, 1986. P. 150—170.

Dreyfus H., Taylor C. Retrieving Realism. Cambridge, MA ; L. : Harvard University Press, 2015.

Forrest K. When Machines Can Be Judge, Jury, and Executioner: Justice in the Age of Artificial Intelligence. New Jersey : World Scientific Publishing House, 2021.

Habermas J. Auch eine Geschichte der Philosophie. Berlin : Suhrkamp, 2019. Bd. 2 : Vernünftige Freiheit: Spuren des Diskurses über Glauben und Wissen.

Herman B. The Practice of Moral Judgment. Cambridge, MA ; L. : Harvard University Press, 1993.

Hochreiter S. "The Algorithm Can Learn Anything — Good Things as Well as Bad Things" // Digital Transformation and Ethics / ed. by M. Hengstschläger. Salzburg ; Munich : Ecowin Verlag, 2020. P. 362—375.

Klingel A., Kraft T., Zweig A. Potential Best Practice Approaches in the Use of an Algorithmic Decision-Making Systems with the Example of the AMS Algorithm // Digital Transformation and Ethics / ed. by M. Hengstschläger. Salzburg ; Munich : Ecowin Verlag, 2020. P. 178—196.

Kurzweil R. The Singularity is Near. L. : Viking Penguin, 2005.

Marantz A. Antisocial. Online Extremists, Techno-Utopian, and the Hijacking of the American Conversation. N. Y. : Viking Penguin, 2019.

Mersch D. Ideen zu einer Kritik 'algorithmischer' Rationalität // Deutsche Zeitschrift für Philosophie. 2019. Bd. 67, № 5. S. 851—873.

Mitchell M. Artificial Intelligence. A Guide for Thinking Humans. N. Y. : Farrar, Straus and Giroux, 2019.

Nagl L. Gesellschaft und Autonomie. Historisch-systematische Studien zur Entwicklungsgeschichte der Sozialtheorie von Hegel bis Habermas. Wien : Verlag der Österreichischen Akademie der Wissenschaften, 1983.

Dreyfus, H. and Dreyfus, S., 1986. Coping with Change: Why Computers Can't and Humans Can. In: L. Nagl and R. Heinrich, 1986. *Wo steht die Analytische Philosophie heute?* Vienna: Oldenbourg, pp. 150-170.

Dreyfus, H. and Taylor, C., 2015. *Retrieving Realism*. Cambridge, MA and London: Harvard University Press.

Forrest, K., 2021. *When Machines Can Be Judge, Jury, and Executioner: Justice in the Age of Artificial Intelligence*. New Jersey: World Scientific Publishing House.

Habermas, J., 2019. *Auch eine Geschichte der Philosophie. Volume 2: Vernünftige Freiheit: Spuren des Diskurses über Glauben und Wissen*. Berlin: Suhrkamp.

Herman, B., 1993. *The Practice of Moral Judgment*. Cambridge, MA & London: Harvard University Press.

Hochreiter, S., 2020. The Algorithm Can Learn Anything — Good Things as Well as Bad Things. In: M. Hengstschläger, ed. 2020. *Digital Transformation and Ethics*. Salzburg & Munich: Ecowin Verlag, pp. 362-375.

Horkheimer, M. and Adorno, T., 2016. *Dialectic of Enlightenment*. London & New York: Verso.

Kant, I., 1996. Groundwork of the Metaphysics of Morals. In: I. Kant, 1996. *Practical Philosophy*. Edited by M. Gregor and A. Wood. Cambridge: Cambridge University Press, pp. 41-108.

Klingel, A., Kraft, T. and Zweig, A., 2020. Potential Best Practice Approaches in the Use of an Algorithmic Decision-Making Systems with the Example of the AMS [Austrian Arbeitsmarktservice] Algorithm. In: M. Hengstschläger, ed. 2020. *Digital Transformation and Ethics*. Salzburg & Munich: Ecowin Verlag, pp. 178-196.

Kurzweil, R., 2005. *The Singularity is Near*. London: Viking Penguin.

Marantz, A., 2019. *Antisocial. Online Extremists, Techno-Utopian, and the Hijacking of the American Conversation*. New York: Viking Penguin.

Mersch, D., 2019. Ideen zu einer Kritik 'algorithmischer' Rationalität. *Deutsche Zeitschrift für Philosophie*, 67(5), pp. 851-873.

Mitchell, M., 2019. *Artificial Intelligence. A Guide for Thinking Humans*. New York: Farrar, Straus and Giroux.

Nagl, L., 1983. *Gesellschaft und Autonomie. Historisch-systematische Studien zur Entwicklungsgeschichte der Sozialtheorie von Hegel bis Habermas*. Vienna: Verlag der Österreichischen Akademie der Wissenschaften.

Nagl, L., 2018. What is it to be a Human Being? Charles Taylor on 'the Full Shape of the Human Linguistic Capacity'. In: B. Buchhammer, ed. 2018. *Re-Learning to be Human in Global Times: Challenges and*

Nagl L. What is it to be a Human Being? Charles Taylor on 'the Full Shape of the Human Linguistic Capacity' // Re-Learning to be Human in Global Times: Challenges and Opportunities from the Perspective of Contemporary Philosophy of Religion / ed. by B. Buchhammer. Washington : The Council for Research in Values and Philosophy, 2018. P. 117—136.

Nagl L. Merits and Limits of AI // Artificial Intelligence and Human Enhancement. Affirmative and Critical Approaches in the Humanities / ed. by H. Nagl-Docekal, W. Zacharasiewicz. Berlin : De Gruyter, 2022. P. 33—50.

Nida-Rümelin J., Weidenfeld N. Digitaler Humanismus. Eine Ethik für das Zeitalter der Künstlichen Intelligenz. München : Piper, 2018.

Nida-Rümelin J. Structural Rationality and Other Essays on Practical Reason. Berlin ; Dordrecht ; Heidelberg ; N. Y. : Springer, 2019.

Putnam H. Brains in a Vat // Putnam H. Reason, Truth and History. Cambridge : Cambridge University Press, 1981. P. 1—21.

Putnam H. The Project of Artificial Intelligence // Putnam H. Renewing Philosophy. Cambridge, MA : Harvard University Press, 1992. P. 1—18.

Spiekermann S. On the Difference between Artificial and Human Intelligence and the Ethical Implications of Their Confusion // Digital Transformation and Ethics / ed. by M. Hengstschläger. Salzburg ; Munich : Ecowin Verlag, 2020. P. 88—117.

Sudmann A. The Democratization of Artificial Intelligence. Net Politics in the Era of Learning Algorithms // The Democratization of Artificial Intelligence. Net Politics in the Era of Learning Algorithms / ed. by A. Sudmann. Bielefeld : transcript, 2019. P. 9—31.

Taylor C. The Language Animal. The Full Shape of Human Linguistic Capacity. Cambridge, MA ; L. : The Belknap Press of Harvard University Press, 2016.

Taylor C., Nanz P., Beaubien-Taylor M. Reconstructing Democracy. How Citizens are Building from the Ground Up. Cambridge, MA ; L. : Harvard University Press, 2020.

Werthner H. The Vienna Manifesto on Digital Humanism // Digital Transformation and Ethics / ed. by M. Hengstschläger. Salzburg ; Munich : Ecowin Verlag, 2020. P. 350—385.

Wood A. Humanity as End in Itself // Kant's Groundwork of the Metaphysics of Morals. Critical Essays / ed. by P. Guyer. Lanham ; Boulder ; N. Y. ; Oxford : Rowman and Littlefield Publishers, 1998. P. 165—187.

Opportunities from the Perspective of Contemporary Philosophy of Religion. Washington, DC: The Council for Research in Values and Philosophy, pp. 117-136.

Nagl, L., 2022. Merits and Limits of AI. In: H. Nagl-Docekal and W. Zacharasiewicz, eds., 2022. *Artificial Intelligence and Human Enhancement. Affirmative and Critical Approaches in the Humanities*. Berlin: De Gruyter, pp. 33-50.

Nida-Rümelin, J. and Weidenfeld, N., 2018. *Digitaler Humanismus. Eine Ethik für das Zeitalter der Künstlichen Intelligenz*. München: Piper.

Nida-Rümelin, J., 2019. *Structural Rationality and Other Essays on Practical Reason*. Berlin, Dordrecht, Heidelberg & New York: Springer.

Putnam, H., 1981. Brains in a Vat. In: H. Putnam, 1981. *Reason, Truth and History*. Cambridge: Cambridge University Press, pp. 1-21.

Putnam, H., 1992. The Project of Artificial Intelligence. In: H. Putnam, 1992. *Renewing Philosophy*. Cambridge, MA: Harvard University Press, pp. 1-18.

Searle, J., 1987. Minds, Brains, and Programs. In: R. Born, ed. 1987. *Artificial Intelligence. The Case Against*. London & Sydney: Croom Helm, pp. 18-40.

Spiekermann, S., 2020. On the Difference between Artificial and Human Intelligence and the Ethical Implications of Their Confusion. In: M. Hengstschläger, ed. 2020. *Digital Transformation and Ethics*. Salzburg & Munich: Ecowin Verlag, pp. 88-117.

Sudmann, A., 2019. The Democratization of Artificial Intelligence. Net Politics in the Era of Learning Algorithms. In: A. Sudmann, ed. 2019. *The Democratization of Artificial Intelligence. Net Politics in the Era of Learning Algorithms*. Bielefeld: transcript Verlag, pp. 9-31.

Taylor, C., 2016. *The Language Animal. The Full Shape of Human Linguistic Capacity*. Cambridge, MA & London: The Belknap Press of Harvard University Press.

Taylor, C., Nanz, P. and Beaubien-Taylor, M., 2020. *Reconstructing Democracy. How Citizens are Building from the Ground Up*. Cambridge, MA & London: Harvard University Press.

Werthner, H., 2020. The Vienna Manifesto on Digital Humanism. In: M. Hengstschläger, ed. 2020. *Digital Transformation and Ethics*. Salzburg & Munich: Ecowin Verlag, pp. 350-385.

Wood, A., 1998. Humanity as End in Itself. In: P. Guyer, ed. 1998. *Kant's Groundwork of the Metaphysics of Morals. Critical Essays*. Lanham, Boulder, New York & Oxford: Rowman and Littlefield Publishers, pp. 165-187.

Об авторе

Людвиг ***Нагль***, доктор философии, профессор, Институт философии, Университет Вены, Австрия.

E-mail: ludwig.nagl@univie.ac.at

Для цитирования:

Нагль Л. Цифровые технологии: размышления о различии между инструментальной рациональностью и практическим разумом // Кантовский сборник. 2022. Т. 41, № 1. С. 60—88.

doi: 10.5922/0207-6918-2022-1-3

The author

Prof. Dr Ludwig ***Nagl***, Institute of Philosophy, University of Vienna, Austria.

E-mail: ludwig.nagl@univie.ac.at

To cite this article:

Nagl, L. 2022. Digital Technology: Reflections on the Difference between Instrumental Rationality and Practical Reason. *Kantian Journal*, 41(1), pp. 60-88.

http://dx.doi.org/10.5922/0207-6918-2022-1-3