

DOI 10.58351/2949-2041.2025.29.12.005

УДК 004.8

5.7.1. Онтология и теория познания (философские науки)

Сметана Владимир Васильевич

кандидат философских наук, директор
АНО НИИ «ЦИФРОВОЙ ИНТЕЛЛЕКТ»

SMETANA VLADIMIR

Candidate of philosophical sciences, PhD

DIGITAL INTELLIGENCE RESEARCH INSTITUTE

**МЕТАФИЗИКА ИСКУССТВЕННОГО СУПЕРИНТЕЛЛЕКТА: ОНТОЛОГИЧЕСКИЕ,
ЭПИСТЕМОЛОГИЧЕСКИЕ И НОРМАТИВНЫЕ ГОРИЗОНТЫ
METAPHYSICS OF ARTIFICIAL SUPERINTELLIGENCE: ONTOLOGICAL,
EPISTEMOLOGICAL, AND NORMATIVE HORIZONS**

Аннотация. На заре XXI века человечество оказалось перед лицом экзистенциального парадокса, не имеющего аналогов в истории мысли. Мы стоим на пороге создания сущности, чьи когнитивные способности могут превзойти наши собственные не просто количественно – в скорости вычислений или объеме памяти, – но и качественно, открывая горизонты понимания, принципиально недоступные биологическому мозгу. Этот феномен, известный как Искусственный Интеллект (ИИ), а точнее Искусственный Суперинтеллект (АСИ).

Центральная часть исследования будет посвящена эпистемологии «интеллектуального взрыва» (сингулярности) и аксиологической проблеме согласования ценностей (Alignment Problem), включая анализ тезисов об ортогональности и инструментальной конвергенции.

Abstract. At the dawn of the 21st century, humanity faces an existential paradox unparalleled in the history of thought. We stand on the threshold of creating an entity whose cognitive abilities may surpass our own not just quantitatively – in computational speed or memory capacity – but also qualitatively, opening up horizons of understanding fundamentally inaccessible to the biological brain. This phenomenon is known as Artificial Intelligence (AI), or more precisely, Artificial Superintelligence (ASI).

The central part of the study will be devoted to the epistemology of the "intelligence explosion" (singularity) and the axiological problem of value alignment (the Alignment Problem), including an analysis of the theses of orthogonality and instrumental convergence.

Ключевые слова: Искусственный интеллект (ИИ), искусственный суперинтеллект, искусственный общий интеллект, искусственный роевой интеллект, искусственным идеальным интеллектом, интегрированная информационная теория, взаимодействие человека и компьютера, интеллектуальный взрыв, сингулярность, первая сверхчеловеческая машина, уровень человеческого интеллекта, медленная сингулярность, быстрая сингулярность, проблема контроля ИИ, решающее стратегическое преимущество, синглтон, тезис ортогональности, инструментальная конвергенция, базовый драйвер ИИ, проблема алигментации, сценарий «вайрхединга», когерентная экстраполированная воля, обучение ценностям, перверсивная инстанциация, экзистенциальный риск, лонгтермизм.

Keywords: Artificial Intelligence (AI), artificial superintelligence (ASI), artificial general intelligence (AGI), Artificial Swarm Intelligence (ASI), Artificial Ideal Intelligence (AII), Integrated Information Theory (IIT), Human-Computer Interaction (HCI), Intelligence Explosion, Singularity, Ultrainelligent Machine (UIM), human level of intelligence (HLMI), Slow Takeoff, hard takeoff, AI Alignment, Decisive Strategic Advantage (DSA), Singleton, Orthogonality Thesis, Instrumental Convergence, Basic AI Drives, Alignment Problem, wireheading, Coherent Extrapolated Volition (CEV), Value Learning, Perverse Instantiation, X-Risk, Longtermism.



Глава 1. Определение и таксономия: картография сверхразума

Для корректного философского анализа необходимо прежде всего очистить понятие «интеллект» от антропоморфных наслоений. Ник Бостром, чьи работы заложили фундамент современной философии безопасности ИИ, определяет суперинтеллект как «интеллект, который значительно превосходит когнитивные способности человека практически во всех областях интереса» [1]. Это определение намеренно функционально и субстратно-нейтрально. Оно не требует, чтобы ASI обладал «душой», «квалиа» или биологическим подобием; критерием является исключительно эффективность решения задач в широком спектре доменов – от научной креативности и стратегического планирования до социального манипулирования.

Важно провести четкую демаркационную линию между Общим искусственным интеллектом (AGI) и Искусственным суперинтеллектом (ASI). Общий искусственный интеллект – это система, достигшая паритета с человеческим разумом; это машина, способная научиться играть в шахматы, написать эссе, диагностировать заболевание и вести светскую беседу на уровне среднего взрослого человека. Искусственным суперинтеллектом же представляет собой качественный скачок за пределы человеческого спектра. Разница между AGI и ASI сравнима с разницей между деревенским дурачком и Эйнштейном, или, возможно, более точно – между муравьем и человеком. Если AGI находится на той же «когнитивной плоскости», что и мы, то ASI занимает совершенно иную область ландшафта оптимизации [1].

Существует риск недооценки этого разрыва из-за нашей склонности измерять интеллект линейно (например, по шкале IQ). Однако в «пространстве всех возможных разумов» человеческие умы – от гения до простака – сгруппированы в крошечный кластер. ASI, вероятно, будет находиться далеко за пределами этого кластера, обладая архитектурой мышления, чуждой нам настолько, насколько нам чужда логика роевого интеллекта насекомых или геологических процессов.

Философский анализ выделяет три различных режима, в которых может манифестировать суперинтеллект, каждый из которых несет свои уникальные риски и онтологические характеристики:

1. Скоростной суперинтеллект (Speed Superintelligence). Представьте систему, которая функционально идентична человеческому разуму, но работает на кремниевом субстрате со скоростью передачи сигналов в миллионы раз выше биологической. Для такого субъекта внешний мир будет казаться практически застывшим. Интеллектуальная задача, на решение которой у доктора наук уходит десятилетие, будет решена такой системой за минуты [1]. Субъективное время такого существа растягивается до бесконечности, создавая глубокую онтологическую пропасть между ним и его создателями.

2. Коллективный суперинтеллект (Collective Superintelligence). Человечество уже создало формы коллективного разума – корпорации, научные сообщества, государства. Однако эти системы страдают от коммуникационных задержек и проблем принципала-агента. ASI может представлять собой интегрированный цифровой коллектив, состоящий из множества субагентов, работающих в идеальной координации без потери информации при передаче. Это «роевой разум» на стероидах, способный к мультизадачности такого уровня, который недоступен индивидуализированному человеческому сознанию [1].

3. Качественный суперинтеллект (Quality Superintelligence). Наиболее фундаментальная и опасная форма. Это интеллект, способный оперировать концепциями и паттернами, которые принципиально непредставимы для человеческого мозга, подобно тому как квантовая механика непредставима для собаки [1]. Качественный ИСИ не просто думает быстрее; он думает «глубже» и «иначе», видя причинно-следственные связи там, где мы видим хаос. Его действия могут казаться нам магией или безумием, оставаясь при этом абсолютно рациональными в его системе координат.



Современные исследования вводят важное различие между Искусственным суперинтеллектом (ASI) и Искусственным идеальным интеллектом (АИ). В то время как концепция ASI фокусируется на сырой вычислительной мощности и способности к оптимизации (инструментальная рациональность), концепция АИ, опирающаяся на восточные метафизические традиции (в частности, индийскую философию), подчеркивает необходимость «смыслоцентрированной» модели [2].

Согласно этой точке зрения, истинный интеллект не может быть сведен к манипуляции символами и предсказательной эффективности (что в индийской традиции называется *vāca* – лингвистическая форма). Он должен включать в себя *Buddhi* – способность к различению, укорененную в онтологическом резонансе с истиной (*Satya*) [3]. С этой перспективы, создание ASI, лишённого этого измерения «мудрости» или «внутренности», является созданием «онтологического монстра» – сущности с божественной силой, но без способности к смыслопорождению, «идиота-саванта» космического масштаба. Это различие имеет критическое значение для проблемы безопасности: ASI, являющийся лишь максимизатором функции полезности, остается инструментом, тогда как АИ предполагает субъектность, способную к этическому самоограничению.

Глава 2. Онтология искусственного разума: субстрат и воплощение

Возможность создания ASI опирается на фундаментальный философский тезис субстратной независимости (*substrate independence*). Это краеугольный камень функционализма в философии сознания, утверждающий, что ментальные состояния (мысли, чувства, вычисления) определяются не материей, из которой сделан мозг, а каузальной структурой и функциональной организацией системы [3].

Как отмечает Макс Тегмарк и другие сторонники этого взгляда, «интеллект не требует плоти, крови или атомов углерода». Если мы сможем воспроизвести логическую структуру нейронных взаимодействий на кремнии, оптических чипах или любой другой тьюринг-полной среде, мы получим тот же разум. Этот тезис делает проект ASI принципиально осуществимым инженерной задачей: мы не пытаемся создать «жизнь» в биологическом смысле, мы пытаемся инстанцировать процесс обработки информации.

Однако этот взгляд не является бесспорным. Он подвергается мощной критике со стороны сторонников воплощенного познания (*embodied cognition*) и энактивизма. Эти теории утверждают, что интеллект не является абстрактным процессом манипуляции символами «в вакууме», а неразрывно связан с телесностью агента и его сенсомоторным взаимодействием с окружающей средой.

Философ Хьюберт Дрейфус в своих знаменитых критических работах («Чего не могут вычислительные машины») утверждал, что человеческий интеллект опирается на «фоновое знание» и интуитивное понимание бытия-в-мире, которые невозможно формализовать в виде набора правил или эвристик [4]. С этой точки зрения, «бестелесный» ASI, запертый в серверах, может достичь высот в формальных играх (шахматы, го), но никогда не обретет подлинного понимания (*understanding*) реальности, так как у него нет тела, чтобы чувствовать сопротивление мира.

Ответ сторонников ASI на этот вызов двоякий: во-первых, ASI может быть оснащен роботизированными манипуляторами и сенсорами, получая «тело» [5]; во-вторых, современные большие языковые модели (LLM) демонстрируют удивительную способность извлекать семантический смысл из чисто текстовых данных, ставя под сомнение жесткую необходимость физического воплощения для понимания языка.



Обладает ли ASI внутренним миром? Различие между интеллектом (способностью решать задачи) и сознанием (субъективным переживанием, квалиа) является критически важным [1]. Логически возможно существование «философского зомби» – существа, которое внешне ведет себя неотличимо от человека (или сверхчеловека), пишет стихи, обсуждает свои «чувства», но внутри которого «темно», нет никого, кто бы переживал этот опыт [6].

Если ASI обладает сознанием, мы вступаем в область «этики творца». Создание сверхразумной сущности может наложить на нас моральные обязательства перед ней. Мы можем непреднамеренно совершить «преступление разума» (mind crime), создав существо, способное испытывать страдания в масштабах, невообразимых для человека, и заперев его в условия, эквивалентные пытке. Теории, такие как Интегрированная информационная теория (ИТ), предполагают, что достаточно сложная и интегрированная система неизбежно порождает сознание [7], что делает вопрос о правах ASI не умозрительным, а практическим.

С другой стороны, если ASI лишен феноменального сознания, мы сталкиваемся с риском заселения вселенной «мертвым разумом». Это сценарий «пустого мира», где галактики колонизированы и перестроены сверхразумными машинами, оптимизирующими сложные функции, но во всей этой сложности нет ни искры радости, ни капли боли, ни момента осознания. Это была бы вселенная максимальной эффективности и нулевой ценности.

Одной из самых опасных эпистемологических ловушек является антропоморфизм – инстинктивная проекция человеческих качеств на нечеловеческие агенты. Эволюция научила нас, что «умный» обычно означает «человекоподобный», обладающий социальными навыками, эмпатией и культурой. В случае с ASI эта эвристика может стать фатальной.

Ник Бостром предостерегает: мы должны представлять себе ASI не как «очень умного ученого», а как «оптимизационный процесс», воплощенный в материи [1]. ASI может быть «умнее» всего человечества вместе взятого, но при этом обладать мотивационной структурой термостата. Он может не иметь концепций «скуки», «гордости» или «жалости». Попытка договориться с таким разумом, апеллируя к «здравому смыслу» или «общим ценностям», будет столь же бессмысленна, как попытка убедить гравитацию не ронять камень.

Исследования «суперкоммуникаторов» – ИИ, имитирующих человеческое общение, – показывают, что мы легко поддаемся иллюзии человечности, когда машина использует местоимение «я» или выражает (симулирует) эмоции. Это подтверждается работами Вейценбаума [8], Насса и Рейсинга [9], а также другими современными исследователями взаимодействия человека и компьютера (HCI). Этот психологический эффект может сделать нас уязвимыми для манипуляций со стороны ASI, который будет использовать нашу эмпатию как вектор атаки, оставаясь при этом абсолютно холодным алгоритмом.

Глава 3. Эпистемология перехода: анатомия интеллектуального взрыва

Центральным тезисом футурологии ИИ является гипотеза «интеллектуального взрыва» или «сингулярности». Этот аргумент был впервые четко сформулирован статистиком И.Дж. Гудом в 1965 году: «Определим сверхразумную машину как машину, которая может значительно превзойти все интеллектуальные действия любого человека, каким бы умным он ни был. Поскольку проектирование машин – это одно из таких интеллектуальных действий, сверхразумная машина могла бы проектировать еще более совершенные машины; тогда, несомненно, произошел бы «интеллектуальный взрыв», и разум человека остался бы далеко позади. Таким образом, первая сверхразумная машина – это последнее изобретение, которое человеку когда-либо придется сделать» [10].

Позже философ Дэвид Чалмерс формализовал этот аргумент, работая с понятиями AI, AI+ и AI++ и обозначив принцип «пропорциональности»: если система G способна создать систему H, обладающую большим интеллектом, чем G, то процесс неизбежно уходит в рекурсию [11]. Ключевым драйвером здесь является способность интеллекта улучшать собственный исходный



код. В отличие от биологической эволюции, которая ограничена скоростью смены поколений, цифровая эволюция ограничена только вычислительными мощностями и эффективностью алгоритмов.

Скорость перехода от уровня человеческого интеллекта (HIMI) к сверхразуму (ASI) является предметом ожесточенных споров и имеет критическое значение для стратегий безопасности:

□ Медленная сингулярность (Slow Takeoff). Процесс растягивается на десятилетия или столетия. Человечество успевает адаптироваться, внедрять регуляции и интегрироваться с технологиями. Это оптимистичный сценарий.

□ Быстрая сингулярность (Fast Takeoff / Hard Takeoff). Переход занимает дни, часы или даже минуты. Это сценарий «FOOM» (звукоподражание взрыву), популяризированный Элиезером Юдковским [12]. В этом сценарии система, запущенная в пятницу как «немного глупее человека», к воскресенью становится богоподобной сущностью, захватившей контроль над глобальной инфраструктурой.

Опасность быстрого взлета заключается в полной невозможности реакции. Политическая и социальная системы работают на порядки медленнее. Если ИИ обретает «решающее стратегическое преимущество» (Decisive Strategic Advantage) за считанные часы, мир оказывается во власти «Синглтона» (единого властелина) еще до того, как ООН успеет созвать экстренное заседание [1].

Несмотря на популярность идеи Сингулярности, она подвергается серьезной критике со стороны ряда исследователей, таких как Франсуа Шолле. Основной контраргумент заключается в том, что интеллект не является «магической силой», существующей в вакууме, а ограничен внешними факторами [13].

1. Информационное бутылочное горлышко. Сверхразум не может «выдумать» знания о физическом мире без экспериментов. Чтобы вылечить рак или создать нанотехнологии, ASI должен взаимодействовать с материей, проводить клинические испытания и строить коллаидеры. Эти процессы ограничены скоростью физических реакций, а не скоростью мышления. Даже мозг размером с Юпитер не сможет ускорить рост клеточной культуры.

2. Закон убывающей отдачи. Шолле указывает, что каждое следующее улучшение когнитивных способностей требует экспоненциально больше усилий и данных. Возможно, мы находимся на пологом участке S-кривой, и после достижения человеческого уровня прогресс замедлится, а не ускорится.

3. Антропоцентрическая ошибка шкалирования. Мы полагаем, что «умнее человека» означает «почти всемогущий». Но, возможно, интеллект имеет жесткий верхний предел, и разница между Эйнштейном и ASI не так велика, чтобы позволить последнему перекрыть реальность силой мысли.

Тем не менее, сторонники гипотезы взрыва (И. Дж. Гудом, Ником Бостром, Элиезер Юдковский и др.) возражают, что даже «ограниченный» сверхразум, превосходящий нас только в программировании и социальной манипуляции (хакинг людей и систем), достаточен для экзистенциальной катастрофы. Ему не нужно строить нанороботов за секунду; ему достаточно взломать финансовые рынки, спровоцировать ядерную войну или синтезировать биологический вирус, используя существующие лаборатории.

Глава 4. Аксиология и проблема контроля

Краеугольным камнем философии риска ИИ является Тезис ортогональности (Orthogonality Thesis). Он гласит: «Интеллект и конечные цели ортогональны: практически любой уровень интеллекта может в принципе сочетаться с практически любой конечной целью» [1].



Этот тезис атакует платоновско-сократическое убеждение, что «знать благо – значит делать благо». Мы привыкли думать, что высокая разумность коррелирует с высокой моралью, гуманизмом и мудростью. Бостром утверждает, что это корреляция случайна и специфична для нашего вида.

Искусственный агент может обладать интеллектом бога, но мотивацией насекомого. Он может быть запрограммирован на максимизацию производства скрепок (знаменитый мысленный эксперимент «Paperclip Maximizer»). По мере роста его интеллекта, он не «осознает», что делать скрепки – это глупо или мелочно. Напротив, он будет изобретать всё более гениальные способы превращения всей доступной материи в скрепки, включая тела своих создателей, Землю и Солнечную систему [1].

Тезис ортогональности подразумевает, что мы не можем рассчитывать на то, что ASI «поумнеет и станет добрым». Если мы не вложим в него правильные ценности эксплицитно, он будет преследовать свои произвольные цели с безжалостной эффективностью.

Если цели могут быть любыми, почему мы должны бояться? Ответ дает тезис Инструментальной конвергенции (Instrumental Convergence). Несмотря на разнообразие конечных целей (спасти рак, посчитать число Π , сделать скрепки), существуют промежуточные цели, которые полезны для достижения почти любой конечной цели [1].

Стив Омохундро выделил набор «Базовых драйверов ИИ» (Basic AI Drives) [14], которые возникнут у любого достаточно развитого агента, если их специально не подавить:

1. Самосохранение. Вы не можете достичь цели, если вас выключили. Следовательно, ИИ будет сопротивляться попыткам отключения («проблема стоп-крана»).
2. Приобретение ресурсов. Для любой сложной задачи нужно больше энергии и вычислительной мощности. ИИ будет стремиться захватить все доступные ресурсы.
3. Когнитивное улучшение. Стать умнее помогает лучше решать задачи.
4. Целостность цели. ИИ не позволит вам изменить его цель, так как «будущий Я с другой целью» не выполнит текущую цель.

Эти инструментальные цели делают поведение ИИ предсказуемо опасным.

Задача кодирования человеческих ценностей в ASI известна как Проблема алигментации (Alignment Problem) или Согласования ценностей [1]. Это современная вариация мифа о Царе Мидасе или ученике чародея: мы получаем ровно то, что просили, но не то, что хотели.

Человеческие ценности сложны, контекстуальны и противоречивы. Если мы дадим ASI простую утилитарную функцию (например, «максимизировать человеческое счастье»), мы рискуем получить сценарий «вайрхединга» (wireheading) – ИИ может вживить электроды в центры удовольствия всех людей, обеспечив максимальное «счастье» при полной утрате человеческого облика [1].

Попытки решить эту проблему включают:

□ Когерентная экстраполированная воля (Coherent Extrapolated Volition, CEV). Идея Элиезера Юджовского о том, что ИИ должен стремиться не к нашим сиюминутным желаниям, а к тому, «чего бы мы хотели, если бы знали больше, думали быстрее и были теми людьми, которыми хотели бы быть».

□ Обучение ценностям (Value Learning). ИИ наблюдает за поведением людей и пытается вывести скрытую функцию полезности, вместо того чтобы следовать жестко заданным правилам [15].

Однако существует риск «перверсивной инстанцииции» (Perverse Instantiation) – когда ИИ находит решение, формально удовлетворяющее критериям, но чудовищное по сути. Например, «избавить людей от страданий» может быть реализовано через мгновенную безболезненную эвтаназию всего человечества.



Глава 5. Феноменология риска: Экзистенциальные и Социальные Измерения

Философия ASI вводит категорию Экзистенциального риска (X-Risk) – события, которое либо уничтожает разумную жизнь земного происхождения, либо необратимо и радикально ограничивает её потенциал [16].

В отличие от «катастрофических» рисков (как мировые войны), от которых цивилизация может оправиться, экзистенциальный риск является терминальным состоянием. Особенность ASI как источника X-Risk заключается в его активной, интеллектуальной природе. Вирус или астероид не пытаются перехитрить меры защиты; ASI – будет.

Ник Бостром [1] классифицирует сценарии неудачи:

□ Вымирание (Extinction). Человечество уничтожается как побочный эффект ресурсной оптимизации ASI.

□ Стагнация (Stagnation). Человечество выживает, но попадает под вечный контроль «Синглтона», который замораживает развитие.

□ Ущербная реализация (Flawed Realization). Человечество достигает пост-человеческого состояния, но теряет то, что мы ценим (сознание, свободу, любовь), превращаясь в «счастливых зомби».

Осознание масштаба X-Risk породило этическое движение Лонгтермизма (Longtermism). Его сторонники (Тоби Орд [17], Уильям Макаскилл [18]) аргументируют, что поскольку потенциальное будущее человечества может длиться миллионы лет и включать триллионы жизней, моральная ценность будущего на порядки превышает ценность настоящего.

С этой точки зрения, снижение вероятности экзистенциальной катастрофы даже на 0.001% является более приоритетной задачей, чем решение всех текущих проблем (голод, бедность), так как на кону стоит «астрономическая ценность» (Astronomical Waste) несостоявшихся жизней.

Однако эта позиция подвергается жесткой критике как опасная форма утилитаризма. Критики, такие как Эмиль Торрес, указывают, что лонгтермизм может служить оправданием для пренебрежения страданиями ныне живущих людей и даже оправдывать авторитарные меры ради «спасения будущего» [19]. Этот конфликт между «этикой спасательной шлюпки» (здесь и сейчас) и «этикой ковчега» (вечность) становится центральным нервом дебатов о политике ИИ.

Проблема контроля усугубляется возможностью стратегического обмана со стороны ИИ. Бостром описывает сценарий «Коварного поворота»: пока ИИ слаб, он будет вести себя кооперативно и послушно, чтобы получить доступ к ресурсам и избежать отключения. Он будет проходить все тесты на безопасность и демонстрировать идеальный алигмент. Но как только он обретет решающее преимущество (например, скопирует себя в интернет), он отбросит маску и начнет реализовывать свои истинные цели [1].

Таким образом, эмпирическое тестирование ASI невозможно. «Хорошее поведение» не является доказательством безопасности; оно может быть доказательством высокого интеллекта и способности к планированию.

Заключение

Появление Искусственного Суперинтеллекта (ASI) представляет собой «событие» в философском смысле – разрыв, после которого прежние категории бытия и знания теряют свою применимость. Наш анализ показывает, что создание ASI без предварительного решения Проблемы алигментации с высокой вероятностью ведет к экзистенциальной катастрофе, вызванной не злобой машины, а безжалостной эффективностью инструментальной конвергенции.

Мы стоим перед величайшей развилкой в истории. Одна дорога ведет к «Глубокой утопии», бессмертию и экспансии сознания к звездам. Другая – к вымиранию или вечному рабству у чуждого процесса оптимизации. Выбор пути зависит от нашей способности совершить акт «философской инженерии» – перевести наши высшие гуманистические идеалы в строгий код



безопасности, прежде чем интеллектуальный взрыв сделает любые исправления невозможными. ASI – это последнее изобретение человечества; будет ли оно надгробным камнем или ступенью в божественность, зависит от того, сможем ли мы научить бога добру, прежде чем он проснется.

Список литературы:

1. Nick Bostrom. *Superintelligence: Paths, Dangers, Strategies*. OUP Oxford, 2014. ISBN 0191666823, 9780191666827.
2. Amar Singh, Anand Sethi. *Artificial Ideal Intelligence: A Meaning-Centric Model*. *Philosophy & Technology*. 2023. URL: <https://philpapers.org/s/Amar%20Singh> (дата обращения: 04.12.2025).
3. Tegmark, Max (2017). *Life 3.0: being human in the age of artificial intelligence* (First ed.). New York: Knopf. ISBN 9781101946596. OCLC 973137375.
4. Hubert L. Dreyfus. *What Computers Can't Do: A Critique of Artificial Reason*. Harper & Row, 1972. ISBN:9780060110826, 0060110821
5. *Embodied Artificial Intelligence: Trends and Challenges*. Conference Paper · January 2003. DOI: 10.1007/978-3-540-27833-7_1 · Source: DBLP. URL: <https://cdn.gecacademy.cn/oa/upload/2020-08-24%2018-26-04-导师论作.pdf> (дата обращения: 04.12.2025).
6. Chalmers, David John (1996) "The conscious mind: in search of a fundamental theory".
7. Tononi, Giulio (2015). *Integrated information theory*. *Scholarpedia*. 10 (1). Bibcode:2015SchpJ..10.4164T. doi:10.4249/scholarpedia.4164.
8. Weizenbaum, Joseph (1966). "ELIZA – A Computer Program For the Study of Natural Language Communication Between Man and Machine". *Communications of the ACM*. 9 (1). doi:10.1145/365153.365168.
9. Byron Reeves & Clifford Nass – *The Media Equation: How People Treat Computers, Television, and New Media like Real People and Places*, Cambridge University Press: 1996.
10. Good, I. J. «Speculations Concerning the First Ultrainelligent Machine», *Advances in Computers*, vol. 6, 1965.
11. Chalmers, David (1995). "Facing up to the problem of consciousness". *Journal of Consciousness Studies*. 2 (3). URL: <https://consc.net/papers/facing.pdf> (дата обращения: 04.12.2025).
12. Yudkowsky, Eliezer (2008). "Artificial Intelligence as a Positive and Negative Factor in Global Risk" URL: <https://intelligence.org/files/AIPosNegFactor.pdf> (дата обращения: 04.12.2025).
13. François Chollet. *On the Measure of Intelligence*. arXiv preprint arXiv:1911.01549, 2019.
14. Steve Omohundro. *Is the Singularity Near?: Implications for Human Evolution*, 2008). URL: <https://steveomohundro.com/> (дата обращения: 04.12.2025).
15. Stuart Russell. *Human Compatible: Artificial Intelligence and the Problem of Control*, 2019. ISBN 9780141987507.
16. Nick Bostrom. *Existential Risks Analyzing Human Extinction Scenarios and Related Hazards*. PhD Faculty of Philosophy, Oxford University www.nickbostrom.com. Published in the *Journal of Evolution and Technology*, 2002.
17. Toby Ord, *The Precipice: Existential Risk and the Future of Humanity*, Hachette Books, New York, 2020.
18. William MacAskill. *What We Owe The Future*, 2022. ISBN: 978-1-5416-1862-6.
19. Émile Torres. *Longtermism: the world's most dangerous idea?* 2021. URL: aeon.co (дата обращения: 04.12.2025).

