Философия науки и техники
2025. Т. 30. № 1. С. 21–33
УДК 168.521+165.12

Philosophy of Science and Technology
2025, vol. 30, no. 1, pp. 21–33
DOI: 10.21146/2413-9084-2025-30-1-21-33

НАУКИ О ЧЕЛОВЕКЕ

С.В. Лаврентьева

Автономия пациента в контексте развития технологий по персонализации ИИ: новые подходы и этические вызовы*

Лаврентьева Софья Всеволодовна – младший научный сотрудник. Институт философии РАН. Российская Федерация, 109240, г. Москва, ул. Гончарная, д. 12, стр. 1; e-mail: sonnig89@gmail.com

В данной статье рассматриваются этические вызовы, связанные с применением больших языковых моделей в принятии медицинских решений в ситуации недееспособности пациентов. Показано, как развитие технологий искусственного интеллекта (ИИ) проблематизирует формулировку критериев принципа автономии пациента в медицине. Возможности нейросетей в выявлении закономерностей поведения на основе больших данных и имитации персональных черт характера способствуют появлению инициатив по технологической поддержке совместного принятия решений на основе ИИ. Но неспособность таких моделей воспроизвести персональные качества и ценности конкретного человека вызывает сомнение в том, что принцип автономии будет соблюден. Усовершенствованием подобного подхода может стать персонализированный предиктор предпочтений пациента (П4). П4 является большой языковой моделью, обученной на личных данных агента, что позволяет учитывать личные предпочтения. Однако создание П4 требует четкой формулировки подхода к отбору материала для обучения. Основной проблемой здесь может стать потенциальная неконсистентность убеждений (beliefs) и оценочных установок (attitudes) агента в переписке, социальных сетях и устных суждениях. Исследования в области социальной психологии указывают на возможность использования психометрических данных для воспроизведения оценочных установок в обучении нейросетей. Подобный подход может стать шагом к разработке расширенной автономии на основе ИИ и инициировать дискуссии о роли характера личности в автономных решениях.

Ключевые слова: искусственный интеллект, автономия, большие языковые модели, предиктор, оценочные установки, убеждения, субъектность

* Работа выполнена при поддержке РНФ, проект № 24-28-01604 «Проблема субъектности в биоэтике: философско-практический анализ».

Биоэтика является особой сферой практической философии, сосредоточенной на оценке рисков внедрения новых научных методов и технологий в медицинской сфере. В настоящее время один из ключевых для биоэтики принципов – принцип автономии пациента – может быть пересмотрен в контексте развития технологий искусственного интеллекта. Основной вызов для регуляции автономии представляет использование нейросетей как инструмента предикции решений недееспособных пациентов. В данной статье мы покажем, как участие ИИ в процессе совместного принятия решений о здоровье таких пациентов соотносится с уже существующими нормами соблюдения принципа автономии в медицине. В исследовании также будут подняты вопросы, связанные с представлением о личности, способной принимать автономные решения.

Понятие автономии в биоэтике связано с четырьмя принципами, сформулированными Т. Бичампом и Дж. Чилдрессом и составляющими этическую основу медицинской практики: уважение автономии пациента, непричинение вреда, благодеяние и справедливость (иначе принципализм, или «джорджтаунская мантра») [Childress, Beauchamp, 2001]. В рамках принципализма этические концепции адаптируются для практического применения и описания реальных медицинских ситуаций. Ориентация описания автономии в биоэтике на регуляторную операциональность данного термина приводит к тому, что он сильно отличается от представлений об автономии в классической философии, ориентированной на формулировку автономии у Иммануила Канта. Принцип уважения автономии пациентов основывается на кантовском понятии автономии, но имеет явные отличия от него.

Кант рассматривает автономию как способность разума самостоятельно устанавливать моральные законы, руководствуясь моральными принципами, которые могут быть признаны универсальными. Автономия противопоставляется Кантом гетерономии – ориентации на внешние потребности, интересы и желания [Артемьева, 2018]. Бичамп и Чилдресс также говорят об автономии как о способности самостоятельного действия, основанного на рациональном принципе, с которым агент согласен. Автономное действие должно соответствовать определенным стандартам рациональности, а автономный агент способен обдумывать свое решение, понимая его причины. Но авторы «джорджтаунской мантры» не развивают тему моральных принципов в автономном решении и не акцентируют отличие автономии от гетерономии. На первом плане для них стоит право индивида принимать решение на основании своих убеждений и ценностей, без давления извне [Белялетдинов, 2019, с. 25].

Автономия в биоэтике не только предполагает наличие рационального основания для принятия решений, но и требует учета правового контекста, в котором эти решения принимаются. Философские дискуссии о принципе автономии всегда будут соотноситься с представлением о пациенте как субъекте, обладающем определенным правовым статусом, компетентностью и со способностью пациента принимать решения, его дееспособностью. При оценке способности пациента автономно принимать решения под дееспособностью понимается когнитивная способность человека понимать и обдумывать свой выбор. То, насколько человек компетентен принимать собственные решения,

определяет его юридические полномочия, его законное право действовать так или иначе.

При подобной трактовке становятся возможными такие формы автономии, как ассистированная или делегированная. Ассистированная автономия предполагает, что агент может нуждаться в помощнике, который может обеспечить соблюдение правил автономии по части информированности и практическую поддержку в процессе принятия решения. Делегированная автономия подразумевает, что в случаях, когда агент недееспособен из-за состояния здоровья или из-за когнитивных нарушений, обязанность принимать решения может быть делегирована третьим лицам, в англоязычной литературе обозначаемым термином «суррогаты» (“surrogates”). Предполагается, что «суррогаты» (чаще всего ими становятся родные и близкие пациентов) принимают решение, ориентируясь на свои знания о предпочтениях и убеждениях пациента, что в конечном итоге помогает соблюсти принцип автономии пациента [DeGrazia, Millum, 2021; Earp et al., 2024, p. 14].

Но и самим «суррогатам» может понадобиться помощь в том, чтобы артикулировать потенциальное решение пациента. В настоящее время идет разработка различных проектов по созданию автоматизированных систем, предикторов, цель которых – предоставить сведения о вероятном выборе пациента и облегчить процесс совместного принятия решений.

Одним из примеров подобных проектов является инициатива по созданию «Предиктора предпочтений пациента», сокр. ППП (“Patient Preference Predictor”, PPP). Предполагается, что ППП будет функционировать на основе эмпирических данных, собранных в ходе репрезентативных опросов о предпочтениях людей относительно совместного принятия решений в периоды недееспособности. Опираясь на выявленные корреляции характеристик пациента (возраст, пол, социальное положение и т.п.) со статистикой опросов, ППП сможет ответить, какой вариант лечения с наибольшей вероятностью предпочтет конкретный недееспособный пациент [Rid, Wendler, 2014].

Автоматизация подхода к выявлению потенциальных предпочтений пациента позволяет нам говорить о более широкой перспективе использования новых технологий в принятии совместных решений. Применение наработок в сфере ИИ предоставляет возможности по усовершенствованию подобных предикторов. Применение технологий больших языковых моделей (БЯМ) может сделать подобные предикторы более точными за счет работы с большими данными. Более того, в рамках использования БЯМ появляется возможность создавать персонализированные сценарии потенциального автономного решения.

Неперсонализированный предиктор предпочтений пациента как реализация «алгоритма автономии»

Большие языковые модели в настоящее время нашли широкое применение во многих сферах – начиная с маркетинга и кончая здравоохранением. Они отвечают за подбор персонализированных предложений на торговых и стриминговых онлайн-платформах. Они же помогают нам достраивать поисковые запросы и адекватно переводить тексты. Благодаря им развиваются различные

платформы, позволяющие получить профессиональные ответы на медицинские вопросы, в частности, Med-PaLM [Singhal et al., 2023].

Вместе с тем использование больших языковых моделей в медицине порождает новые вызовы и ставит вопрос о необходимости ограничения возможностей ИИ. Сбор данных для обучения нейросетей может привести к нарушению приватности, неправильная настройка ИИ может способствовать распространению дезинформации. При отсутствии должного контроля нейросети проявляют тенденцию к воспроизведению общих предрассудков [Li, Bamman, 2022].

Помимо возможностей, связанных с распространением информации или анализом данных, способность к автогенерации текстов (впервые представленная нейросетью GPT-2 в 2018 г.) может сыграть существенную роль во внедрении функционала БЯМ в медицинские практики. Разговорная функция нейросетей может быть задействована в создании искусственных помощников, сопровождающих процедуры информированного согласия. Подобное использование БЯМ позволит пациентам получить доступ к релевантной информации о медицинских процедурах и улучшить процесс информированного согласия [Allen et al., 2024].

Интерактивные свойства больших языковых моделей и возможность обрабатывать большие пласты информации также могут быть использованы для решения проблем, связанных со статусом полностью или частично недееспособных пациентов (в силу заболевания, травмы, когнитивных проблем или даже просто преклонного возраста).

В подобных случаях решение относительно лечения зачастую приходится принимать врачам или «суррогатам». Делегирование полномочий по принятию решения при этом рассматривается как необходимый шаг, который, однако, не дает полной уверенности в соблюдении принципа автономии. В ситуации, когда не было дано четких инструкций в виде заблаговременного распоряжения или плана ухода, принимать решения за другого человека особенно трудно. Находясь под давлением моральных обязательств перед недееспособным агентом и испытывая эмоциональное напряжение, «суррогаты» могут принимать решения, не соответствующие реальным предпочтениям и ценностям человека [Lamanna, Byrne, 2018, p. 903–904].

Разработка автоматизированных помощников для принятия решений представляется как шаг на пути к устранению данных проблем и как способ снять три бремени: этическое – способствуя соблюдению принципа автономии; эмоциональное – помогая «суррогатам» в принятии сложного решения; и экономическое, так как позволяет не проводить лечение, от которого бы пациент отказался [Ibid., p. 902].

Возможности, предоставляемые ИИ, позволяют создавать предикторы, охватывающие большие объемы информации и работающие в более понятной интерактивной форме.

Возможности больших языковых моделей по работе с большими данными позволяют добавить к результатам таргетных эмпирических исследований возможных предпочтений пациентов более широкую информацию за счет использования сведений из медицинских электронных карт и социальных сетей.

Данные медицинских карт могут быть использованы для анализа информации о решениях дееспособных пациентов, находившихся в схожей с недееспособным агентом ситуации. И в случае необходимости, например, назначения определенной медицинской процедуры, подобная аналитика позволит дать оценку вероятности согласия или несогласия на ее проведение. Данные из социальных сетей могут оказать помощь в выявлении соотношения черт характера человека и его возможных предпочтений. Благодаря такому подходу можно будет говорить о создании «алгоритма автономии», играющего важную роль в повышении точности и доступности предикции решения пациентов относительно лечения [Lamanna, Byrne, 2018, p. 904–906].

Вместе с тем есть основания сомневаться в возможностях «алгоритма автономии» должным образом поддерживать принцип уважения автономии.

Сомнение в первую очередь вызывает техническая сторона идеи алгоритма автономии – в условиях неполной ясности работы нейросетей (т.н. «черного ящика нейросетей») непонятно, насколько может быть верифицирован и проверен тот или иной вывод большой языковой модели [Biller-Andorno, Biller, 2019].

Во-вторых, не очень понятно, насколько может быть реализована этическая составляющая задачи «алгоритма автономии». Основной проблемой здесь является тенденция воспроизведения нейросетями «мудрости толпы», общих предрассудков. Указание на статистически высокую вероятность определенного типа решений для конкретного пациента не означает, что он действительно примет это решение или что он обязан его принять.

В контексте регулятивной функции концепции автономии ключевым аспектом является соблюдение права человека на принятие персонального решения. Решение может считаться персональным, если оно основано на индивидуальных предпочтениях агента, а не только на общих рассуждениях о том, какой сценарий является для него оптимальным или соответствует усредненному образу человека его возраста, профессии и других характеристик.

Важность индивидуальных предпочтений в принятии автономных решений подчеркивает необходимость более глубокого анализа концепции личности в рамках принципа автономии. Ярким доказательством является полемика о контрактах Улисса (когда люди утверждают решения относительно будущего лечения, пока они дееспособны, получая гарантию, что эти решения будут выполнены, даже если впоследствии они не смогут делать рациональный выбор [Антипов, 2023]), в рамках которой во главу угла ставится проблема личного выбора. Обоснование операциональности контракта Улисса в соблюдении принципов биоэтики в рамках принятия решений по большей части зависит от формулировки роли личности в автономии [Quante, 1999].

Персонализированный ИИ и проблема неконсистентности суждений

Немаловажным приобретением, сопутствующим развитию и распространению больших языковых моделей, стала способность обучаться на уже имеющихся текстах любого авторства с возможностью воспроизводить не только

стиль письма, но и направление мысли конкретного человека [Schwitzgebel et al., 2024].

Наиболее популярной нишей применения данной способности стало конструирование ботов, обученных на текстах именитых авторов и философов (GPT, основанный на работах Дэниела Деннета, Лукреций-GPT, Платон-GPT). Одной из целей создания подобных цифровых двойников является воспроизведение испытаний наподобие теста Тьюринга [Ibid., p. 238]. Такие нейросети могут выполнять и развлекательную функцию, предоставляя возможность побеседовать или получить совет от философа или исторического персонажа.

Важной особенностью подобных персонализированных нейросетей является то, что они основываются на методе обучения «тонкой настройки», достаточно экономичном в плане финансов и времени. «Тонкая настройка» позволяет адаптировать большую языковую модель для решения конкретных задач, не требуя ее разработки с нуля. Для этого берется модель БЯМ, уже натренированная на большом массиве данных. Верхние слои такой базовой модели для конкретных задач донастраиваются на небольшом количестве данных [Church et al., 2021]. Хорошим примером является бот Деннет-GPT, базовая модель которого была обучена на текстах из Википедии, а верхний слой был донастроен за счет обучения на работах философа [Ibid., p. 239].

Использовать данные возможности по воспроизведению персональных суждений при помощи больших языковых моделей, а также простоту их настройки предлагают авторы концепции проекта программы персонализированного предиктора предпочтений пациента автономии, или сокращенно – П4 (“A Personalized Patient Preference Predictor”, “P4”) [Earp et al., 2024].

Как и предиктор предпочтений пациента (ППП) и «алгоритм автономии», основанные на статических данных, П4 также направлен на то, чтобы облегчить процесс принятия решения в ситуации недееспособности пациента. При этом проект П4 ориентирован на воспроизведение предпочтений и ценностей конкретного человека за счет тонкой настройки БЯМ на основе личной информации, доступной о нем. Таким образом, «суррогаты» и медицинские специалисты смогут получить доступ к боту, который в идеале сможет отвечать на поставленные вопросы так же, как и недееспособный пациент [Ibid., p. 16–19]. Стоит, однако, учитывать, что создание подобного типа языковой модели требует четко сформулированного подхода к подбору информации о пациенте: как и откуда она может быть извлечена, какие источники могут считаться релевантными, а какие – нет.

В дополнение к данным, которые могут быть использованы в неперсонализированных стратегиях прогноза, авторы П4 предлагают несколько вариантов сбора личной информации о пациенте, из которых можно выделить четыре основных:

1) на основе текстов, созданных индивидуумом (например, электронные письма, посты в блогах);

2) на основе ответов на опросы или интервью для выявления предпочтений относительно лечения;

3) на основе данных, предоставленных «суррогатами», – нарративов близких людей о предпочтениях недееспособного пациента;

4) на основе выборов специально сконструированных мысленных экспериментов, направленных на выявление ценностей участника.

Авторы П4 также предлагают «геймифицированый» вариант последнего способа – наподобие эксперимента «Moral machine», участники которого решали задачи наподобие дилеммы вагонетки, выбирая, кого спасти – пассажиров автомобиля или пешеходов [Earp et al., p. 16].

Когда речь идет о воспроизведении персональных суждений, значимой становится проблема выборки входного материала. Авторы П4 также останавливаются на этом вопросе, отмечая важность сохранения приватности и учета личных пожеланий относительно того, какие данные стоит использовать, а какие – нет. Но даже если пациент дал согласие на использование информации из соцсетей заранее, все равно сохраняется проблема спецификации материала для обучения.

В критике проекта П4 отмечается, что данные из реакций и текстов, опубликованных в соцсетях, нуждаются в дополнительной обработке и прояснении. Только после этого их можно использовать в качестве материала для обучения нейросетей. Мнение, высказанное в соцсетях о ситуации с лечением другого человека, – например, «я считаю, что в таком состоянии лечение должно быть прекращено», – может не соответствовать пожеланиям относительно себя [Meier, 2024].

Хорошим примером такого противоречия может послужить исследование, направленное на выявление предпочтений людей в отношении моральных принципов автономных транспортных средств. В ходе опроса, в котором моделировались сценарии аварий с участием автономных автомобилей, исследователи обнаружили, что несмотря на то, что большинство участников изначально предпочитали утилитарные варианты решений для минимизации количества жертв, их взгляды полностью менялись, как только по сценарию они сами оказывались в машине (зачастую не в пользу пешеходов) [Cecchini et al., 2023].

Данный пример показывает, что для текстов и реакций в онлайн-сетях может понадобиться классификация и ранжировка, чтобы провести различение: какие решения человек считает правильными в общем и в целом и какие он предпочтет, чтобы приняли относительно него лично.

При этом стоит понимать, что идиосинкратические предпочтения также могут противоречить друг другу. Мы можем столкнуться с несоответствием между тем, что человек полагает правильным для себя, и тем, что человеку нравится или что его отвращает. Данное противоречие, рассмотренное как противопоставление убеждений и оценочных установок, будет разобрано в следующем разделе.

Измерение оценочных установок как часть персонализации ИИ

Обсуждение критериев реализации принципа автономии часто сопровождается вопросами, связанными с персональной идентичностью. Нужно ли обращать внимание на целостность автобиографического нарратива в формулировке автономного решения? Если решение сильно отличается от «характера»

человека и от его предыдущих действий и высказываний, надо ли проявлять беспокойство о нарушении автономии в силу каких-то «внешних» сил? Что делать, если человек часто будет высказывать противоречивые суждения и пожелания относительно своего лечения в разные моменты времени? [Childress, 1990].

Технологии наподобие П4 призваны помочь «суррогатам» и медицинским специалистам преодолеть сформулированные выше затруднения. Но хотя в критике проекта П4 и обсуждается противоречие между тем, что человек считает правильным решением, и тем, как бы он поступил, пока нет инструкции по выбору высказываний и суждений, которые могут стать основой для автономного решения.

Одним из ярких примеров, когда конкретизация персонального решения в условиях недееспособности агента могла бы снять груз ответственности с родственников и врачей, является казус «кофейной эвтаназии» (в англоязычной литературе – “Dutch euthanasia case”).

«Кофейная эвтаназия» является знаковой историей с участием недееспособной пациентки с диагностированной тяжелой деменцией. 74-летняя пациентка выразила желание об эвтаназии по наступлению состояния тяжелой деменции в предварительном распоряжении (т.н. advance euthanasia directive, AED) после того, как в 2012 г. ей поставили диагноз «болезнь Альцгеймера». Незадолго до ухудшения состояния она устно сообщила врачу, что еще не готова умирать, но не отозвала предварительное распоряжение. Когда ее состояние ухудшилось, врач проконсультировался с другими медицинскими работниками, подтвердил волю пациентки, ориентируясь на предварительное распоряжение об эвтаназии. Перед эвтаназией врач ввел в организм пациентки успокоительное (смешал его с ее утренним кофе). Все шаги заранее оговаривались с семьей пациентки, но фактическая эвтаназия в тот конкретный момент времени с пациенткой не обсуждалась. Во время процедуры пациентка пришла в сознание, и до завершения ее удерживали члены семьи [Asscher et al., 2020].

Данный казус, так же как и проблема выборки материала для обучения П4, может быть раскрыт через сложившуюся в социальной психологии традицию сопоставления оценочных установок (attitudes) и убеждений (beliefs) – особых психологических конструктов, позволяющих описать поведение человека, в том числе и в этическом контексте. Оценочные установки представляют собой психологическую тенденцию или предрасположенность, выраженную в оценке конкретного объекта с той или иной степенью симпатии или антипатии («мне нравится» или «я не одобряю»). Убеждения подразумевают, что человек полагает что-то истинным [Guerin, 1994] или же (в этическом контексте данного понятия) что человек считает что-то правильным [Chignell, 2018].

Если рассматривать с этой точки зрения казус «кофейной эвтаназии», то мы можем увидеть несовпадение убеждений, сформулированных в предварительном распоряжении, и оценочных установок, выраженных в конкретный момент времени уже до наступления недееспособного состояния. Можно предположить, что если обучение П4 было бы основано на суждениях

пациентки, иллюстрирующих убеждения о благой жизни и достойной смерти, то он бы порекомендовал поступить так же, как и врачи. Однако, если в контур обучения были бы включены такие индивидуальные оценочные установки, как избегание необратимых решений или любовь к простым рутинным практикам, решение П4 могло бы стать противоположным.

Важным в контексте потенциальной подготовки данных для П4 является то, что в социальной психологии оценочные установки представляются как измеримые свойства. Впервые методика измерения ценностных установок была предложена еще 1928 г. Луисом Леоном Тёрстоуном, пионером в области психометрики [Thurstone, 1928]. Хотя предложенная им методика на данный момент практически не используется в силу громоздкости, данная статья является началом движения в данном направлении.

Уже через несколько лет Ренсис Лайкерт значительно упростил его подход и предложил инструмент для измерения оценочных установок, которым часто пользуются в социальной психологии и по сей день. Помимо этого, Лайкерт описал связь ценностных установок с представлением о природе человеческой личности. Признавая, что оценочная установка представляет собой привычку, которая достаточно компактна и устойчива и потому может рассматриваться как единица, Лайкерт развивает дискуссию о конкретности и общности персональных черт. Первая позиция исходит из того, что человеческая личность является набором разнородных специфических черт. Вторая позиция настаивает на единстве человеческого характера. Благодаря данному единству социальный психолог может сформировать общую картину характера человека по нескольким точно определенным оценочным установкам. Это также должно позволить предсказывать поведение личности в широком ряду разнообразных контекстов [Likert, 1932].

Несмотря на то что традиции сопоставления убеждений и оценочных суждений в социальной философии уже более полувека, все же до сих пор нет общего согласия относительно того, можно ли напрямую рядополагать их пропозициональное содержание и, соответственно, ставить вопрос об их консистентности [Fishbein, 1966]. Эту неконсистентность нельзя обосновать незнанием этических норм – что подтверждают исследования моральных позиций академических философов. Этот конфликт может быть обнаружен в жизни любого человека [Schwitzgebel, Rust, 2014].

В контексте выборки данных для тренировки П4 важными свойствами концепции оценочных суждений является то, что они могут представлять собой суждения, противоречащие убеждениям. Они являются стабильной частью личности, связанной с другими чертами характера, что позволяет говорить об их важности для соблюдения критериев автономного выбора. В социальной психологии есть подходы, благодаря которым личные оценочные установки могут быть измерены, что может способствовать повышению прозрачности обучения и настройки П4.

Заключение

Развитие технологий искусственного интеллекта и их использование ставит новые задачи на пути артикуляции критериев принципа автономии пациента.

С одной стороны, технологии наподобие «алгоритма автономии» могут помочь сохранению принципа автономии при недееспособности пациента, предоставляя информацию о вероятностных сценариях решения пациента. Например, такие технологии могут быть вписаны в регламент реализации делегированной или ассистированной автономии в качестве дополнительного источника сведений для совместного принятия решений. «Алгоритм автономии» представлен его авторами как способ облегчить груз моральной ответственности и избавить от эмоционального выгорания «суррогатов», принимающих решение. Тем не менее технологии ИИ, работающие на основе статистических корреляций и анализа больших данных, могут не учитывать такие важные для принципа автономии свойства, как личные ценности агента и необходимость реализации свободного выбора.

Персонализированный предиктор предпочтений пациента, наоборот, ориентирован на воспроизведение личных ценностей агента. Искусственное воспроизведение черт характера и стиля мышления акцентирует проблему личных предпочтений в автономном решении. Задачи по выборке материала для обучения персонализированных нейросетей ставят перед нами проблему противоречивости высказываний пациента, которую можно представить как неконсистентность убеждений и оценочных установок.

Оценочные установки могут не соответствовать представлениям самого агента о правильном решении – его убеждениям, но являются частью его личности и, соответственно, могут повлиять на процесс принятия решений. Наработки социальной психологии в области измерения оценочных установок позволяют с осторожностью говорить о том, что данные, собранные при помощи подобной психометрики, могут быть использованы в обучении П4 для придания личностных свойств персонализированной нейросети. Данное решение (конечно, не окончательное и не единственное) может стать еще одним шагом на пути к разработке расширенной за счет технологии ИИ автономии, а также стать поводом для дискуссий о роли характера личности в автономных решениях.

Список литературы

Антипов, 2023 – *Антипов А.В.* Биоэтические аспекты контракта Улисса (на примере превенции суицида) // Вестник Томского государственного университета. Философия. Социология. Политология. 2023. № 73. С. 71–78.

Артемьева, 2018 – *Артемьева О.В.* Универсальность и автономия в этике И. Канта // Философские науки. 2018. № 11. С. 86–102.

Белялетдинов, 2019 – *Белялетдинов Р.Р.* Риски современных биотехнологий: социогуманитарный анализ: монография. М.: ООО «4 Принт», 2019. 212 с.

Allen et al., 2024 – *Allen J.W., Earp B.D., Koplin J., Wilkinson D.* Consent-GPT: is it ethical to delegate procedural consent to conversational AI? // Journal of Medical Ethics. 2024. No. 50. P. 77–83.

Asscher, Alida, Vathorst, 2020 – *Asscher E., Alida C., van de Vathorst S.* First prosecution of a Dutch doctor since the Euthanasia Act of 2002: what does the verdict mean? // Journal of medical ethics. 2020. Vol. 46. No. 2. P. 71–75.

Biller-Andorno, Biller, 2019 – *Biller-Andorno N., Biller A.* Algorithm-Aided Prediction of Patient Preferences – An Ethics Sneak Peek // The New England journal of medicine. 2019. Vol. 381. No. 15. P. 1480–1485.

Cecchini, Brantley, Dubljević, 2023 – *Cecchini D., Brantley S., Dubljević V.* Moral judgment in realistic traffic scenarios: moving beyond the trolley paradigm for ethics of autonomous vehicle // AI & Soc. 2023. URL: https://link.springer.com/article/10.1007/s00146-023-01813-y (accessed on: 28.09.2024).

Chignell, 2018 – *Chignell A.* The Ethics of Belief // The Stanford Encyclopedia of Philosophy. 2018. URL: https://plato.stanford.edu/archives/spr2018/entries/ethics-belief/ (accessed on: 28.09.2024)

Childress, 1990 – *Childress J.F.* The Place of Autonomy in Bioethics // The Hastings Center Report. Vol. 20. No. 1. 1990. P. 12–17.

Childress, Beauchamp, 2001 – *Childress J.F., Beauchamp T.L.* Principles of Biomedical Ethics. 5th ed. Oxford: Oxford University Press, 2001. 455 p.

Church, Chen, Ma, 2021 – *Church K.W., Chen Z., Ma Y.* Emerging trends: A gentle introduction to fine-tuning // Natural Language Engineering. 2021. Vol. 27. No. 6. P. 763–778.

DeGrazia, Millum, 2021 – *DeGrazia D., Millum J.* Autonomy // A Theory of Bioethics. Cambridge: Cambridge University Press, 2021. P. 97–137.

Earp et al., 2024 – *Earp B.D., Porsdam Mann S., Allen J., Salloch S., Suren V., Jongsma K., Braun M. et al.* A Personalized Patient Preference Predictor for Substituted Judgments in Healthcare: Technically Feasible and Ethically Desirable // The American Journal of Bioethics. 2024. Vol. 24. No. 7. P. 13–26.

Fishbein, 1966 – *Fishbein M.* The relationships between beliefs, attitudes and behavior // Cognitive consistency, motivational antecedents and behavioral consequents / Ed. by S. Feldman. New York: Academic Press, 1966. P. 199–223.

Guerin, 1994 – *Guerin B.* Attitudes and beliefs as verbal behavior // The behavior analyst. 1994. No. 17. P. 155–163.

Lamanna, Byrne, 2018 – *Lamanna C., Byrne L.* Should Artificial Intelligence Augment Medical Decision Making? The Case for an Autonomy Algorithm // AMA journal of ethics. 2018. Vol. 20. No. 9. P. 902–910.

Li, Bamman, 2022 – *Li L., Bamman D.* Gender and Representation Bias in GPT-3 Generated Stories // Proceedings of the Third Workshop on Narrative Understanding. Stroudsburg: ACL, 2022. P. 48–55.

Likert, 1932 – *Likert R.* A technique for the measurement of attitudes. New York: New York University, 1932. 55 p.

Meier, 2024 – *Meier L.J.* Predicting Patient Preferences with Artificial Intelligence: The Problem of the Data Source // The American Journal of Bioethics. 2024. Vol. 24. No. 7. P. 48–50.

Quante, 1999 – *Quante M.* Precedent autonomy and personal identity // Kennedy Institute of Ethics journal. 1999. Vol. 9. No. 4. P. 365–381.

Rid, Wendler, 2014 – *Rid A., Wendler D.* Treatment decision making for incapacitated patients: is development and use of a patient preference predictor feasible? // J Med Philos. 2014. Vol. 39. No. 2. P. 130–152.

Schwitzgebel E., Schwitzgebel D., Strasser, 2024 – *Schwitzgebel E., Schwitzgebel D., Strasser A.* Creating a large language model of a philosopher // Mind & Language. 2024. Vol. 39. No. 2. P. 237–259.

Schwitzgebel, Rust, 2014 – *Schwitzgebel E., Rust J.* The moral behavior of ethics professors: Relationships among self-reported behavior, expressed normative attitude, and directly observed behavior // Philosophical Psychology. 2014. Vol. 27. No. 3. P. 293–327.

Singhal et al., 2023 – *Singhal K., Azizi S., Tu T., Mahdavi S.S., Wei J., Chung H.W., Scales N., Tanwani A., Cole-Lewis H. et al.* Large language models encode clinical knowledge // Nature. 2023. No. 7972. P. 172–180.

Thurstone, 1928 – *Thurstone L.L.* Attitudes can be measured // American Journal of Sociology. 1928. No. 33. P. 529–554.

Patient autonomy in the context of AI personalization technologies: new approaches and ethical challenges

Sofya V. Lavrentyeva

RAS Institute of Philosophy. 12/1 Goncharnaya Str., Moscow, 109240, Russian Federation; e-mail: sonnig89@gmail.com

This article explores the ethical challenges related to using large language models in medical decision-making for incapacitated patients. It highlights how advancements in artificial intelligence (AI) raise questions about the application of the principle of patient autonomy in healthcare. AI systems, which analyze behavioral patterns from big data and simulate personal traits, can be used to support decision-making. However, these models cannot fully capture an individual's unique values and personality, raising concerns about whether the principle of autonomy is truly upheld. One solution could be the development of a personalized patient preference predictor (P4) – a large language model trained on personal data to reflect the individual's preferences. Yet, the creation of 4P requires careful consideration of the data used for training. A key challenge is the potential inconsistency in the individual's beliefs and attitudes across different contexts, such as written communications, social media, and verbal statements. Social psychology research suggests that psychometric data could be used to better model attitudes in training AI. This approach could pave the way for AI-enhanced autonomy and spark discussions about the role of personality in autonomous decision-making.

Keywords: artificial intelligence, autonomy, large language models, predictor, attitudes, beliefs, subjectivity

Acknowledgments: The reported study was funded by the Russian Science Foundation according to the research project No. 24-28-01604.

References

Antipov, A.V. "Bioeticheskiye aspekty kontrakta Ulissa (na primere preventsii suitsida)" [Bioethical aspects of the Ulysses contract (on the example of suicide prevention)], *Vestnik Tomskogo gosudarstvennogo universiteta. Filosofiya. Sotsiologiya. Politologiya*, 2023, no. 73, pp. 71–78. (In Russian)

Artemyeva, O.V. "Universal'nost' i avtonomiya v etike I. Kanta" [Universality and Autonomy in Kant's Moral Philosophy], *Filosofskie nauki*, 2018, no. 11, pp. 86–102. (In Russian)

Asscher, E., Alida, C., van de Vathorst, S. "First prosecution of a Dutch doctor since the Euthanasia Act of 2002: what does the verdict mean?", *Journal of Medical Ethics*, 2020, vol. 46, no. 2, pp. 71–75.

Belyaletdinov, R.R. *Riski sovremennykh biotekhnologiy: sotsiogumanitarnyy analiz* [Risks of modern biotechnologies: sociohumanitarian analysis]. Moscow: "4 Print" Publ., 2019. 212 pp. (In Russian)

Biller-Andorno, N., Biller, A. “Algorithm-Aided Prediction of Patient Preferences An Ethics Sneak Peek”, *The New England Journal of Medicine*, 2019, vol. 381, no. 15, pp. 1480–1485.

Cecchini, D., Brantley, S., Dubljević, V. “Moral judgment in realistic traffic scenarios: moving beyond the trolley paradigm for ethics of autonomous vehicle”, *AI & Society*, 2023. URL: https://link.springer.com/article/10.1007/s00146-023-01813-y (accessed on: 28.09.2024).

Chignell, A. “The Ethics of Belief”, *The Stanford Encyclopedia of Philosophy*, 2018. URL: https://plato.stanford.edu/archives/spr2018/entries/ethics-belief/ (accessed on: 28.09.2024).

Childress, J.F. “The Place of Autonomy in Bioethics”, *The Hastings Center Report*, vol. 20, no. 1, 1990, pp. 12–17.

Childress, J.F., Beauchamp, T.L. *Principles of Biomedical Ethics*. 5th ed. Oxford: Oxford University Press, 2001. 455 pp.

Church, K.W., Chen, Z., Ma, Y. “Emerging trends: A gentle introduction to fine-tuning”, *Natural Language Engineering,* 2021, vol. 27, no. 6, pp. 763–778.

DeGrazia, D., Millum, J. “Autonomy”, in: *A Theory of Bioethics*. Cambridge: Cambridge University Press, 2021, pp. 97–137.

Earp, B.D., Porsdam Mann, S., Allen, J., Salloch, S., Suren, V., Jongsma, K., Braun, M. et al. “A Personalized Patient Preference Predictor for Substituted Judgments in Healthcare: Technically Feasible and Ethically Desirable”, *The American Journal of Bioethics*, 2024, vol. 24, no. 7, pp. 13–26.

Fishbein, M. “The relationships between beliefs, attitudes and behavior”, in: *Cognitive consistency, motivational antecedents and behavioral consequents*, ed. by S. Feldman. New York: Academic Press, 1966, pp. 199–223.

Guerin, B. “Attitudes and beliefs as verbal behavior”, *The Behavior Analyst*, 1994, vol. 17, pp. 155–163.

Lamanna, C., Byrne, L. “Should Artificial Intelligence Augment Medical Decision Making? The Case for an Autonomy Algorithm”, *AMA Journal of Ethics*, 2018, vol. 20, no. 9, pp. 902–910.

Li, L., Bamman, D. “Gender and Representation Bias in GPT-3 Generated Stories”, in: *Proceedings of the Third Workshop on Narrative Understanding*. Stroudsburg: ACL, 2022, pp. 48–55.

Likert, R. *A technique for the measurement of attitudes*. New York: New York University Press, 1932. 55 pp.

Meier, L.J. “Predicting Patient Preferences with Artificial Intelligence: The Problem of the Data Source”, *The American Journal of Bioethics*, 2024, vol. 24, no. 7, pp. 48–50.

Quante, M. “Precedent autonomy and personal identity”, *Kennedy Institute of Ethics Journal*, 1999, vol. 9, no. 4, pp. 365–381.

Rid, A., Wendler, D. “Treatment decision making for incapacitated patients: is development and use of a patient preference predictor feasible?”, *Journal of Medical Philosophy*, 2014, vol. 39, no. 2, pp. 130–152.

Schwitzgebel, E., Rust, J. “The moral behavior of ethics professors: Relationships among self-reported behavior, expressed normative attitude and directly observed behavior”, *Philosophical Psychology*, 2014, vol. 27, no. 3, pp. 293–327.

Schwitzgebel, E., Schwitzgebel, D., Strasser, A. “Creating a large language model of a philosopher”, *Mind & Language*, 2024, vol. 39, no. 2, pp. 237–259.

Singhal, K., Azizi, S., Tu, T., Mahdavi, S.S., Wei, J., Chung, H.W., Scales, N., Tanwani, A., Cole-Lewis, H. et al. “Large language models encode clinical knowledge”, *Nature*, 2023, no. 7972, pp. 172–180.

Thurstone, L.L. “Attitudes can be measured”, *American Journal of Sociology*, 1928, no. 33, pp. 529–554.