Цифровая экономика и искусственный интеллект
Digital economy and artificial intelligence

УДК 004.8; 004.89; 316.77

DOI: 10.55959/MSU2070-1381-116-2026-100-111

Методология обучения ИИ-агентов для оценки видеороликов с имитацией оценки человеком: социологический аспект

Григорьева Наталия Сергеевна
Доктор политических наук, профессор, заведующий кафедрой социологии управления, SPIN-код РИНЦ: 9017-7352, ORCID: 0000-0002-7707-6754, grigorieva@spa.msu.ru

Факультет государственного управления МГУ имени М.В. Ломоносова. Москва, РФ.

Крупенко Мария Анатольевна
Соискатель, SPIN-код РИНЦ: 7625-7269, ORCID: 0009-0006-9999-0228, mmmasha1999@yandex.ru

Факультет государственного управления МГУ имени М.В. Ломоносова. Москва, РФ.

Аннотация
В статье представлена социологически обоснованная методология обучения ИИ-агентов для оценки видеороликов, имитирующей оценку, данную человеком (человеческой оценки), с учетом социально обусловленных моделей восприятия контента реальными пользователями. Следует отметить, что статей непосредственно по методологии обучения ИИ-агентов для имитации человеческой оценки видеороликов с социологической оценкой в журналах за 2021–2026 годы практически нет, но есть близкие публикации (в зарубежных и русских источниках), авторы которых фокусируются на ИИ-оценке социальных ситуаций в видеоиграх и социологической симуляции поведения. В данной статье затронута проблема расхождения между алгоритмической оценкой и восприятием видео различными социальными группами, возникающая из-за ориентации алгоритмов на формализованные метрики и игнорирования социокультурных особенностей зрительского восприятия. В основе методологии лежат теоретические подходы символического интеракционизма, теории социальных представлений и социального конструктивизма, а также концепции цифровой социологии и теории медиавосприятия. Методология опирается на четыре ключевых принципа: социальную репрезентативность данных, моделирование социальных процессов, учет вариативности восприятия и прозрачность решений искусственного интеллекта (ИИ). Представлена система социологических критериев для оценки человекоподобности решений ИИ. Предложены механизмы валидации результатов, включающие расчет коэффициента согласия между оценками ИИ и социальных групп; определение доли решений ИИ, которые пользователи не могут отличить от человеческих; анализ процента снижения апелляций по сравнению с традиционными системами, а также оценку индекса культурной адаптивности, а именно способности модели корректно работать в разных социокультурных средах. Такой подход позволяет преодолеть разрыв между алгоритмической и социальной оценкой видеоконтента, а ее внедрение повысит релевантность ИИ-систем за счет учета групповых различий в зрительском восприятии, что поспособствует созданию более сбалансированных и социально адекватных решений в цифровом медиапространстве.

Ключевые слова
ИИ-агенты, оценка видеоконтента, социальная оценка, символический интеракционизм, теория социальных представлений, цифровая социология, медиавосприятие.

Для цитирования
Григорьева Н.С., Крупенко М.А. Методология обучения ИИ-агентов для оценки видеороликов с имитацией оценки человеком: социологический аспект //Государственное управление. Электронный вестник. 2026. № 115. С. 100–111. DOI: 10.55959/MSU2070-1381-116-2026-100-111

Methodology of Training AI Agents To Evaluate Videos with Imitation of Human Evaluation: A Sociological Aspect

Natalia S. Grigorieva
DSc (Political Sciences), Professor, Head of the Department of Management Sociology, ORCID: 0000-0002-7707-6754, grigorieva@spa.msu.ru

School of Public Administration, Lomonosov Moscow State University, Moscow, Russian Federation.

Maria A. Krupenko
PhD applicant, ORCID: 0009-0006-9999-0228, mmmasha1999@yandex.ru

School of Public Administration, Lomonosov Moscow State University, Moscow, Russian Federation.

Abstract
This article presents a sociologically grounded methodology for training AI agents to rate videos, simulating human ratings while taking into account socially conditioned models of content perception by real users. It should be noted that there are practically no articles directly on the methodology of training AI agents to simulate human evaluation of videos with sociological evaluation in journals for 2021–2026, but there are similar publications (in foreign and Russian sources), the authors of which focus on AI assessment of social situations in video games and sociological simulation of behaviour. The study addresses the discrepancy between algorithmic ratings and video perception by different social groups, which arises due to the algorithms' reliance on formalized metrics

and their ignorance of the sociocultural characteristics of viewer perception. The methodology draws on theoretical approaches from symbolic interactionism, social representation theory, and social constructivism, as well as concepts from digital sociology and media perception theory. It relies on four key principles: social representativeness of data, modeling of social processes, accounting for variability in perception, and transparency of artificial intelligence (AI) decisions. A system of sociological criteria for assessing the humanlikeness of AI decisions has been developed. Mechanisms for validating results are proposed, including calculating the agreement coefficient between AI and social group assessments, determining the proportion of AI decisions that users cannot distinguish from human ones, analyzing the percentage reduction in appeals compared to traditional systems, and assessing the cultural adaptability index, namely, the model's ability to operate correctly in different sociocultural contexts. This methodology will bridge the gap between algorithmic and social assessment of video content, and its implementation will increase the relevance of AI systems by taking into account group differences in viewer perception, thereby contributing to the creation of more balanced and socially appropriate decisions in the digital media space. In the long term, this will allow for more sustainable content evaluation practices that focus not only on formal criteria, but also on the dynamics of social norms and audience values.

Keywords
AI agents, video content evaluation, social evaluation, symbolic interactionism, social representation theory, digital sociology, media perception.

For citation
Grigorieva N.S., Krupenko M.A. (2026) Methodology of Training AI Agents To Evaluate Videos with Imitation of Human Evaluation: A Sociological Aspect. *Gosudarstvennoye upravleniye. Elektronnyy vestnik*. No. 115. P. 100–111. DOI: 10.55959/MSU2070-1381-116-2026-100-111

Дата поступления/Received: 23.05.2026

Введение

Современное медиапространство претерпевает существенную трансформацию: видеоконтент становится преобладающим типом информационного контента и ключевым каналом передачи смыслов, а алгоритмы искусственного интеллекта (ИИ) активно включаются в изучение того, как разные аудитории воспринимают и оценивают такие материалы. В то же время развитие ИИ-систем традиционно смещено в сторону технической оптимизации, сами алгоритмы совершенствуются для более точного подсчета метрик вовлеченности (длительность просмотра, лайки, репосты), но не для понимания смыслов, которые аудитория вкладывает в оценку контента. Принципиальный недостаток подобных подходов состоит в том, что они не учитывают социокультурную обусловленность восприятия и что восприятие является не механической реакцией на аудиовизуальные стимулы, а сложным процессом социального взаимодействия. Одно и то же видео может получить абсолютно противоположные оценки в разных социальных группах из-за различий в культурных кодах, ценностных установках, возрастных и эстетических предпочтениях. Расхождение между алгоритмической и социальной оценкой приводит к росту ошибочных решений, усилению поляризации мнений и ограничению культурного разнообразия. В современной практике подготовки ИИ-агентов доминируют методики, опирающиеся на массивы данных, фиксирующих лишь очевидные, лежащие на поверхности поведенческие шаблоны пользователей. При этом глубинные процессы, посредством которых в обществе складываются коллективные оценочные суждения, остаются за рамками анализа. Существующие алгоритмы не только не преодолевают, но зачастую закрепляют и даже усиливают уже имеющиеся когнитивные искажения. Их архитектура не предусматривает механизмов отслеживания подвижности социальных норм, равно как и инструментов для учета разнородности зрительского восприятия, ведь опыт взаимодействия с контентом существенно варьируется в зависимости от социокультурного контекста аудитории. Предлагаемая в работе методологическая альтернатива исходит из необходимости воспроизвести в логике работы ИИ-систем сам механизм человеческой оценки. Это предполагает целенаправленный учет комплекса социальных переменных от культурно обусловленных смыслов и групповых нормативных ожиданий до динамики коллективных представлений и моделей социального научения.

Ключевая цель исследования заключается в формировании такого инструментария оценки видеоконтента, при котором выводы ИИ-системы отражали бы не абстрактный усредненный

взгляд типичного зрителя, а реальные паттерны коллективных суждений, характерные для конкретных социальных групп. Это обусловливает необходимость концептуализировать процесс социальной оценки видеоконтента, выявить факторы, влияющие на восприятие разных аудиторий, сформулировать принципы учета социокультурных различий при обучении ИИ и разработать инструменты валидации человекоподобности его решений. Разработка такого инструментария может заложить основу для создания более справедливых, адаптивных и социально ответственных ИИ-систем, способных учитывать многообразие зрительского опыта.

Социологическая концептуализация оценки видеоконтента

Видеоконтент в современном обществе выступает не просто как информационный носитель, но как сложный семиотический объект, порождающий множественные смыслы в процессе социального взаимодействия. Соответственно, на восприятие и оценку контента влияют не только его объективные параметры, но и широкий круг социальных факторов, формирующих способы интерпретации у аудитории. Исследованием этих социальных факторов и занимается социология с учетом того, что большие языковые модели (LLM) способны имитировать ответы людей, принадлежащих к отдельным социальным группам, и такие модели можно использовать для проведения соцопросов, так как они могут имитировать наиболее типичные ответы респондентов с заданными характеристиками[1] [Пузанова и др. 2025].

Ключевым аспектом понимания видеоконтента является его семиотическая природа: видео содержит систему знаков и символов, интерпретация которых зависит от культурного кода зрителя. Согласно семиотическому подходу, визуальные образы несут конвенциональные значения, усвоенные в процессе социализации. Одни и те же жесты или мимика могут иметь различную трактовку в разных культурных контекстах: то, что в одной культуре воспринимается как дружелюбие, в другой может считываться как агрессия. В результате это создает предпосылки для вариативности оценок видеоконтента в зависимости от культурной принадлежности аудитории. В качестве примера межкультурных различий в невербальной коммуникации можно привести специфику жестов согласия и несогласия в Болгарии: в данной культурной традиции вертикальное движение головы (кивок сверху вниз) кодируется как выражение несогласия («нет»), а горизонтальное (поворот головы влево-вправо) — как выражение согласия («да»). Эта система жестов инвертирована относительно той, что распространена в большинстве мировых культур.

В рамках конструктивистского подхода [Бергер, Лукман 1995] подчеркивается, что зритель не пассивно воспринимает видео, а активно конструирует его значение через собственный опыт и социальные установки. Так, процесс просмотра видео — это не просто потребление информации, а взаимодействие между содержанием ролика и системой представлений зрителя. Более того, конструирование смыслов происходит на нескольких уровнях — от непосредственной интерпретации визуальных образов до соотнесения увиденного с социальными нормами и ценностями.

Особую роль в современных условиях играет визуальная коммуникация. В контексте «визуального поворота» [Mitchell 1994] видео становится доминирующим каналом социальной коммуникации, заменяя традиционные текстовые форматы в передаче культурных кодов. Визуальные образы обладают высокой степенью эмоциональности и запоминаемости, что делает их мощным способом формирования коллективных представлений. Соответственно, видеоконтент не только отражает существующие социальные нормы, но и активно участвует в их конструировании, задавая образцы поведения, эстетические предпочтения и ценностные ориентиры.

[1] Искусственный интеллект в социологии // SocioLogos [Электронный ресурс]. URL: https://sociologos.ru/blog/iskusstvennyy-intellekt-v-sotsiologii/ (дата обращения: 18.02.2026).

Социальные механизмы формирования оценок также оказывают существенное влияние на восприятие видео: возрастные, гендерные, профессиональные и этнические группы обладают специфическими паттернами восприятия. Например, молодежь будет больше оценивать видео по критериям креативности, оригинальности и соответствия актуальным трендам, тогда как старшее поколение в большинстве своем — по критериям информативности, полезности и соответствия традиционным ценностям.

Важную роль в формировании коллективных оценок играют референтные группы и лидеры мнений. Согласно теории двухступенчатого потока коммуникации [Lazarsfeld et al. 1944; Katz 1987], оценки лидеров мнений транслируются в широкие массы через социальные сети, формируя коллективные предпочтения. В цифровой среде этот механизм усиливается за счет алгоритмов рекомендаций, которые распространяют контент, получивший одобрение влиятельных пользователей; тем самым первичная оценка, данная лидером мнений, может стать основой для массового восприятия видео.

Выработка и установление критериев оценки видео также связаны с механизмами социального научения [Бандура 2000]. Процесс усвоения зрителями норм восприятия медиаконтента осуществляется посредством механизмов наблюдения и имитации поведения референтных групп, к которым относятся близкие социальные связи (друзья, родственники) и значимые публичные фигуры (кумиры). В цифровой среде данный социально-психологический механизм приобретает специфическую форму — он проявляется через распространение вирусных реакций на видеоматериалы. При этом тип оценки, получивший широкое одобрение в онлайн-сообществе, стремительно тиражируется другими пользователями, что в результате приводит к формированию устойчивого паттерна восприятия контента.

Культурные нормы очерчивают допустимые границы медиапотребления. Социальные правила и табу задают критерии приемлемости отдельных элементов видеоконтента в рамках конкретного социокультурного контекста. Например, сцены эротического характера в одной культуре могут считаться допустимыми, а в другой — подпадать под строгий запрет.

Эстетические предпочтения играют заметную роль в восприятии видеоконтента. Согласно концепции культурного капитала [Бурдье 2004], они выстраивают иерархию «высокого» и «низкого» искусства. Так, представители элитарных кругов чаще ценят артхаусное кино, документалистику и экспериментальные форматы, в то время как широкая аудитория тяготеет к развлекательным шоу, комедиям и блокбастерам. Эти различия в эстетических ориентирах напрямую сказываются на том, как люди оценивают тот или иной контент.

Моральные и этические нормы неодинаковы в разных социальных группах, и это напрямую влияет на то, как люди воспринимают видеоконтент. К примеру, в религиозных сообществах с настороженностью относятся к материалам, критикующим религиозные догмы, тогда как в светской среде такой контент могут оценить нейтрально или даже одобрить.

В цифровой среде механизмы формирования социальных оценок меняются. Коллективное мнение складывается в ходе обсуждений и взаимного влияния внутри онлайн-сообществ, где постепенно вырабатываются общие нормы восприятия контента. Существенные препятствия для адекватного понимания механизмов оценки контента создают эффекты поляризации и феномен «эхо-камер». Суть проблемы в том, что рекомендательные алгоритмы выстраивают ленту пользователя вокруг тематически близких материалов, что в результате приводит к тому, что внутри отдельных сообществ оценочные суждения постепенно выравниваются, тогда как иные, не совпадающие с доминирующей линией позиции фактически вытесняются из поля зрения.

Таким образом, анализ оценки видеоконтента демонстрирует, что восприятие и оценка видеороликов представляют собой не просто индивидуальную реакцию на аудиовизуальные стимулы, но и сложный, социально обусловленный процесс.

Ошибки автоматической оценки видеороликов

При конструировании методик обучения систем автоматической оценки видеоконтента, ориентированных на моделирование человеческого восприятия, существенное значение приобретает систематизация характерных ошибок, допускаемых в ходе автоматической модерации. Они показывают, где расходятся алгоритмический и человеческий подходы к анализу контента. В таких случаях программа реагирует на формальные признаки нарушения, а человек учитывает контекст, культурные нормы и намерения автора. Причина расхождений кроется в разном способе обработки информации, алгоритмы опираются на статистические закономерности и числовые показатели, тогда как люди воспринимают смысл, понимают цели высказывания и ориентируются на социальные правила. Анализ ошибок модерации помогает выявить слепые зоны системы и скорректировать обучающие данные и архитектуру модели.

Социологический аспект здесь выступает важным ориентиром, так как он позволяет соотнести ИИ-решения с социальными нормами и ценностными установками конкретных сообществ, а также выявить зоны наибольшего риска ошибочной интерпретации.

Проанализируем на примере, какие ошибки чаще всего допускают современные системы модерации, созданные с применением машинного обучения. Допустим, автоматизированная система анализирует набор видеороликов разной тематики и в ряде случаев проявляет ограниченность: реагирует на отдельные визуальные или текстовые элементы, но не понимает общего смысла сообщения. Например, видео кулинарного мастер-класса, где показан нож для нарезки овощей, может быть ошибочно заблокировано, так как система воспримет нож как демонстрацию оружия. Аналогично фрагмент военно-исторического фильма могут удалить из-за того, что он якобы пропагандирует насилие.

Следующим типом ошибок являются мультимодальные сбои; они возникают из-за того, что алгоритмы раздельно обрабатывают аудиовизуальные и текстовые компоненты видео [Новые подходы к оцениванию 2025]. Например, нейтральный закадровый комментарий может быть неверно соотнесен с визуальным рядом, что приводит к ложной маркировке ролика как нарушающего правила платформы.

Серьезное влияние на качество работы ИИ-агента оказывает отравление данных: если в обучающий набор попали специально искаженные примеры, модель усваивает ошибочные паттерны.

Распространенным сценарием является также избыточная строгость фильтров: стремясь минимизировать юридические риски, разработчики настраивают алгоритмы на гиперчувствительность. В итоге видео с косвенной отсылкой к спорной теме, например просветительский ролик о профилактике заболеваний, может быть автоматически заблокирован без глубокого анализа содержания.

Значимым искажением является распознавание юмора или иронии, и проявляется оно в неспособности алгоритмов идентифицировать сатирические, пародийные материалы: ИИ-система интерпретирует их в буквальном ключе, игнорируя коммуникативное намерение автора. В результате комедийный скетч, целенаправленно высмеивающий экстремистские взгляды для их дискредитации, может быть классифицирован как прямая пропаганда радикальных идей. Аналогичным образом мем с нарочито утрированным визуальным образом или провокационным

текстом, созданный исключительно ради юмористического эффекта, нередко получает маркировку «дезинформация», так как алгоритм фиксирует отдельные маркеры, но не учитывает жанровую специфику [Блумер 2017].

Некорректная интерпретация культурного кода возникает из-за вариативности семантики жестов, символов и речевых формул в разных социокультурных контекстах. ИИ-система, обученная преимущественно на данных, репрезентирующих нормы одной культурной среды, неизбежно допускает ошибки при анализе контента из иных регионов: безобидные локальные традиции или устоявшиеся формы коммуникации квалифицируются как нарушение правил платформы.

Ошибки в распознавании возрастных ограничений связаны с трудностями комплексной оценки контента для определения корректного возрастного рейтинга. Алгоритмы ориентируются на формальные признаки, пренебрегая информационной ценностью материала. К примеру, дисбаланс в алгоритмической модерации цифрового контента проявляется в избыточной фильтрации, образовательные ресурсы с научно-визуальной составляющей (биологические видео с анатомическими схемами) ограничиваются для аудитории младше 18 лет, а развлекательный контент с латентными референциями к взрослым темам остается вне зоны действия ограничительных механизмов и может быть рекомендован несовершеннолетним. Это создает парадоксальную ситуацию, где доступ к научному знанию ограничивается строже, чем к потенциально проблематичному развлекательному контенту.

При высокой нагрузке в многоуровневых системах модерации возникают каскадные сбои: первичная экспресс-проверка с упрощенными критериями фактически становится окончательным вердиктом, а последующие этапы с привлечением модераторов задействуются выборочно. В результате ошибки начального этапа закрепляются, снижая общую надежность системы.

Проведенный анализ на примере типичных ошибок ИИ-модерации наглядно демонстрирует, что современные алгоритмы оценки видеороликов пока не способны в полной мере воспроизвести логику и глубину человеческой интерпретации контента. Основная трудность кроется в фундаментальном несоответствии между технической фиксацией отдельных элементов (слов, объектов, визуальных паттернов) и способностью постигать объемный контекст, в котором эти элементы обретают смысл.

Соответственно, эффективная методология обучения ИИ-агентов оцениванию видеороликов должна строиться на симбиозе передовых технологических инструментов и глубокого понимания социальных процессов.

Преодоление ограничений ИИ-оценки: подходы международных платформ

В научных коллективах и на крупных международных площадках активно ведутся поиски путей повышения эффективности модерации контента. Среди апробированных решений особое место занимает многоуровневая схема проверки, предполагающая обязательное участие человека на ключевых этапах. Наглядным примером служит практика YouTube, где как раз используются гибридные системы. На первом этапе алгоритмы отсеивают явно недопустимый контент, а на втором уже спорные случаи отдают экспертам. Подобный подход помогает реже ошибаться, особенно когда нужно разобраться в чем-то тонком: понять культурный подтекст или отличить шутку от провокации [Липпман 2004].

Второй подход связан с тем, чтобы адаптировать системы модерации под культурные особенности. Для этого компании создают отдельные настройки для разных регионов и обучают соответствующие модули на данных, собранных локально. Например, TikTok внедряет локализованные фильтры, которые учитывают особенности местной речи и устоявшиеся речевые обороты.

Еще одно перспективное направление — это кросс-модальный анализ. Специалисты из Google работают над системами, которые могут одновременно изучать текст, аудиодорожку, видеоряд и сопутствующие данные. Такой подход заметно улучшает распознавание иронии: например, система сопоставляет, как звучит голос диктора, с тем, что происходит на экране, или замечает преувеличение, когда слова явно не совпадают с изображением и графикой.

Не менее важна и система обратной связи от пользователей: платформы внедряют механизмы быстрого оспаривания ИИ-решений, позволяя пользователям подавать апелляции, после чего модераторы проверяют спорный контент, тем самым алгоритмы дообучаются, уточняя границы допустимого с учетом жанровых и культурных особенностей [Krishnan 2025].

Методологические принципы обучения ИИ-агентов

В рамках исследования формируется методология, опирающаяся на четыре взаимосвязанных принципа учета социальных факторов при обучении ИИ-агентов[2] для оценки видеоконтента.

Согласно первому принципу — принципу социальной репрезентативности данных, обучающая выборка должна адекватно отражать социальную структуру аудитории видеоплатформ; этого можно достичь через стратификацию обучающей выборки по ключевым социально-демографическим признакам (возраст, пол, уровень образования, профессия, регион), учет культурного и регионального разнообразия, включение в обучающий набор видеоконтента и оценочных данных из разных культурных контекстов и обеспечение пропорционального представительства различных социальных групп в данных для предотвращения смещения в сторону доминирующих групп за счет квотирования долей разных категорий респондентов [Santavirta et al. 2025].

Второй принцип, моделирование социальных процессов[3], направлен на воспроизведение в ИИ-системах ключевых механизмов формирования коллективных оценок [Сафонова и др. 2023]. Его реализация предполагает моделирование группового обсуждения и консенсуса с помощью ансамбля ИИ-агентов, каждый из которых отражает позицию отдельной социальной группы, чтобы итоговая оценка выводилась на основе согласования их решений. Важным элементом здесь является учет социального влияния через весовые коэффициенты: мнения экспертов, блогеров и инфлюенсеров получают повышенный вес, что воспроизводит реальную асимметрию авторитета в обществе. Важно обеспечить также адаптивность системы при помощи непрерывного обучения, регулярного обновления данных и донастройки модели, это все позволит ей отслеживать динамику социальных норм и актуальных тем [Московичи 1995].

Третьим принципом является учет социальной вариативности, он предполагает отказ от унифицированных критериев в пользу дифференцированной настройки ИИ-систем с учетом социокультурных различий. Модели обучают на репрезентативных данных и при необходимости дополняют специализированными подмоделями для отдельных групп. Оценочные критерии локализуют с учетом культурных и социальных контекстов, а для отражения коллективного суждения используют механизмы голосования и агрегации оценок [Argyle et al. 2023]. Такой подход позволяет ИИ-системам учитывать специфику разных сообществ и избегать унифицированных оценочных критериев.

Четвертый принцип заключается в объяснимости и прозрачности решений ИИ, чтобы их понимали и специалисты, и обычные пользователи. Для этого в моделях выделяют ключевые

[2] ИИ-агент (AI agent) — автономная программа с элементами ИИ, способная воспринимать данные, принимать решения и обучаться. В исследовании используется для оценки видеоконтента с учетом социальных факторов.
[3] В конце 1990-х – начале 2000-х был разработан и читался на химическом факультете МГУ имени М.В. Ломоносова авторский спецкурс «Социальные технологии и моделирование социальных процессов» (Григорьева Н.С., Моргунов Е.Б., Чубарова Т.В. Учебные программы по специализации «Социальный менеджмент». М.: Изд-во «Дело» 1998. С. 27–34).

факторы, влияющие на итог, а результаты сопровождают простыми пояснениями на естественном языке. Дополнительно внедряют механизмы обратной связи, где через специальные интерфейсы пользователи могут оспорить оценку и помочь исправить ошибки [Qu et al. 2025]. В итоге люди больше доверяют системе, а неточности в ее работе удается устранять быстрее.

Предложенные принципы позволяют выстроить подход к созданию ИИ-систем для оценки видео, которые действительно учитывают потребности общества. В отличие от простых алгоритмов, проверяющих лишь формальные признаки, такие системы понимают, как люди воспринимают контент в реальной жизни с учетом социальных норм, культурных кодов и особенностей интерпретации.

Методы сбора и анализа данных для обучения ИИ

Реализация методологии требует применения комплексного подхода к сбору и анализу данных с обязательным учетом социальной природы оценки видеоконтента. Использование социологических методов сбора данных дает возможность выявить глубинные механизмы формирования оценочных суждений, а также определить критерии, которыми руководствуются различные социальные группы при оценке видеоматериалов.

На первом этапе целесообразно применять качественные методы исследования. Так, особенно полезными для выявления критериев оценки видео различными социальными группами будет метод фокус-группы. В ходе дискуссий фиксируются не только итоговые оценки, но и аргументация, динамика мнений, противоречия между участниками. Например, при обсуждении юмористического ролика молодежная группа может акцентировать внимание на оригинальности шуток, тогда как старшее поколение — на этичности содержания. Дополнением к фокус-группам служат глубинные интервью с представителями различных демографических и культурных групп, позволяющие понять глубинные мотивы оценок, скрытые установки и ценности.

Для масштабирования качественных данных и получения репрезентативной картины применяется социологический опрос с многомерной шкалой оценки, что позволяет зафиксировать не только реакцию, но и мотивационные основания оценки, характеристики эмоциональной реакции, а также степень соответствия контента социальным и культурным нормам [Silverstone 2007]. Таким образом, респонденты оценивают каждое видео по ряду параметров: информативность, креативность, этичность, соответствие культурным традициям целевой аудитории и т. п.

Следующий этап анализа связан с изучением коллективного измерения оценки контента. Важную роль здесь играет сетевой анализ формирования коллективных оценок в онлайн-сообществах: так изучаются структуры влияния, выявляются лидеры мнений, кластеры пользователей с похожими оценками, механизмы распространения оценок. Построение графов взаимодействия позволяет смоделировать процессы группового принятия решений для имитации в ансамблях ИИ-агентов.

В итоге, чтобы реализовать персонализированный подход на практике, применяется кластерный анализ для выделения типологических групп зрителей со схожими паттернами оценки. На основе многомерных данных (оценки видео, социально-демографические характеристики, ценностные установки) формируются профили целевых аудиторий, и в результате эти профили используются для калибровки подмоделей ИИ под конкретные социальные группы.

Таким образом, последовательное применение качественных методов (фокус-групп и глубинных интервью), опроса, сетевого и кластерного анализа создает целостную систему сбора и обработки данных, которая отражает многоуровневую природу восприятия видеоконтента

от индивидуальных мотивов до коллективных процессов, и служит надежной основой для обучения социально ориентированных ИИ-агентов [Lupton 2017].

Такой комплексный подход обеспечивает не просто сбор данных, а их осмысленную трансформацию в формат, пригодный для обучения ИИ с учетом социальной природы восприятия видеоконтента. Получаемые модели способны имитировать не абстрактного среднего зрителя, а дифференцированные оценки реальных социальных групп, что повышает релевантность и доверие к решениям ИИ-агентов.

Валидация и оценка эффективности методологии

Для подтверждения работоспособности разработанной методологии необходима комплексная валидация, учитывающая как технические, так и социальные аспекты оценки видеоконтента. Критерии социальной валидности позволяют оценить, насколько решения ИИ соответствуют реальным моделям восприятия пользователей. Прежде всего, это степень соответствия оценок ИИ распределению оценок репрезентативной выборки пользователей, которая измеряется через статистическое сравнение распределений, а также уровень согласия между решениями ИИ и групповым консенсусом (доля совпадений в ключевых категориях оценок), свидетельствующий о способности ИИ учитывать групповые нормы восприятия. Еще один значимый критерий — это способность воспроизводить типичные когнитивные искажения, характерные для социальных групп; к таким искажениям относятся эффект первого впечатления или эффект ореола (когда один яркий элемент влияет на восприятие всего ролика). Дополнительно можно ввести интегральные показатели: индекс социальной релевантности, учитывающий соответствие оценок ИИ ключевым социальным нормам и ценностям целевых групп (рассчитывается на основе опросов пользователей), и метрику культурной адаптивности, то есть способность модели корректно оценивать контент в разных культурных контекстах без дополнительной настройки.

Методы эмпирической проверки обеспечивают оценку соответствия ИИ-агентов социальным моделям восприятия. В их число входит сравнительный анализ оценок ИИ и репрезентативных социальных групп на идентичных видеоматериалах; соответственно, ИИ-агенты и группы пользователей оценивают один и тот же набор видеороликов, после чего результаты сравниваются с использованием статистических тестов.

Еще одним методом выступает проведение фокус-групп с оценкой решений ИИ, когда участникам предлагаются пары оценок (одна от человека, другая от ИИ-агента) без указания источника, а их задача определить, какое решение принадлежит человеку, при этом доля правильных ответов ниже 50% свидетельствует о высокой человекоподобности решений.

Дополнительно можно применить метод кейс-стади для реальных ситуаций оценки видеоконтента с участием пользователей: в рамках таких исследований моделируются сценарии модерации, где пользователи взаимодействуют с системой на основе ИИ, а исследователи фиксируют поведенческие реакции, например готовность принять решение ИИ без апелляции и субъективные оценки «понятность решений».

Метрики социальной валидности позволяют формализовать качественные критерии оценки в виде количественных показателей и служат инструментом объективной интерпретации результатов эмпирических методов. В их число входят коэффициент согласия между оценками ИИ и репрезентативными социальными группами, который измеряет степень согласованности решений с учетом вероятности случайного совпадения. Так, значение выше 0,8 интерпретируется как высокий уровень согласия, что подтверждает адекватность работы ИИ с точки зрения социальных ожиданий: доля случаев, в которых объяснения решений ИИ воспринимаются как человеческие

(по результатам фокус-групп), — это процент участников, которые не смогли отличить решения ИИ от человеческих; процент снижения апелляций по сравнению с традиционными системами, отражающий рост доверия к ИИ-решениям (чем реже пользователи оспаривают выводы системы, тем выше их удовлетворенность и готовность полагаться на ответы ИИ); индекс культурной адаптивности (от 0 до 1), рассчитываемый как средневзвешенное соответствие оценок ИИ нормам различных культурных групп. Значение выше 0,8 считается хорошим результатом и демонстрирует универсальность модели, ее способность корректно работать в разных социокультурных контекстах [Spearman 1904; Cohen 1960; Fleiss 1971].

Таким образом, сочетание эмпирических методов сбора данных и формализованных метрик социальной валидности создает комплексную систему оценки эффективности ИИ-агентов, которая позволяет не только фиксировать текущее соответствие решений ИИ социальным моделям восприятия, но и выявить направления для дальнейшего совершенствования системы.

Заключение

Проведенное исследование позволило преодолеть ключевой разрыв между алгоритмической оценкой видеоконтента и реальным восприятием такого контента различными социальными группами. Предложенная методология обучения ИИ-агентов опирается на социологические теории и учитывает многообразие культурных кодов, социальных норм, возрастных и ценностных различий аудитории. В ее основе лежат принципы социальной репрезентативности данных, моделирования социальных процессов, учета вариативности восприятия и прозрачности решений ИИ. Благодаря комплексному подходу к сбору и анализу данных, сочетанию качественных (фокус-групп, глубинные интервью) и количественных (опросы, сетевой и кластерный анализ) методов, удалось создать систему, способную имитировать не абстрактного среднего зрителя, а дифференцированные оценки реальных социальных групп. Предложенная методология предусматривает систему целевых показателей для будущей валидации ее эффективности.

Разработанную методологию можно масштабировать на другие типы медиаконтента — тексты, фото, аудио, а также адаптировать к динамике социальных норм. В перспективе это позволит создавать ИИ-решения, которые не просто анализируют формальные параметры информации, но и учитывают сложную природу социального восприятия, обеспечивая более гармоничное взаимодействие человека и технологий в цифровом пространстве.

Таким образом, исследование не только решает актуальную научную проблему расхождения между алгоритмическими и социальными оценками контента, но и закладывает основы для развития нового поколения социально ориентированных, культурно адаптивных и объяснимых ИИ-систем.

Список литературы:

Бандура А. Теория социального научения. СПб.: Евразия, 2000.

Бергер П., Лукман Т. Социальное конструирование реальности: трактат по социологии знания. М.: Медиум, 1995.

Блумер Г. Символический интеракционизм: перспектива и метод. М.: Элементарные формы, 2017.

Бурдье П. Различение: социальная критика суждения // Экономическая социология. 2005. Т. 6. № 3. С. 25–48.

Липпман У. Общественное мнение. М.: Институт Фонда «Общественное мнение», 2004.

Московичи С. Социальные представления: исторический взгляд // Психологический журнал. 1995. Т. 16. № 1. С. 3–18.

Новые подходы к оцениванию: искусственный интеллект как драйвер изменений в образовании / под науч. ред. Е.Ю. Кардановой. М.: НИУ ВШЭ, 2025.

Пузанова Ж.В., Кожоридзе Г.Г., Кожоридзе Д.Г. ИИ и социология: анализ технологических возможностей виртуальных респондентов // Социология: методология, методы, математическое моделирование. 2025. № 60. С. 216–246. DOI: 10.19181/4m.2025.34.1.6

Сафонова Ю.А., Субочева О.Н., Коршкова А.С. Агентность искусственных автономных систем как фактор трансформации социума // Социология. 2023. № 6. С. 116–122.

Argyle L.P., Busby E.C., Fulda N., Gubler J.R., Rytting C., Wingate D. Out of One, Many: Using Language Models to Simulate Human Samples // Political Analysis. 2023. Vol. 31. Is. 3. P. 337–351. DOI: 10.1017/pan.2023.2

Cohen J. A Coefficient of Agreement for Nominal Scales // Educational and Psychological Measurement. 1960. Vol. 20. Is. 1. P. 37–46. DOI: 10.1177/001316446002000104

Fleiss J.L. Measuring Nominal Scale Agreement among Many Raters // Psychological Bulletin. 1971. Vol. 76. Is. 5. P. 378–382. DOI: 10.1037/h0031619

Katz E. Communications Research since Lazarsfeld // Public Opinion Quarterly. 1987. Vol. 51. Special Issue. P. S25–S45.

Krishnan N. AI Agents: Evolution, Architecture, and Real-World Applications // arXiv Preprint. 2025. DOI: 10.48550/arXiv.2503.12687

Lazarsfeld P., Berelson B., Gaudet H. The People's Choice: How the Voter Makes Up His Mind in a Presidential Campaign. Princeton: Princeton University Press, 1944.

Lupton D. Digital Sociology. London: Routledge, 2017.

Mitchell W.J.T. Picture Theory: Essays on Verbal and Visual Representation. Chicago: University of Chicago Press, 1994.

Qu X., Damoah A., Sherwood J., Liu P., Jin Ch., Chen L., Shen M., Aleisa N., Hou Z., Zhang Ch., Gao L., Li Y., Yang Qu., Wang Qu., De Souza Ch. A Comprehensive Review of AI Agents: Transforming Possibilities in Technology and Beyond // arXiv Preprint. 2025. DOI: 10.48550/arXiv.2508.11957

Santavirta S., Wu Y., Suominen L., Nummenmaa L. GPT-4V Shows Human-Like Social Perceptual Capabilities at Phenomenological and Neural Levels // Imaging Neuroscience. 2025. Vol. 3. DOI: 10.1162/IMAG.a.134

Silverstone R. Media and Morality: The Rise of Mediated Public Conscience. Cambridge: Polity Press, 2007.

Spearman C. The Proof and Measurement of Association between Two Things // The American Journal of Psychology. 1904. Vol. 15. Is. 1. P. 72–101.

References:

Argyle L.P., Busby E.C., Fulda N., Gubler J.R., Rytting C., Wingate D. (2023) Out of One, Many: Using Language Models to Simulate Human Samples. *Political Analysis*. Vol. 31. Is. 3. P. 337–351. DOI: 10.1017/pan.2023.2

Bandura A. (2000) *Social Learning Theory*. St. Petersburg: Yevraziya.

Berger P., Luckmann T. (1995) *The Social Construction of Reality: A Treatise in the Sociology of Knowledge*. Moscow: Medium.

Blumer G. (2017) *Symbolic Interactionism: Perspective and Method*. Moscow: Elementarnye formy.

Bourdieu P. (2005) La Distinction: Critique sociale du jugement. *Ekonomicheskaya sotsiologiya*. Vol. 6. No. 3. P. 25–48.

Cohen J.A (1960) Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*. 1960. Vol. 20. Is. 1. P. 37–46. DOI: 10.1177/001316446002000104

Fleiss J.L. (1971) Measuring Nominal Scale Agreement among Many Raters. *Psychological Bulletin*. Vol. 76. Is. 5. P. 378–382. DOI: 10.1037/h0031619

Kardanova Ye.Yu. (ed.) (2025) *Novyye podkhody k otsenivaniyu: iskusstvennyy intellekt kak drayver izmeneniy v obrazovanii* [New approaches to assessment: Artificial Intelligence as a driver of change in education]. Moscow: NIU VSHE.

Katz E. (1987) Communications Research since Lazarsfeld. *Public Opinion Quarterly*. Vol. 51. Special Issue. P. S25–S45.

Krishnan N. (2025) AI Agents: Evolution, Architecture, and Real-World Applications. *arXiv Preprint*. DOI: 10.48550/arXiv.2503.12687

Lazarsfeld P., Berelson B., Gaudet H. (1944) *The People's Choice: How the Voter Makes Up His Mind in a Presidential Campaign*. Princeton: Princeton University Press.

Lippmann W. (2004) *Public Opinion*. Moscow: Institut Fonda "Obshchestvennoye mneniye".

Lupton D. (2017) *Digital Sociology*. London: Routledge.

Mitchell W.J.T. (1994) *Picture Theory: Essays on Verbal and Visual Representation*. Chicago: University of Chicago Press.

Moscovici S. (1995) Social Representations: A Historical Perspective. *Psikhologicheskiy zhurnal*. Vol. 16. No. 1. P. 3–18.

Puzanova Zh.V., Kozhoridze G.G., Kozhoridze D.G. (2025) Generative AI and Sociology: Analyzing Virtual Respondent Technology. *Sotsiologiya: metodologiya, metody, matematicheskoye modelirovaniye*. No. 60. P. 216–246. DOI: 10.19181/4m.2025.34.1.6

Qu X., Damoah A., Sherwood J., Liu P., Jin Ch., Chen L., Shen M., Aleisa N., Hou Z., Zhang Ch., Gao L., Li Y., Yang Qu., Wang Qu., De Souza Ch. (2025) A Comprehensive Review of AI Agents: Transforming Possibilities in Technology and Beyond. *arXiv Preprint*. DOI: 10.48550/arXiv.2508.11957

Safonova Yu.A., Subocheva O.N., Korshkova A. S. (2023) Agency of Artificial Autonomous Systems as a Factor in the Transformation of Society. *Sotsiologiya*. No. 6. P. 116–122.

Santavirta S., Wu Y., Suominen L., Nummenmaa L. (2025) GPT-4V Shows Human-like Social Perceptual Capabilities at Phenomenological and Neural Levels. *Imaging Neuroscience*. Vol. 3. DOI: 10.1162/IMAG.a.134

Silverstone R. (2007) *Media and Morality: The Rise of Mediated Public Conscience*. Cambridge: Polity Press.

Spearman C. (1904) The Proof and Measurement of Association between Two Things. *The American Journal of Psychology*. Vol. 15. Is. 1. P. 72–101.