

■ Концептуальная исследовательская статья / Conceptual article

Искусственный интеллект как антропологический проект: от редукции к становлению (ч.1)¹

Д. Е. Николаев

АНО Исследовательский центр «Аналитик». Екатеринбург, Россия

АННОТАЦИЯ Современные исследователи активно обсуждают, чем искусственный интеллект (ИИ) является для человека: культурным орудием, средством цифровой экстернализации, цифровым значимым Другим или альтернативой человека. Гораздо реже встречаются размышления о том, кем является сам человек для технологий и их создателей. В работе показано, что технологии ИИ, представляющие собой не единое инженерное решение, а конгломерат различных подходов, не только базируются на принципиально разных образах человека, но и транслируют их через собственную архитектуру, поддерживая развитие соответствующих практик. Тем самым данные решения влияют на культуру, конструируют новые миры и нового человека в них. Архитектура большинства внедряемых и обсуждаемых исследователями типов ИИ (генеративный, человекоцентричный, «заслуживающий доверия», объяснимый, интерпретируемый, воплощенный, ситуативный, энактивистский, социальный, а также проектируемый в рамках социотехнического и акторно-сетевого подходов) опирается на редуцированные образы человека. Одним из ключевых рисков, связанных с широким внедрением подобных систем, является опасность расчеловечивания – «вымывания» из человеческих практик контекстуального понимания, смыслов творчества, моральной и нравственной ответственности, глубоких межличностных связей, со-действия, вариативности, автономии и самодетерминации. В результате человек становится не более чем исполнителем социальной роли (пользователь, клиент, работник, пациент, студент и т.п.). Преодоление редукции человека в системах ИИ возможно через изменение дизайна архитектуры. Соответствующие технологии должны выходить за пределы упрощенных антропологий и строиться на сложных образах, способствуя не только становлению человека и развитию его потенциала, но и формированию по-настоящему человекоцентричной культуры – культуры достоинства (А. Г. Асмолов). В работе сформулированы практические ориентиры, которые могут лечь в основу проектирования подобных архитектур.

КЛЮЧЕВЫЕ СЛОВА искусственный интеллект, потенциал человека, человекоцентричность, образы человека, антропология технологий

ССЫЛКА ДЛЯ ЦИТИРОВАНИЯ Николаев Д. Е. (2025). Искусственный интеллект как антропологический проект: от редукции к становлению (ч.1). *Образовательная политика*, 23(4), 8–18



© Николаев Д. Е. 2025

Контент доступен под лицензией
Creative Commons Attribution 4 License.
The content is available under
a Creative Commons Attribution 4 License.

Поступила: 20.11.2025
Принята: 01.11.2025
Дата публикации: 30.12.2025



Дмитрий Евгеньевич НИКОЛАЕВ

к. псих. н., заместитель директора,
АНО Исследовательский центр «Аналитик»
(620219, РФ, Екатеринбург, пр. Ленина, 38а),
<https://orcid.org/0000-0002-3385-1640>,
e-mail: dnpsy2019@gmail.com

¹ Продолжение см. в журнале «Образовательная политика» 24(5), 2026.

Введение

Среди исследователей и ведущих разработчиков в последнее время все чаще обсуждается опасность «финансового пузыря» вокруг искусственного интеллекта (ИИ), а также беспокойство по поводу темпов его дальнейшего развития ввиду ограниченности необходимых ресурсов². Несмотря на это, сомнений в ценности технологий ИИ не возникает. Напротив, эти дискуссии лишь усиливают интерес к пониманию того, какое место искусственный интеллект занимает и может занять в человеческой культуре и деятельности. Тренды свидетельствуют, что его позиция окажется значимой. Например, опубликованные в октябре 2025 г. результаты проведенного в США опроса показывают, что 85% учителей и 86% школьников используют ИИ. При этом 42% учеников в ходе такого взаимодействия рассчитывают получить психологическую поддержку, еще столько же используют технологию как способ сбежать от реальной жизни. Многие пользователи воспринимают ИИ в качестве друга или компаньона. В романтические отношения с ним вовлечены 19% респондентов³. Исследования демонстрируют у молодежи конвергенцию реального и цифрового миров в рамках единого переживания личностью смешанной онлайн-офлайн реальности (Солдатова, Чигарькова & Илюхина, 2025).

В научном дискурсе сформировались несколько доминирующих подходов к определению роли и места систем

искусственного интеллекта в жизни человека. Сущность этих подходов часто формулируется с помощью взаимоисключающих метафор: ИИ как «замена человека», «культурное орудие», «средство цифровой экстернализации», «элемент человека расширенного или дополненного» и т.п. Исследователями подробно обсуждалось, что то, чем являются технологии по отношению к человеку, во многом определяется режимом (социотехнической конфигурацией) их внедрения. Одна и та же технология в различных режимах внедрения может становиться либо вытесняющим, либо развивающим человека фактором (Acemoglu & Restrepo, 2018, 2020). Гораздо меньше внимания исследователи уделяют тому, что сама архитектура многих технических систем уже имплицитно предполагает определенный характер взаимодействия с человеком, а значит, задает вполне конкретное распределение ролей в диаде «человек – машина». Конечно, можно попробовать есть вилкой и ножом, как палочками для еды (т.е. изменить «режим внедрения»), но это будет не слишком функциональным и эффективным решением, поэтому большинство людей будет использовать столовые приборы именно так, как задумано их производителями. Каждая проектируемая для взаимодействия с людьми техническая система по умолчанию опирается на неявную теорию человека – на способ представления его природы и возможностей. Хотят ли те, кто разрабатывает и внедряет современные технологии, заместить человека машиной или, напротив, содействовать его развитию (Acemoglu & Restrepo, 2018, 2020; Асмолов, 2024; Ушаков & Валуева, 2022), минимальная рефлексия над проблемой взаимодействия человека и техники в любом случае сталкивается с ключевым вопросом: какой именно образ человека скрыто или

МНОГИЕ ПОЛЬЗОВАТЕЛИ
ВОСПРИНИМАЮТ ИИ В КАЧЕСТВЕ
ДРУГА ИЛИ КОМПАЬОНА.
В РОМАНТИЧЕСКИЕ ОТНОШЕНИЯ
С НИМ ВОВЛЕЧЕНЫ
19% РЕСПОНДЕНТОВ

² Heath, A. (2025, January 24). I talked to Sam Altman about the GPT-5 launch fiasco. The Verge. <https://www.theverge.com/command-line-newsletter/759897/sam-altman-chatgpt-openai-social-media-google-chrome-interview>

³ Center for Democracy & Technology (CDT). (2025, January 30). *Hand in hand: Schools' embrace of AI connected to increased risks to students*. <https://cdt.org/insights/hand-in-hand-schools-embrace-of-ai-connected-to-increased-risks-to-students/>

явно заложен в архитектуре конкретной технологии? Этот образ — больше, чем техническая подробность. Он определяет, что именно в человеке следует развивать или замещать, какие практики будут поощряться, а какие маргинализироваться, и в конечном счете, какой тип субъекта и мира будет конструироваться. Поэтому в рамках дискуссии о замещении или развитии целесообразно прямо поставить вопрос: кто является человеком для проектировщика и для самой технической системы? Иными словами, прежде чем спорить о пользе или вреде той или иной технологии и о режимах ее внедрения, следует раскрыть ее имплицитную антропологию.

Архитектура информационных систем и тот образ человека, на который они явно или неявно опираются, не являются нейтральными. Напротив, и архитектура, и образ выступают как проектные решения, структурирующие поведение индивидов, задающие нормативные ожидания и в результате субъективирующие человека. Значимыми оказываются формальные и интерфейсные решения, протоколы сбора и обработки данных, встроенные цели оптимизации и даже материальная организация аппаратной части. Широкий корпус исследований в области философии техники, STS (Science and technology studies), антропологии и дизайна показывает, что устройства и системы задают поведение пользователей, конструируют ожидания и устанавливают нормы практик также, как законодательство, традиции, культура и институты (Akrich, 1992; Woolgar, 1990; Winner, 1980). Более того, архитектура систем и встроенный в них образ человека не только задают возможности и ограничения практик, но и участвуют в конструировании нового человека и новых миров (Асмолов, 2024; Фаликман, 2020; Skulmowski, 2023; Гудмен, 2001; Солдатова, Чигарькова & Илюхина, 2024а, 2024б; Hayles, 1999). В связи с этим вопрос о влиянии ИИ-архитектур и лежащих в их основе антропологических моделей на человеческие практики и на самого субъекта заслуживает центрального места в исследованиях взаимодействия человека и информационных систем. Именно рассмотрение такого влияния является целью данной работы.

Методы исследования

Данная теоретическая работа опирается на междисциплинарную методологию, сочетающую философско-антропологический анализ, социотехнический подход и элементы критических исследований технологий. В центре нашего внимания находится не функциональная эффективность систем искусственного интеллекта как таковая, а архитектурные решения, через которые в технологии имплицитно встраиваются определенные образы человека. В качестве основного аналитического инструмента используется концептуальная реконструкция, главной целью которой выступает выявление и сопоставление антропологических предпосылок, заложенных в различных типах архитектур ИИ, и анализ того, какие формы субъектности и практик они делают вероятными.

Также применяется сравнительно-типологический метод, позволяющий систематически сопоставить ключевые подходы к проектированию ИИ (от генеративных и человекоцентричных до воплощенных, социотехнических и реляционных) с точки зрения их отношения к человеческой автономии, становлению и смыслообразованию. Материалами для исследования служат опубликованные архитектурные описания, нормативные рамки, кейсы внедрения и результаты междисциплинарных трудов, интерпретируемые в логике социотехнического анализа. Тем самым автор транслирует позицию «рефлексивной инженерии», в которой архитектура ИИ рассматривается как культурно и антропологически нагруженный объект, а не как нейтральное техническое средство.

Влияние архитектуры ИИ на человеческие практики

Под архитектурой ИИ мы понимаем совокупность проектных решений, которые определяют форму функционирования системы: ее модели и алгоритмы, схемы представления и потоков данных,

интерфейсные протоколы взаимодействия, механизмы обучения и контроля, а также правила развертывания и интеграции в организационные и инфраструктурные контексты. Каждое решение несет представление о том, кем является человек, взаимодействующий с системой, какие у него цели, какими ресурсами и потенциалом он располагает и какие формы взаимодействия возможны. Таким образом, архитектура опосредует представление о человеческой природе, определяющее, какие практики будут поощряться, а какие подавляться.

Проблема влияния технологий (в самом широком понимании термина) на человеческие практики и человека обсуждалась представителями различных научных дисциплин и подходов. В культурно-деятельностной психологии развиваются идеи Л. С. Выготского о культурных орудиях как звене, опосредующем воздействие на субъекта внешнего, физического мира (Выготский, 1983; Асмолов, 2023). Целый ряд исследователей, представляющих различные школы психологической науки (К. Левин, К. Дункер, Н. А. Бернштейн, А. Н. Леонтьев, Л. И. Божович и др.), отмечает, что орудие во многом структурирует и даже конструирует деятельность, использующего его человека (Асмолов, 2023). Психофизиолог И. М. Фейгенберг (2011) пишет о «человеке достроенном», имея в виду и достраивание с помощью технологий. Культуролог М. Маклюэн рассматривает любые технологии и медиа как своеобразные «протезы», «расширения» или «продолжения» человека (McLuhan, 1964). Э. Кларк и Д. Чалмерс развивают теорию активного экстернализма или расширенного сознания, предполагающую, что внешние, в т. ч. технические объекты участвуют в формировании ментального (Clark & Chalmers, 1998). В рамках постфеноменологии

РЕКОМЕНДАТЕЛЬНЫЕ АЛГОРИТМЫ МЕДИА И СОЦИАЛЬНЫХ ПЛАТФОРМ, СПРОЕКТИРОВАННЫЕ ПОД МАКСИМИЗАЦИЮ ВОВЛЕЧЕННОСТИ, ПООЩРЯЮТ КРАТКОВРЕМЕННЫЕ, ИНТЕНСИВНЫЕ ФОРМЫ ВНИМАНИЯ И СНИЖАЮТ СКЛОННОСТЬ К РЕФЛЕКСИИ

Д. Ихде и П. — П. Вербик подчеркивают, что технологии выступают медиаторами нашего переживания мира и тем самым трансформируют восприятие и способы действия. Кроме того, технологии выполняют «морализирующую» роль: устройства и системы не просто облегчают выполнение задач, но и меняют нравственные ориентиры и формируют новые возможности для оценки и выбора (Ihde, 1990; Verbeek, 2011). Л. Виннер показывает, что в артефакты может быть заложена определенная политика; их форма и архитектура могут конституировать властные отношения и институциональные практики (Winner, 1980). В свою очередь, исследуя власть, знание и субъективацию в рамках постструктурализма, М. Фуко показывает, как дискурсы и практики конструируют людей как субъектов со своими особенностям, такими, как дисциплинированность, нормированность, ответственность (Foucault, 1977). В совокупности названные подходы дают базовую методологическую установку: архитектуру систем ИИ и соответствующие проектные допущения следует рассматривать как ключевое основание конструирования определенных социальных практик и субъектности.

Воздействие архитектуры ИИ на человека и его практики проявляется через несколько взаимосвязанных механизмов:

- через предметную детерминацию образа мира (культурно-деятельностная психология) или функциональную фиксированность (К. Дункер). Иными словами, через свойства или функции системы, подсказывающие характер взаимодействия с ней (аффордансы). Архитектура интерфейса, временные лаги отклика, доступность информации и способы обратной связи формируют

- определенные возможности взаимодействия, буквально «требуя» от пользователя вести себя конкретным образом. Эти «вшитые» в архитектуру образцы поведения быстро становятся шаблонами/ ментальными моделями/ навыками и т. п. и транслируются в культуре;
- через поощрение к совершению определенных действий. Алгоритмы оптимизации и экономика продукта создают стимулы, переориентирующие поведение индивидов на то, чтобы поддерживать активность, делиться данными, следовать рекомендованными маршрутами. Это приводит к нормализации определенных практик;
 - через фильтрацию и предписывание. Архитектура классифицирует, ранжирует и отсекает информацию. Таким образом, она формирует для пользователя его «мир вокруг» и предписывает, что важно, а что нет (т. н. эхо-камера, информационный или цифровой «пузырь»);
 - через перераспределение ответственности и субъектности. Когда часть решений автоматизируется, люди могут перекладывать на систему ответственность за выбор и проявлять меньшую субъектность в сферах, где ранее демонстрировали активную рефлексию и развитые навыки самодетерминации.

Все эти механизмы в совокупности влияют на формирование социальных ролей (пользователь, клиент, работник, пациент, студент и т. п.), языка самосознания и особенностей субъектности.

На уровне практик это можно продемонстрировать на нескольких примерах. Так, рекомендательные алгоритмы медиа и социальных платформ, спроектированные под максимизацию вовлеченности, поощряют кратковременные, интенсивные формы внимания и снижают склонность к рефлексии. Они конструируют пользователей, ориентированных на мгновенное подкрепление. Другой пример: образовательные адаптивные системы. Многие из них предполагают модель

учащегося-потребителя и оценивают успех по краткосрочным тестовым результатам, тем самым формируя практику «обучения под тест», сужая образовательные траектории и превращая учащегося из субъекта в оптимизируемый объект. Третий пример: большие языковые модели, имитирующие диалог, создают ситуацию, когда человек может перестать развивать способность четко формулировать и критически переосмыслять свою речь, что делает его «потребителем готового смысла» (Bommasani et al., 2021; Bender et al., 2021). Все эти примеры показывают, что архитектуры, ориентированные на определенные KPI и метрики, делают соответствующие формы субъектности более вероятными.

Важно подчеркнуть, что подобные эффекты приобретают дополнительную устойчивость при институциональной поддержке (через культурные и законодательные нормы, модели монетизации, образовательные программы и т. п.). При такой поддержке фасилитируемое поведение может быстро конвертироваться в норму, и тогда трансляция культуры начинает происходить уже системно, путем повсеместного формирования новых практик профессиональной деятельности, общения, образования и т. д. В этом смысле технология выполняет роль «нормативного катализатора», поскольку она не только предлагает новые действия, но и маркирует их как адекватные, рациональные (Winner, 1980; Jasanoff, 2016). Таким образом, субъективация — это эффект не только самой информационной системы, но и конструируемых вокруг нее смыслов.

Ключевые типы архитектуры искусственного интеллекта

Под типом архитектуры ИИ в этой работе предлагается понимать согласованный набор предпосылок:

а) онтологического характера (что собой представляет технология — актер, инструмент, медиатор, инфраструктуру, часть расширенного «мы»?);

- б) эпистемологического характера (что она «знает» и как это «знание» конституируется?)
- в) нормативного характера (на ком лежит ответственность? Чьи интересы она отражает? Какие ценности встраиваются в проект?).
- Такой подход переводит типологию с чисто технического уровня на уровень социальных практик и смыслов, делая возможным синтез работ по психологии, философии информации, критической теории технологий, антропологии и прикладной этике.
- Рассмотрим ключевые типы архитектуры ИИ — как уже реализуемые на практике, так и только обсуждаемые исследователями.

Генеративный ИИ (Generative AI)

Появление больших языковых моделей (LLM) усилило интерес к особому классу ИИ, который описывается как генеративный. Такие системы не столько «решают задачи», сколько порождают и перестраивают тексты, изображения, музыку, формируют гипотезы, то есть участвуют в символическом производстве и интерпретации. Технология здесь выступает как герменевтический субъект, не мыслящий в человеческом смысле, но способный встраиваться в культурную семиосферу (Ю. М. Лотман).

Отличительная черта генеративного ИИ — обучение на огромных, разнотипных корпусах данных и последующая адаптация к множеству задач, что обеспечивает одновременно и универсальность применения, и специфические риски (смещение, утрата контекста, непредсказуемые эмерджентные «способности», а также «галлюцинации») (Bommasani et al., 2021). Критики генеративного режима ИИ показывают, что такой подход чреват репродукцией предвзятостей и иллюзией понимания. Знаменитая статья «О рисках «стохастических попугаев»: не слишком ли велики языковые модели?» предостерегает

против бездумного наращивания размеров моделей (Bender et al., 2021).

Генеративные системы стали материалом для быстрого прототипирования, создания ассистентов, инструментов для творчества и корпоративных сервисов. LLM-платформы широко внедряются в разнообразные продукты и процессы, что породило новую волну исследований об их социальных и институциональных последствиях. С научной точки зрения, сейчас дискуссия переходит от вопроса «можно ли получить хорошие результаты?» к вопросам управления экосистемой, т. е. к проблемам верификации, внешних эффектов, экологических аспектов и обладания правами на операции с данными. Практически LLM используются повсеместно: в корпоративных чат-ботах, системах поддержки принятия решений, образовании и медиа. Это актуализирует вопросы о долгосрочных социальных последствиях применения LLM.

Архитектура генеративного ИИ обычно включает в себя уровни представления, механизмы обработки последовательностей и модуль для дообучения под конкретные задачи. Такая архитектура предполагает, что единая модель усваивает широкие статистические закономерности языка или изображений, а затем выдает варианты вывода путем вероятностной генерации. Управлять выходом модели необходимо при помощи внешних протоколов фильтрации, дообучения и инструкций от человека. Вследствие этой конструкции способности модели к генерации тесно связаны с масштабом и свойствами исходных данных, с архитектурными гиперпараметрами и с процедурами постобработки, а не с внутренним пониманием смысла или намерений (Bommasani et al., 2021; Brown et al., 2020).

АРХИТЕКТУРА ЧЕЛОВЕКОЦЕНТРИЧНОГО ИИ ПРЕДПОЛАГАЕТ ИНТЕГРАЦИЮ АЛГОРИТМИЧЕСКИХ ПОДСИСТЕМ С ИНТЕРФЕЙСАМИ ПОЯСНЕНИЙ, УПРАВЛЕНИЕМ ОШИБКАМИ И ИНСТРУМЕНТАРИЕМ СОВМЕСТНОЙ РАБОТЫ

Появление LLM усилило представление об ИИ как о герменевтическом акторе: машина генерирует тексты, перестраивает контексты и тем самым участвует в культурном мире смыслов. В этой перспективе человек предстает не как пассивный потребитель сервиса, а как участник дискурсивного поля (Bommasani et al., 2021; Bender et al., 2021). Образ человека здесь — преимущественно культурный интерпретатор, с той оговоркой, что технологии, усиливающие имитационную герменевтику (симуляцию «понимания»), подменяют глубинные процессы формирования смысла их имитацией.

Человекоцентричный ИИ (Human-Centered AI, HCAI)
Одна из наиболее распространенных нормативных рамок — человекоцентричный ИИ. Этот подход предполагает проектирование систем, приоритетом для которых являются человеческие ценности, потребности и развитие. Данный тип архитектуры характеризуется вниманием к механизмам контроля человека над системой. В нормативной политике и практике обозначенный подход получил институциональное выражение в документах регуляторов и в академических работах, декларирующих внимание к человеку как принцип проектирования. Суть подхода заключается в требовании, чтобы системы соответствовали юридическим и этическим стандартам на протяжении всего жизненного цикла, а не только оптимизировали заданные технические метрики⁴.

Парадигма HCAI на сегодняшний день достигла значительной зрелости: она представлена в академических институтах (например, Stanford HAI), стандартах проектирования и ряде отраслевых руководств, а также стала доминировать во многих корпоративных и государственных документах, направленных на разработку «ответственного» дизайна ИИ⁵. На практике

это выражается в специальных техниках проектирования взаимодействия, процедурах оценки удобства и вовлечения пользователей, а также интеграции междисциплинарных команд в разработки продуктов (инженеры, UX-дизайнеры, социологи, юристы). Вместе с тем HCAI остается фрагментарной практикой, т. к. пока показывает только локальный успех (инструменты, ориентированные на конкретные пользовательские группы). Системное внедрение человекоцентричных принципов сталкивается с инерцией рынка, сложностью монетизации и отсутствием единых стандартов. Следствием этого является то, что ориентация на человека только декларируется, а при воплощении ценностей в измеримые инженерные требования возникают трудности.

Архитектура человекоцентричного ИИ предполагает интеграцию алгоритмических подсистем с интерфейсами пояснений, управлением ошибками и инструментарием совместной работы. Технически это означает применение адаптивных моделей, механизмов обратной связи от пользователей и дизайнера взаимодействия, ориентированного на прозрачность и вовлечение. При этом архитектура учитывает разные уровни пользователей. Ее главная цель состоит в том, чтобы обеспечить условия, при которых система усиливает человеческие способности и сохраняет контроль человека над результатами (Amershi et al., 2019; Friedman, Kahn & Borning, 2006).

Под термином «человекоцентричность» здесь обычно скрывается образ человека как функционального субъекта, чьи потребности и задачи подлежат оптимизации. Это рациональный субъект, в интересах которого система должна обеспечивать эффективность, удобство и безопасность взаимодействия. Человек рассматривается как «пользователь», т. е. носитель задач, когнитивный ресурс и источник требований к системе. Такое представление оправдывает метрики производительности, скорости, точности и удовлетворенности пользователя, проектные усилия направлены на и повышение «юзабилити». Критика этого образа отмечает его редукционизм: образ человека как

«пользователя», несмотря на декларируемую человекоцентричность, не включает в себя глубинные измерения становления личности, смыслообразования и автономии.

«Заслуживающий доверия» (ответственный) ИИ (Trustworthy / Responsible AI)
Этот подход во многом пересекается с человекоцентричным, но имеет другой фокус внимания. Здесь исследователи видят основную проблему в согласовании целей ИИ с человеческими ценностями. Теоретики безопасности и философы задают ключевой вопрос: можно ли гарантировать, что автономные системы будут реализовывать желаемые человеком цели без опасных побочных эффектов? С. Расселл предложил принципиальную перестройку парадигмы: вместо априорной фиксации целей важно наделять систему неопределенностью относительно человеческих предпочтений и учить ее экстрагировать эти предпочтения из конкретного поведения людей (Russell, 2019). Такая нетривиальная задача создает многочисленные методологические трудности и в то же время стимулирует появление междисциплинарных работ в области безопасности и этики ИИ.

В последнее время исследовательское направление, сосредоточенное на согласовании («выравнивании») поведения систем с целями и ограничениями человека, переживает период бурного развития. Традиционные подходы дополнились новыми методами: обучение с подкреплением на основе человеческой оценки (reinforcement learning from human feedback, RLHF), а также методы и схемы, когда сама модель участвует в поддержании ограничений (в определенной мере реализуется в проектах Anthropic, OpenAI и пр.). Наблюдается переход от чисто теоретических задач к практической реализации (пилотные и производственные развертывания моделей

с многоуровневыми процедурами обеспечения безопасности). Однако теоретики и практики отмечают фундаментальные трудности, например, масштабируемость процессов «выравнивания», конфликт между полезностью и «безвредностью», а также сохранение формальных свойств при сложных способах обучения. На уровне отрасли это означает интенсивное развитие практик, но пока преимущественно внедряются далекие от универсальности технические решения.

Архитектура систем, претендующих на «выравнивание» ценностей, сочетает на формальные компоненты (спецификации ограничений, функции потерь с учетом этических критериев), механизмы обратной связи от людей (human-in-the-loop) и слои мониторинга безопасности. Часто в нее встроены процедуры

многократной адаптации и проверки — от RLHF до правил, управляющих генерацией. Технически задача архитектуры состоит в формализации ценности в корректируемые параметры и обеспечении возможности верификации и контроля. Соответственно, архитектура ИИ должна включать в себя протоколы аудита, модульные барьеры и механизмы обновления политик поведения (Russell, 2019; Bai et al., 2022).

Описываемый подход базируется на образе человека как носителя устойчивых (хотя и частично неопределенных) предпочтений и ценностей, которые можно формализовать, извлечь и встроить в целевые функции систем. Этот образ близок к homo economicus, но дополнен эмпирическим пониманием того, что предпочтения проявляются через поведение. Вследствие этого проектирование упирается в проблемы наблюдаемости, амбивалентности и динамики предпочтений. Оценка успеха системы сводится к тому, насколько поведение системы согласуется с ценностями человека (Russell, 2019). Подобная

РЕЖИМ «ОДНО ОБЪЯСНЕНИЕ ДЛЯ ВСЕХ» НЕ РАБОТАЕТ, И ОБЪЯСНИТЕЛЬНЫЕ ИНТЕРФЕЙСЫ ДОЛЖНЫ БЫТЬ АДАПТИРОВАНЫ К ОСОБЕННОСТЯМ ПОЛЬЗОВАТЕЛЯ И К ЦЕЛЯМ ВЗАИМОДЕЙСТВИЯ

4 European Commission. (2019). Ethics guidelines for trustworthy AI. High-Level Expert Group on Artificial Intelligence. <https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai>

5 Там же.

редукция игнорирует историчность субъекта и его способность к саморазвитию. Человек в этой модели — лишь объект согласования, но не субъект становления. Также игнорируются аффективные, телесные и неосознаваемые проявления.

Объяснимый и интерпретируемый ИИ (Explainable / Interpretable AI)
Это направление сформировалось в ответ на проблему черного ящика в системах ИИ. Здесь объединяются два близких, но не тождественных направления: с одной стороны, инженерные и методологические практики, целью которых является построение моделей, интуитивно понятных человеку (интерпретируемый ИИ); с другой стороны, техники и процедуры объяснений, призванные дать осмысленное представление о том, почему модель сформулировала тот или иной ответ (объяснимый ИИ) (Lipton, 2016; Doshi-Velez & Kim, 2017).

ИИ здесь играет несколько ролей, коррелирующих с разными группами адресатов. Разработчики и инженеры нуждаются в объяснениях для отладки и улучшения моделей; регуляторы и аудиторы — для проверки соответствия нормативам и выявления предвзятостей; конечные пользователи — для принятия или оспаривания решений, восстановления контроля. Каждая роль предъявляет свой набор требований. Для инженера объяснимость — это возможность проследить работу системы на разных этапах, тогда как для обычного пользователя «объяснение» должно быть кратким, четким, релевантным его ценностям (Doshi-Velez & Kim, 2017; Miller, 2019). Отсюда следует важный практический вывод: режим «одно объяснение для всех» не работает, и объяснительные интерфейсы должны быть адаптированы к особенностям пользователя и к целям взаимодействия.

Исследования в этом подходе активно развиваются с конца 2010-х и уже привели к разработке широкого набора инструментов — от визуализаций внимания модели до локальных интерпретаций (Gunning & Aha, 2019). В некоторых отраслях, например, в медицине и праве, объяснимый ИИ стал

обязательным элементом проектирования, поскольку без объяснимости его применение ограничено этически и юридически. Однако исследования показывают, что большинство пользователей воспринимают объяснения скорее как средство повышения доверия, чем как инструмент понимания (Miller, 2019). Это ставит под сомнение их когнитивную ценность и требует развития более глубоких форм взаимодействия человека и машины, выходящих за рамки поверхностных интерпретаций.

Архитектурные решения в объяснимом/интерпретируемом ИИ направлены на обеспечение прозрачности и понятности выводов. На практике такая архитектура объединяет модель, дающую рекомендации, надстройку объяснений, адаптированную к профилю пользователя, и инфраструктуру логирования и представления сведений об источнике и логике данных или решений. Ключевая инженерная проблема — как соотнести технические объяснения с теми формами, которые действительно помогают людям понимать, критиковать и оспаривать выводы (Gunning, 2019; Miller, 2019; Doshi-Velez & Kim, 2017).

В большинстве реализаций объяснимого/интерпретируемого ИИ присутствует сочетание двух образов человека. Во-первых, технические решения часто предполагают, что человек — рациональный субъект, обладающий ограниченным знанием, нуждающийся в аргументах ради принятия решения. Объяснение рассматривается как информационный материал, который при корректной подаче позволит пользователю рационально исправить или подтвердить вывод. Во-вторых, в нормативной и регуляторной повестке такой тип ИИ апеллирует к образу человека как правового и морального субъекта, то есть гражданина или клиента, обладающего правом на понимание и оспаривание автоматизированных решений. Оба этих образа важны, но они редуцируют человека.

Редукция до рационального оператора, способного обработать и корректно интерпретировать формальные объяснения, игнорирует реальные особенности того,

как люди на самом деле воспринимают и используют объяснения, например, роль бессознательного и телесности, когнитивные предубеждения, эмоциональные реакции, социальную зависимость и контекстуальные влияния. Редукция до юридического субъекта, чьи интересы достаточно защитить формальными процедурами обоснования, игнорирует необходимость использования для оспаривания решений права на контекстуальную поддержку, доступа к вторичным источникам информации и т. п. В обоих случаях ИИ склонен предполагать, что информация об алгоритмическом процессе равнозначна способности человека критически оценить этот процесс, однако, это часто не так (Miller, 2019; Wachter, Mittelstadt & Floridi, 2017).

Список источников

Асмолов, А. Г. (2023). *Психология личности. Культурно-историческое понимание развития человека* (6-е изд.). Смысл.

Асмолов, А. Г. (2024). Кризис как тест на психологическую совместимость. От Войны миров — к Содружеству разумов. *Образовательная политика*, 2(98), 6–12. <https://www.elibrary.ru/lrgjup>

Выготский, Л. С. (1983). История развития высших психических функций. В Л. С. Выготский, *Собрание сочинений* (Т. 3, с. 5–328). Педагогика.

Гудмен, Н. (2001). *Способы создания миров*. Идея-Пресс, Логос, Праксис.

Солдатова, Г. У., Чигарькова, С. В., & Илюхина, С. Н. (2024а). Метаморфозы идентичности человека достроенного: от цифрового донора к цифровому кентавру. *Социальная психология и общество*, 15(4), 40–57. <https://www.elibrary.ru/dddlsg>

Солдатова, Г. У., Чигарькова, С. В., & Илюхина, С. Н. (2024б). Технологически расширенная личность: разработка и апробация шкалы самоуправления цифровой повседневностью. *Вестник Московского университета. Серия 14. Психология*, 47(2), 177–178. <http://elibrary.ru/hdxghe>

Солдатова, Г. У., Чигарькова, С. В., & Илюхина, С. Н. (2025). Цифровые предикторы психологического благополучия молодежи в реальном и виртуальном мирах. *Вестник Московского университета. Серия 14. Психология*, 48(1), 78–100. <https://www.elibrary.ru/ufxouj>

Ушаков, Д. В., & Валуева, Е. А. (2022). Вызовы искусственного интеллекта для психологии. *Человек и системы искусственного интеллекта* (с. 107–127). Юридический центр.

Фейгенберг, И. М. (2011). *Человек Достроенный и этика. Цивилизация как этап развития жизни Земли*. Медицинское информационное агентство.

References

Acemoglu, D., & Restrepo, P. (2018). Artificial intelligence, automation, and work. In A. Agrawal, J. Gans, & A. Goldfarb (Eds.), *The economics of artificial intelligence: An agenda* (pp. 197–236). University of Chicago Press.

Acemoglu, D., & Restrepo, P. (2020). The wrong kind of AI? Artificial intelligence and the future of labor demand. *Cambridge Journal of Regions, Economy and Society*, 13(1), 25–35. <https://doi.org/10.1093/cjres/rsz022>

Akrich, M. (1992). The de-scription of technical objects. In W. E. Bijker & J. Law (Eds.), *Shaping technology/building society: Studies in sociotechnical change* (pp. 205–224). MIT Press.

Amershi, S., Weld, D., Vorvoreanu, M., Fournery, A., Nushi, B., Collisson, P., Suh, J., Iqbal, S., Bennett, P. N., Inkpen, K., Teevan, J., Kikin-Gil, R., & Horvitz, E. (2019). Guidelines for human-AI interaction. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Paper 3), 1–13. <https://doi.org/10.1145/3290605.3300233>

Bai, Y., Jones, A., Ndousse, K., Askell, A., et al. (2022). *Constitutional AI: Harmlessness from AI feedback*. <https://doi.org/10.48550/arXiv.2212.08073>

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (FAccT '21), 610–623. <https://doi.org/10.1145/3442188.3445922>

Bommasani, R., Hudson, D. A., Adeli, E., et al. (2021). *On the opportunities and risks of foundation models*. <https://doi.org/10.48550/arXiv.2108.07258>

Brown, T. B., Mann, B., Ryder, N., et al. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33, 1877–1901.

Clark, A., & Chalmers, D. (1998). The extended mind. *Analysis*, 58(1), 7–19. <https://doi.org/10.1093/analys/58.1.7>

Doshi-Velez, F., & Kim, B. (2017). *Towards a rigorous science of interpretable machine learning*. <https://doi.org/10.48550/arXiv.1702.08608>

Foucault, M. (1977). *Discipline and punish: The birth of the prison*. Vintage Books.

Friedman, B., Kahn, P. H., Jr., & Borning, A. (2006). Value sensitive design and information systems. In K. E. Himma & H. T. Tavani (Eds.), *The handbook of information and computer ethics* (pp. 69–101). John Wiley & Sons.

Gunning, D., & Aha, D. W. (2019). DARPA’s explainable artificial intelligence (XAI) program. *AI Magazine*, 40(2), 44–58. <https://doi.org/10.1609/aimag.v40i2.2850>

Hayles, N. K. (1999). *How we became posthuman: Virtual bodies in cybernetics, literature, and informatics*. University of Chicago Press.

Ihde, D. (1990). *Technology and the lifeworld: From garden to earth*. Indiana University Press.

Jasonoff, S. (2016). *The ethics of invention: Technology and the human future*. W. W. Norton & Company.

Latour, B. (2005). *Reassembling the social: An introduction to actor-network-theory*. Oxford University Press.

Lipton, Z. C. (2018). The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue*, 16(3), 31–57. <https://doi.org/10.1145/3236386.3241340>

McLuhan, M. (1964). *Understanding media: The extensions of man*. McGraw-Hill.

Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267, 1–38.
<https://doi.org/10.1016/j.artint.2018.07.007>

Russell, S. (2019). *Human compatible: Artificial intelligence and the problem of control*. Viking.

Skulmowski, A. (2023). The cognitive architecture of digital externalization. *Educational Psychology Review*, 35(4), 101.
<https://doi.org/10.1007/s10648-023-09820-7>

Verbeek, P.-P. (2011). *Moralizing technology: Understanding and designing the morality of things*. University of Chicago Press.

Wachter, S., Mittelstadt, B., & Floridi, L. (2017). Why a right to explanation of automated decision-making does not exist in the General Data Protection Regulation. *International Data Privacy Law*, 7(2), 76–99.
<https://doi.org/10.1093/idpl/ix005>

Winner, L. (1980). Do artifacts have politics? *Daedalus*, 109(1), 121–136.

Woolgar, S. (1990). Configuring the user: The case of usability trials. *The Sociological Review*, 38(1_suppl), 58–99.
<https://doi.org/10.1111/j.1467-954X.1990.tb03349.x>

Artificial Intelligence as an nthropological Project: From Reduction to Becoming (Part 1)

Dmitry E. Nikolaev
PhD in Psychology., Deputy director of ANO Research Centre “Analytic” (38a, Lenin Ave., Yekaterinburg, Russian Federation, 620219), <https://orcid.org/0000-0002-3385-1640>, e-mail: dnpsy2019@gmail.com

ABSTRACT Scholars actively debate what artificial intelligence is for humans – a cultural tool, a means of digital externalization, a digital significant Other, or a future alternative, replacement for the human. Far less frequently discussed is who the human is for technologies and their designers. The paper shows that AI technologies – understood not as a single engineering solution but as a system of heterogeneous approaches – not only rest on fundamentally different conceptions of the human, but also transmit these images through their architectures, thereby facilitating corresponding practices and shaping culture, new worlds and a new forms of being. The architectures of most types of AI both, implemented in practice and only discussed in the literature (generative, human-centered, trustworthy, explainable, interpretable, embodied, situated, enactive, social, as well as designs developed within sociotechnical and actor-network frameworks) rely on reduced, incomplete images of the human being. Among the principal risks associated with broad deployment of such systems is the risk of dehumanization – the elimination from human practices of those features that make them genuinely human: contextual understanding, meaning-making, moral and ethical responsibility, deep interpersonal bonds, co-action, variability, autonomy and self-determination. In these architectures the human too often appears only as a performer of a social role (user, client, worker, patient, student, etc.). Overcoming the reduction of the human in AI architectures is possible through redesign of those architectures. Corresponding technologies must move beyond simplified anthropologies and be built on complex images of the human being. Such technologies can not only facilitate human development and the realization of human potential, but also contribute to the formation of a truly human-centered culture, a culture of dignity (A.G. Asmolov). Practical tenets intended to underlie the design of non-reductionist architectures are formulated.

KEY WORDS artificial intelligence; human potential; human-centricity; conceptions of the human; anthropology of technology

TO CITE Nikolaev, D. E. (2025). Artificial Intelligence as an Anthropological Project: From Reduction to Becoming (Part 1). *Educational Policy*, 23(4), 8–18

© Nikolaev D. E
Received: 01.11.2025
Accepted: 20.11.2025
Date of publication: 30.12.2025

