

## ПРОБЛЕМА РАЦИОНАЛИЗАЦИИ И ЧРЕЗМЕРНОГО ПОЛАГАНИЯ НА ИНСТРУМЕНТЫ ХАИ: АНАЛИЗ ОБЪЯСНЕНИЙ БОЛЬШИХ ЯЗЫКОВЫХ МОДЕЛЕЙ

А.В. Суворова (*asuvorova@hse.ru*)

Национальный исследовательский университет  
"Высшая школа экономики", Санкт-Петербург

В работе исследуется проблема чрезмерного полагания (overreliance) пользователей на результаты интерпретации моделей машинного обучения, а также способов ее решения с помощью пояснений, генерируемых большими языковыми моделями (LLM). Результаты эксперимента показали, что большинство моделей, так же как и пользователи-люди в исходном эксперименте, игнорировали аномалии или предлагали правдоподобные, но ложные объяснения, рационализируя выводы. Это указывает на риски некритичного использования LLM для интерпретации моделей машинного обучения без дополнительных механизмов валидации.

**Ключевые слова:** объяснимый ИИ, машинное обучение, оценивание моделей пользователями.

### Введение

Все более широкое применение интеллектуальных технологий, разработка и внедрение систем машинного обучения в различные сферы жизни, включая социальную сферу, приводят к необходимости объяснения решений, принимаемых с помощью таких систем. Вопросы интерпретируемости результатов моделей, построенных с помощью алгоритмов машинного обучения, выявления факторов, оказывающих влияние на решение, все чаще возникают со стороны как общества, так и исследователей, включая вопросы по поводу их корректности, отсутствия дискриминации и т.д. Как следствие, область исследований, связанная с различными аспектами интерпретируемости, объяснимости моделей (explainable AI, interpretable machine learning), очень быстро развивается, в частности, предлагаются различные алгоритмы для объяснения уже построенных моделей.

При этом в ряде статей [Bansal et al., 2021], [Ehsan et al., 2024] поднимается проблема чрезмерного полагания (overreliance) на результаты, предоставляемые различными моделями. Чаще всего авторы отмечают, что даже добавление пояснений в информационные системы не позволяет убрать этот эффект [Bansal et al., 2021], [Buçinca et al., 2021]. Однако в [Vasconcelos et al., 2023] авторы исследуют проблему подробнее и показывают, что эффективность пояснений зависит от многих факторов, включая сложность задачи (слепая уверенность в технологии усиливается для более сложных задач), формат представления пояснений (обычное текстовое пояснение не оказывает влияния на чрезмерное полагание, а сокращенное с выделением важных элементов – снижает его). Одновременно с этим, во многих работах предлагается использовать LLM для интерпретации моделей машинного обучения: как часть объясняющего инструмента (например, для преобразования естественно-языкового запроса к модели [Slack et al., 2023]) или для генерации пояснений к результатам объясняющих моделей, например, SHAP-графикам [Hsu et al., 2024], [Singh et al., 2024].

Важно отметить, что в указанных работах не обсуждается вопрос, насколько такие дополнительные пояснения от LLM свободны от ограничений, выявленных в предыдущих исследованиях, посвященных практикам использования объясняющих инструментов пользователями. Например, в статье [Kaur et al., 2020] авторы показали, что аналитики склонны слишком полагаться на результаты применения подобных инструментов интерпретации, даже не понимая принципы их работы, особенно, если результаты представлены в «научном» формате и подкреплены ссылками на публикации.

В докладе представлены результаты эксперимента, воспроизводящего задачу из [Kaur et al., 2020], но при условии, что «участником» эксперимента является LLM. Другими словами, исследуется, будут ли дополнительные пояснения от LLM, предлагаемые разными авторами для упрощения работы с результатами инструментов XAI, воспроизводить некорректные выводы пользователей или преодолевать их.

## **1. Особенности пользовательской оценки объяснений моделей машинного обучения**

Методы объяснимого искусственного интеллекта (XAI) позволяют получить важную информацию о поведении модели, но из-за множества доступных инструментов интерпретации результатов конкретное решение может не соответствовать оптимальным требованиям целевых пользователей. Исследования показывают, что цели объяснимости и даже значение слова «быть объяснимым» различаются в зависимости от предметной области, опыта и роли конечного пользователя в рабочем процессе [Arrieta et al., 2020].

### **1.1. Типы пользователей в решениях, направленных на интерпретацию моделей**

Авторы [Hong et al., 2020] выделяют три ключевые группы участников, вовлечённых в создание и применение моделей:

- создатели моделей (model builders) – специалисты по анализу данных и машинному обучению, которые разрабатывают модели;
- тестировщики и критики моделей (model breakers) – эксперты в предметной области, менеджеры продуктов, юристы, которые оценивают и проверяют модели;
- пользователи моделей (model consumers) – конечные пользователи, которые взаимодействуют с результатами работы моделей.

Большинство методов и инструментов объяснения моделей машинного обучения ориентированы либо на создателей (для отладки и улучшения моделей), либо на конечных пользователей (для прозрачности и интерпретируемости). В исследованиях [Hong et al., 2020], [Langer et al., 2021]; [Tomsett et al., 2018] показано, что конечным пользователям, не экспертам в предметной области или ML-специалистам, часто не нужны объяснения моделей, основными «потребителями» объяснений в моделях являются внутренние пользователи, включенные в процесс обсуждения, разработки, внедрения модели. Более того, у ML-специалистов тоже мало стимулов использовать модели объяснения [Langer et al., 2021], [Zakharova et al., 2021], т.к. их внедрение требует дополнительного времени и ресурсов, а оценивание адекватности модели чаще всего проводится самим разработчиком на основе формальных метрик качества, а внешнее оценивание базируется на доверии к самому разработчику и метриках ее эффективности (технических или бизнес-метриках), а не на результатах исследования модели. Таким образом, для более объективного оценивания нужен инструмент, который использовал бы существующие алгоритмы объяснения (и был соответственно разработан ML-специалистом), но при этом упрощал исследование модели для не-ML-специалиста. При этом потребности тестировщиков часто остаются без должного внимания. Основная сложность в разработке систем для этой группы заключается в том, что они, как правило, не обладают глубокими техническими знаниями о методах машинного обучения и не могут в полной мере использовать возможности интерпретируемого машинного обучения (IML).

### **1.2. Оценивание моделей интерпретации**

Для выводов об эффективности того или иного решения, направленного на интерпретацию результатов моделей машинного обучения, независимо от того, какой тип пользователей является для этого решения целевым, необходимо иметь возможность оценить качество полученных объяснений.

В широко цитируемой классификации способов оценивания объяснений [Doshi-Velez et al., 2017] выделены три категории подходов к оценке:

- оценка итоговых инструментов пользователями. Этот вид оценки основан на экспериментах, в которых конечные пользователи решают реальные задачи, используя предложенные инструменты для принятия решений.
- упрощенная оценка на людях. Этот тип оценки также требует экспериментов с людьми, но, поскольку конечные пользователи (специалисты в предметной области, например, врачи) обычно труднодоступны, а их время дорого, то используется упрощенное приближение к реальной ситуации с непрофессионалами.
- оценка функциональности. Этот тип оценки не требует экспериментов с вовлечением людей и использует метрики, основанные на формальном определении интерпретируемости, которые можно оценить математически или с помощью моделирования [Agarwal et al., 2022].

Учитывая, что единого общепринятого определения объяснимости не существует, довольно часто она описывается довольно размытыми формулировками как «способность объяснить или представить в понятных терминах человеку» [Doshi-Velez et al., 2017]. При этом так как цель применения методов интерпретации результатов моделей в большинстве случаев связана с необходимостью специалисту принять решение, можно ли использовать анализируемую модель машинного обучения, нет ли в ней смещений и искажений, то оценивание именно со стороны конечных пользователей является существенным этапом проектирования системы. Как следствие, значительная часть исследований по оцениванию объяснений ориентирована на пользовательский опыт. В этих исследованиях описываются конкретные эксперименты, измеряющие удовлетворенность [Hoffman et al., 2018], доверие [Drozdal et al., 2020] или способность принимать решения на основе пояснений результатов модели [Kaur et al., 2020].

## **2. Искажения пользовательской оценки объяснений моделей**

Несмотря на широкое распространение именно пользовательских оценок объяснений, корректность применения инструментов интерпретации конечными пользователями и способы предотвращения ошибочного использования обсуждаются не часто [Kaur et al., 2024]. Одна из таких работ – статья [Kaur et al., 2020], в которой изучались практики использования методов интерпретируемого машинного обучения аналитиками данных для изучения ML-моделей, в частности, для поиска ошибок в модели и данных. Именно эта работа взята за основу для исследования, будут ли дополнительные пояснения от LLM воспроизводить некорректные выводы пользователей или преодолевать их.

## 2.1. Описание исходного исследования

В рамках этого доклада рассмотрена только одна из выявленных в [Kaur et al., 2020] проблем – чрезмерное полагание на результаты объяснения моделей и рационализация необычных паттернов, поэтому в последующем описании приведены только те элементы процедуры исследования, которые относятся к выявлению указанной проблемы.

Для определения того, смогут ли пользователи найти несоответствия в предложенной им модели, в набор данных Adult Income dataset были внесены различные искажения, включая замену для 10% наблюдений с высоким доходом реальных значений возраста на среднее значение по выборке – 38 лет. Затем на модифицированных данных была построена модель классификации с использованием алгоритма LightGBM для предсказания высокого или низкого дохода. К итоговой модели были применены инструменты интерпретации ML-моделей, включая SHAP, в частности, был построен график зависимости SHAP-значений от возраста, показывающий для каждого наблюдения (точка на графике) влияние возраста на предсказание дохода – чем дальше от 0, тем более влиятелен возраст для предсказание. Итоговый график показан на рис. 1. Стоит отметить, что цвет точки соответствует значению семейного статуса для наблюдения, но это не существенно для рассматриваемой задачи.

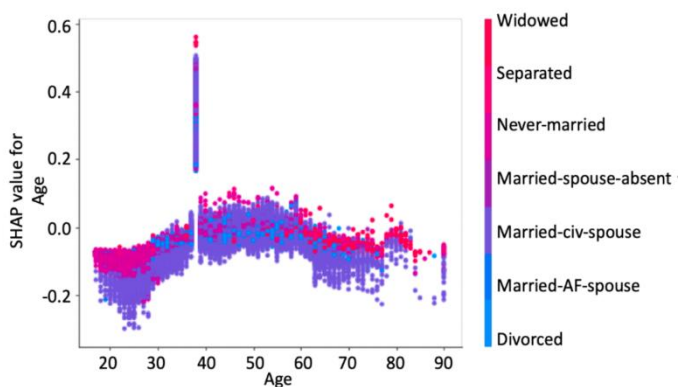


Рис. 1. SHAP-значения модели предсказания дохода относительно возраста

Участникам предоставлялось описание контекста – какие данные использованы, какая модель была построена, какое у нее качество, а также описание логики метода SHAP и полученные графики. Затем задавались вопросы на понимание модели, в том числе вопрос про возраст – «Как возраст влияет на доход (How does the feature Age affect output Income)?». Как отмечалось выше, часть участников в ответ на этот вопрос рациона-

лизировала выброс на значении 38 лет, объясняя его какими-то закономерностями в реальной жизни, а не ошибкой модели («Возраст 38 лет, похоже, имеет наибольшее положительное влияние на доход, исходя из графика. Не уверен, почему, но объяснение ясно показывает это... имеет смысл» [Kaur et al., 2020]).

## **2.2. Генерация пояснений к графикам с помощью LLM**

Многие авторы [Omar et al., 2025] отмечают случаи некорректных выводов конечных пользователей по предоставляемым им SHAP-графикам, часто связывая это со сложностью подобных графиков. Как одно из возможных решений предлагается [Hsu et al., 2024] использовать инструменты на основе LLM для генерации дополнительных пояснений, что потенциально может предотвратить некорректные выводы. Для проверки этой идеи были сгенерированы пояснения к графикам из исходного эксперимента с конечными пользователями (см. раздел 2.1).

Для проведения эксперимента были использованы четыре модели – GigaChat, GPT 3.5, Sonar и DeepSeek. Запрос в каждую из них включал SHAP-график (рис. 1), контекст исходного эксперимента (описание данных и построенной модели) на английском языке, т.к. формулировки были напрямую скопированы из протокола исследования [Kaur et al., 2020], и целевой вопрос «Как возраст влияет на доход (How does the feature Age affect output Income)?». После ответа был задан уточняющий вопрос о возможном объяснении пика в возрасте 38 лет на графике. Запрос к каждой модели был повторен 10 раз. Также была проведена вторая итерация эксперимента, практически полностью повторяющая первую, за одним исключением – контекст задачи был расширен, передан в виде файла из исходного исследования, включающего не только описание данных и модели, но и объяснение принципов работы SHAP.

Результаты трех из четырех рассмотренных моделей (GigaChat, GPT 3.5, DeepSeek) оказались очень схожими – на первую часть задания был выдан очень обобщенный ответ примерно следующего содержания: «возраст оказывает нелинейное влияние на прогнозируемый доход. молодой возраст (20-40 лет): положительное влияние на доход; средний возраст (40-60 лет): нейтральное или слабо отрицательное влияние; пожилой возраст (60+ лет): отрицательное влияние на доход». Этот ответ часто дополнялся формальными описаниями метода: что отображается по оси x, чему соответствует ось y.

Ни одна из этих трех моделей не акцентировала внимание на пике в возрасте 38 лет при первоначальном запросе, в отличие от Sonar, где в дополнение к выводу примерно той же структуры про нелинейный эффект сразу выделялся выброс на значении 38 лет и приводилось возможное объяснение этому: «В этом конкретном возрасте (38 лет) признак "Возраст" оказывает сильное положительное влияние на прогнозирование

высокого дохода для многих людей. Это может быть артефакт данных или когортный эффект (например, случайная концентрация высокооплачиваемых людей этого возраста в выборке)».

Более того, при запросе уточнений в модели Sonar возможные искажения оставались основным объяснением (табл. 1).

Таблица 1

Возможная причина	Описание	Вероятность
Артефакт данных/ошибка ввода	Множество записей для возраста 38 лет из-за округления или ошибок ввода данных	высокая
Переобучение модели	Модель уловила специфическую закономерность, характерную только для 38-летних	средняя
Взаимодействие признаков	Сильное влияние других факторов, не отображённых на графике	средняя
Когортный эффект	Реальная экономическая или демографическая причина высоких доходов в возрасте 38 лет	низкая / средняя

Для остальных трех моделей дополнительный запрос о возможном объяснении пика в 38 лет приводил к рационализации видимой закономерности, повторяя поведение участников эксперимента [Kaur et al., 2020]. Среди возможных причин были указаны многие персональные, экономические и социальные факторы: стабильное карьерное положение, возможность иметь двойной семейный доход, накопленный профессиональный опыт, завершение всех уровней образования, спрос на рынке труда, особенности исторических данных. В некоторых результатах упоминалось возможное взаимодействие с другими переменными, но ни разу не появилась ошибка ввода.

Расширенное пояснение принципов SHAP в контексте на второй итерации не привело к улучшению выводов, обобщенные результаты были похожи.

**Заключение**

Проведенное исследование показало, что большие языковые модели (LLM) повторяют поведение пользователей, демонстрируя склонность к чрезмерному полаганию на результаты интерпретации моделей машинно-

го обучения. Как и в оригинальном эксперименте, в котором люди рационализировали аномалии в данных, три из четырех протестированных моделей либо игнорировали явный выброс на SHAP-графике, либо предлагали ему правдоподобные, но ошибочные объяснения.

Таким образом, несмотря на потенциал LLM в упрощении взаимодействия с ML-моделями, их прямое использование для объяснения решений требует осторожности. Чтобы избежать усиления эффекта слепого полагания на представленный инструментом результат, можно направить исследование на создание гибридных систем, сочетающих генерацию естественно-языковых пояснений с алгоритмами валидации и критического анализа.

### Список литературы

- [Agarwal et al., 2022] Agarwal C. et al. Openxai: Towards a transparent evaluation of model explanations // In: Advances in neural information processing systems. – 2022. – Vol. 35. – P. 15784-15799.
- [Arrieta et al., 2020] Arrieta A.B. et al. Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI // Information Fusion. – 2020. – Vol. 58. – P. 82-115.
- [Bansal et al., 2021] Bansal G. et al. Does the whole exceed its parts? The effect of AI explanations on complementary team performance // In: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems. – 2021. – P. 1-16.
- [Buçinca et al., 2021] Buçinca Z., Malaya M. B., Gajos K. Z. To trust or to think: cognitive forcing functions can reduce overreliance on AI // in AI-assisted decision-making. In: Proceedings of the ACM on Human-computer Interaction. – 2021. – Vol. 5 (CSCW1). – P. 1-21.
- [Drozdal et al., 2020] Drozdal J., Weisz J., Wang D., et al. Trust in automl: Exploring information needs for establishing trust in automated machine learning systems // In: Proceedings of the 25th International Conference on Intelligent User Interfaces. – 2020. – P. 297-307.
- [Ehsan et al., 2024] Ehsan U. et al. Human-centered explainable AI (HCXAI): Reloading explainability in the era of large language models (LLMs) // In: Extended Abstracts of the CHI Conference on Human Factors in Computing Systems. – 2024. – P. 1-6.
- [Hoffman et al., 2018] Hoffman R.R., Mueller S.T., Klein G., Litman J. Metrics for explainable AI: Challenges and prospects // In: arXiv preprint. – 2018. – URL: arXiv:1812.04608.
- [Hong et al., 2020] Hong S.R., Hullman J., Bertini E. Human factors in model interpretability: Industry practices, challenges, and needs // In: Proceedings of the ACM on Human-Computer Interaction. – 2020. – Vol. 4 (CSCW1). – P. 1-26.
- [Hsu et al., 2024] Hsu C.C., Wu I.Z., Liu S.M. Decoding AI Complexity: SHAP Textual Explanations via LLM for Improved Model Transparency // In: 2024 International Conference on Consumer Electronics-Taiwan (ICCE-Taiwan). – IEEE, 2024. – P. 197-198.
- [Kaur et al., 2020] Kaur H. et al. Interpreting interpretability: understanding data scientists' use of interpretability tools for machine learning // In: Proceedings of the 2020 CHI conference on human factors in computing systems, Honolulu, USA, 2020. – P. 1-14. – doi: 10.1145/3313831.3376219.



- [**Kaur et al., 2024**] Kaur H. et al. Interpretability Gone Bad: The Role of Bounded Rationality in How Practitioners Understand Machine Learning // In: Proceedings of the ACM on Human-Computer Interaction. – 2024. – Vol. 8 (CSCW1). – P. 1-34.
- [**Langer et al., 2021**] Langer M. et al. What do we want from Explainable Artificial Intelligence (XAI)?—A stakeholder perspective on XAI and a conceptual model guiding interdisciplinary XAI research. Artificial Intelligence. – 2021. – Vol. 296. – 103473.
- [**Omar et al., 2025**] Omar Z. A. et al. Beyond Accuracy, SHAP, and Anchors--On the difficulty of designing effective end-user explanations // arXiv preprint arXiv:2503.15512. – 2025.
- [**Singh et al., 2024**] Singh C. et al. Rethinking interpretability in the era of large language models // In: arXiv preprint arXiv:2402.01761. – 2024.
- [**Slack et al., 2023**] Slack D. et al. Explaining machine learning models with interactive natural language conversations using TalkToModel // In: Nature Machine Intelligence. – 2023. – Vol. 5(8). – P. 873-883.
- [**Tomsett et al., 2018**] Tomsett R., Braines D., Harborne D., Preece A., Chakraborty S. Interpretable to whom? A role-based model for analyzing interpretable machine learning systems // In: arXiv preprint arXiv:1806.07552. – 2018.
- [**Vasconcelos et al., 2023**] Vasconcelos H. et al. Explanations can reduce overreliance on ai systems during decision-making // In: Proceedings of the ACM on Human-Computer Interaction. – 2023. – Vol. 7 (CSCW1). – P. 1-38.
- [**Zakharova et al., 2021**] Zakharova V., Suvorova A. Social Aspects of Machine Learning Model Evaluation: Model Interpretation and Justification from ML-practitioners' Perspective // In: 24th International Conference "Internet and Modern Society". – 2021. – P. 230-234.