

ВЛИЯНИЕ ОБУЧАЮЩИХ ДАННЫХ НА ПРИНЯТИЕ РЕШЕНИЙ В СИСТЕМАХ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА В СФЕРЕ БЕЗОПАСНОСТИ ЖИЗНЕДЕЯТЕЛЬНОСТИ

Активное внедрение систем искусственного интеллекта (ИИ) в важные области, такие как финансы, здравоохранение, управление персоналом, транспорт и государственное администрирование, чрезвычайные ситуации, привлекло внимание к вопросу об этической нейтральности алгоритмов. Долгое время алгоритмические системы считались объективными инструментами для обработки данных, свободными от субъективных оценок. Однако современные исследования показывают, что алгоритмы машинного обучения не являются нейтральными по своим ценностям и в значительной степени отражают характеристики обучающих данных, включая социальные и поведенческие искажения.

По мнению ассистента кафедры искусственного интеллекта Факультета информационных технологий и анализа больших данных Финансовый университет при Правительстве Российской Федерации Чупревой Алёны Николаевны ключевым фактором, влияющим на характер принимаемых ИИ-решений, являются обучающие данные. Историческая предвзятость в данных отражает социальные привычки и ограничения прошлого. В результате алгоритмы, обученные на таких выборках, не только воспроизводят, но и усугубляют существующие асимметрии. Как отмечают эксперты в области этики ИИ, «алгоритм не находит справедливость; он оптимизирует статистическую закономерность, заданную данными и целевой функцией».

Важным методологическим аспектом анализа является понимание того, что процесс разработки и внедрения алгоритмических систем включает ряд решений, основанных на нормах. Эти решения принимаются на этапах постановки задачи, отбора данных, выбора модели и метрик оценки её эффективности. Каждый из этих этапов предполагает определенные допущения относительно допустимости ошибок, приоритетов оптимизации и интерпретации результатов. Эти аспекты нельзя рассматривать только с технической точки зрения. Важно понимать алгоритмическое решение как нечто большее, чем просто результат вычислительной процедуры. Оно является производным от набора социальных и институциональных выборов, которые фиксируются в формализованной модели.

Ряд инцидентов, связанных с использованием систем автоматизированного вождения, наглядно демонстрирует ограничения алгоритмического принятия решений при отсутствии контекстуальной и этической интерпретации ситуации. В качестве примера можно рассмотреть широко обсуждавшийся случай, при котором автомобиль Tesla в режиме ассистированного управления продолжительное время двигался на высокой скорости, несмотря на отсутствие активного контроля со стороны водителя. Данный эпизод, свидетельствует о том, что алгоритм, эффективно решая задачу навигации и реагирования на дорожную обстановку, оказался неспособен корректно интерпретировать поведенческий контекст – в частности, состояние водителя и уровень его фактической вовлечённости в процесс управления.

По мнению старшего преподавателя кафедры безопасности жизнедеятельности Финансового университета при Правительстве Российской Федерации Зельского Алексея Георгиевича подобное поведение системы не является сбоем в техническом смысле, а представляет собой следствие логики, заложенной в обучающих данных и сценариях использования. Алгоритм действует строго в рамках формализованных допущений, не располагая механизмами оценки потенциальных социальных и этических последствий принимаемых решений, если они изначально не были включены в модель.

Рассматриваемый кейс, иллюстрирует более широкую системную проблему: алгоритмические системы принимают решения в пределах заданных параметров оптимизации, не учитывая моральные и социальные эффекты своих действий, если соответствующие критерии не формализованы на этапе проектирования. При этом ответственность за итоговое поведение системы оказывается распределённой между разработчиками, пользователями и регуляторными институтами, что существенно усложняет процедуру оценки рисков и последствий внедрения подобных технологий.

Дополнительное усложнение ситуации связано с нерепрезентативностью обучающих данных и использованием прокси-признаков. Даже при исключении явных чувствительных параметров алгоритмы способны выявлять устойчивые косвенные корреляции, приводящие к нежелательным эффектам. В этой связи этическая проблематика алгоритмов, по её мнению, заключается не столько в наличии формально запрещённых признаков, сколько в доминировании логики оптимизации эффективности, не сопровождаемой нормативной оценкой допустимости получаемых решений.

Таким образом, вопрос этической нейтральности алгоритмов выходит далеко за пределы сугубо технических дискуссий и становится отражением более широкой общественной проблемы – перераспределения ответственности в условиях цифровизации. Алгоритмы всё чаще принимают решения, влияющие на безопасность, благополучие и возможности людей, при этом оставаясь «невидимыми участниками» социальных процессов. Проблема заключается не в ошибках как таковых, а в том, что автоматизированные системы действуют строго в рамках заданной логики, не обладая способностью к моральному суждению и оценке последствий своих действий. В этой связи становится очевидным, что вопрос нейтральности алгоритмов является, по сути, вопросом человеческого выбора: какие ценности закладываются в системы искусственного интеллекта, кто определяет границы их допустимого поведения и, кто несёт ответственность за последствия их применения. Именно от ответов на эти вопросы зависит, станет ли искусственный интеллект инструментом общественного развития или источником новых форм риска и неравенства.

*Чупрева Алёна Николаевна,
ассистент кафедры искусственного интеллекта
Факультета информационных технологий и анализа больших данных
Финансовый университет при Правительстве Российской Федерации*

*Зельский Алексей Георгиевич
старший преподаватель кафедры безопасности жизнедеятельности
Финансовый университет при Правительстве Российской Федерации*