УДК 351

Вандышев А.В.
магистрант ИГСУ РАНХиГС
Россия, г. Москва

АКТУАЛЬНЫЕ ВОПРОСЫ РЕАЛИЗАЦИИ СИСТЕМ СОЦИАЛЬНОГО РЕЙТИНГА НА ОСНОВЕ ТЕХНОЛОГИЙ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА В ГОСУДАРСТВЕННОМ УПРАВЛЕНИИ

Аннотация: *Статья посвящена вопросу реализации систем социального рейтинга на основе технологий искусственного интеллекта в государственном управлении. Рассмотрены важнейшие, по мнению автора, проблемы внедрения указанных систем. С этой точки зрения проанализирован проект платформы социального рейтинга «Мы», разрабатываемой в Российском государственном социальном университете. Выявлены некоторые проблемы реализации и внедрения платформы.*

Ключевые слова: *государственное управление, искусственный интеллект, социальный рейтинг, социальная справедливость, защита персональных данных, платформа «Мы».*

Annotation: *The article is devoted to the issue of implementing social rating systems based on artificial intelligence technologies in public administration. The most important, in the author's opinion, problems of implementing social rating systems based on artificial intelligence are considered. From this point of view, the project of the social rating platform "We", developed at the Russian State Social University, is analyzed. Some problems in the implementation and implementation of the platform have been identified.*

Key words: *public administration, artificial intelligence, social rating, social justice, personal data protection, "We" platform.*

Указ Президента РФ от 10 октября 2019 года № 490 "Об искусственном интеллекте в Российской Федерации" утверждает "Национальную стратегию развития искусственного интеллекта до 2030 года" [1], впервые определяя понятие искусственного интеллекта на уровне нормативно-правового акта, а также другие ключевые понятия в этой области. Согласно стратегии, искусственный интеллект - это комплекс технологий, позволяющий имитировать когнитивные функции человека (включая самообучение и решение задач без заранее заданного алгоритма) и достигая результатов, сравнимых с интеллектуальной деятельностью человека. Этот комплекс включает в себя информационно-коммуникационную инфраструктуру, программное обеспечение (в том числе методы машинного обучения), процессы и сервисы по обработке данных и поиску решений.

В настоящее время существуют технологические возможности для решения специализированных задач с применением методов машинного обучения. Особый интерес представляет создание искусственного интеллекта (ИИ), способного самостоятельно решать разнообразные задачи в условиях внешних воздействий, требующих адаптации. Решение данной проблемы находится на пересечении естественно-научных, технических и социально-гуманитарных областей научного знания [2].

Применение ИИ представляет собой инновационный подход к совершенствованию государственной и муниципальной системы управления в Российской Федерации. Однако, на данный момент проблема регулирования использования ИИ на всех этапах его жизненного цикла, включая вопросы информационной безопасности, разработку этических стандартов, уменьшение неопределенностей и рисков, остается решенной лишь частично.

Разделом IV Национальной стратегии установлено, что использование ИИ и его технологий носит общий или «сквозной» характер и в том числе обеспечивает «создание условий для повышения качества жизни людей, в том числе путем повышения уровня предоставления услуг в социальной сфере

(образование, здравоохранение), а также повышение качества предоставления государственных и муниципальных услуг» [1].

Одним из способов улучшения качества социальных услуг и повышения эффективности деятельности государственных и муниципальных служб, который появился в последнее время, является целевое оказание услуг на условиях и в порядке зависимости от уровня «социального рейтинга».

Социальный рейтинг - это система оценки граждан, основанная на анализе различных данных об их деятельности и поведении. Данные могут включать онлайн-поведение, кредитную историю, уплату налогов, участие в общественных мероприятиях и другие аспекты. Эта система оценки определяет общий балл, отражающий степень ответственности гражданина и уровень доверия государства к нему.

Одна из самых известных систем социального рейтинга предложена правительством Китая. Однако это не единственный пример. Множество других организаций, таких как различные страховые компании по всему миру, анонсировали или уже внедрили системы, которые позволяют их операторам эффективно собирать информацию о поведении людей и использовать ее при принятии важных решений. Например, страхователям жизни в Нью-Йорке разрешено принимать решения в отношении своих клиентов, используя информацию из социальных сетей. Индустрия автострахования и каршеринга во многом опирается на информационные системы, которые рассчитывают итоговые сборы на основе данных, полученных от устройства слежения, специально установленного в автомобиле.

Сервисы совместного использования, такие как Airbnb, службы такси или доставки, имеют своего рода систему оценки, которая оценивает участников сервиса с различных точек зрения.

Идея общенациональной системы социального оценивания сама по себе относительно проста: каждый гражданин для начала получает определенный

балл, а определенные действия либо снижают, либо увеличивают этот балл. Например, пожертвование на благотворительность увеличит рейтинг, в то время как покупка сигарет снизит его. После этого люди могут быть либо вознаграждены, либо наказаны в зависимости от их рейтинга. Правительство, например, может ограничить поездки человека или помешать ему поступить в лучшие университеты.

Несмотря на кажущиеся преимущества, при создании и внедрении платформ социального рейтинга остаются нерешенными на текущий момент 4 фундаментальных, по мнению автора, вопроса:

Искусственный интеллект далек от совершенства. Как и любая другая компьютерная система, он подвержен ошибкам принятия решения, подвержен предвзятости разработчика, в обучающих данных могут быть ложные корреляции
и циклы обратной связи, и, если это специально не оговорено разработчиком, алгоритмы не учитывают этические соображения. Простой ввод огромного количества информации в систему машинного обучения, а затем принятие результата без какой-либо критической оценки может привести к целому ряду непреднамеренных последствий, которые в конечном итоге приведут к нарушению прав определенных граждан.

Объективно необходимо ограничивать доступ к алгоритмам принятия решений и обрабатываемым данным ввиду того, что чем больше информации о системе ИИ общедоступно, тем больше в ней уязвимостей. С технической точки зрения система социального начисления баллов делает ее особенно уязвимой для искусственных манипуляций, нацеленных на занижение или завышение чьего-либо рейтинга
в корыстных целях.

Сами создатели систем ИИ, особенно глубоких искусственных нейронных сетей, не до конца понимают, как работает их система ввиду слабой интерпретируемости принимаемых решений. В госуправлении наоборот

необходимо обеспечивать прозрачность принимаемых решений во избежание дискриминации
и предвзятости.

Если контроль над такой системой будет сосредоточен в руках только одной социальной группы, имеющей возможность изменять правила работы системы,
то это может существенно повлиять на жизнь тех социальных групп, которые не в состоянии влиять на правила оценки.

Пути решения второго и третьего вопроса по сути противоречат друг другу
и представляют наибольший интерес для изучения. Их раскрытию и будет посвящена статья.

Для обеспечения безопасного внедрения систем социального рейтинга на основе ИИ в государственное управление крайне важно ограничить доступ к информации о структуре ядра этих систем. Это необходимо для предотвращения киберинцидентов и защиты персональных данных граждан.

Помимо стандартных программных уязвимостей, обнаруживаемых в любой распределенной цифровой системе, механизм оценивания, основанный на машинном обучении, несет дополнительный риск атак на его алгоритмы. Значительная часть задач в системе социального оценивания требует использования алгоритмов автоматического анализа больших объемов информации.
Это достигается с помощью нейронных сетей, основанных на современных методах глубокого обучения, которые очень уязвимы к состязательным атакам [3]. В этой связи целесообразно ограничение сведений о применяемых алгоритмах, используемых фреймворках, обучающих выборках данных, непосредственно весов обученных моделей ИИ.

С другой стороны, открытость алгоритмов и интерпретируемость принимаемых решений представляют существенную значимость для

избежания социальных рисков, связанных с обеспечением прозрачности, доверия и справедливости в обществе. Такие системы оказывают влияние на различные аспекты жизни людей, поэтому для избежания возможного злоупотребления и некорректной оценки социального рейтинга, особое внимание должно быть уделено открытости данных систем. Одной из важных составляющих систем социального рейтинга являются алгоритмы, которые позволяют оценивать и ранжировать информацию пользователей. Алгоритмы должны быть объективными, справедливыми и надежными, учитывая разнообразные критерии. При этом важно соблюдать прозрачность и открытость при определении весов критериев. Разработчики систем социального рейтинга должны публично объяснять, какие факторы принимаются во внимание и как они влияют на расчет рейтинга. Это позволит предотвратить произвольность и повысит уровень доверия пользователей к системе. Открытость систем социального рейтинга также требует публичности, что означает, что пользователи должны знать, что их оценивают и каким образом. Более того, должна быть обеспечена возможность пользователям просматривать собственные данные и вносить корректировки в случае ошибочной информации. Поэтому, при разработке и регулировании данных систем необходимо уделять особое внимание открытости и прозрачности, чтобы обеспечить справедливость и соответствие законодательству.

В Европе и Америке активно разрабатывается законодательство, обязывающее компании предоставлять людям понятную информацию о том, почему их рейтинг снизился или повысился [4, 5]). Таким образом предотвращается превращение системы в "черный ящик", когда пользователи и разработчики понимают, что должно делать программное обеспечение, но не понимают, как это происходит. Однако подобные нормативные акты разрабатываются не во всех странах. Это означает, что множество решений

может быть полностью передано в руки ИИ, что влечет за собой реальные риски.

Последствия решений, принятых системой социального оценивания, которая может определять будущее граждан, будучи одновременно простой в манипулировании и уязвимой для компьютерных атак – потенциально опасны.

Жизненный цикл системы социального рейтинга на базе технологий ИИ включает в себя четыре ключевых этапа. На каждом из них возникает потребность в выборе обоснованных стратегий для минимизации рисков, связанных с социальной справедливостью и кибербезопасностью. Особенно важно своевременно принять оптимальные меры по устранению этих потенциальных угроз, что представляется научно значимой задачей.

К основным потенциальным угрозам системы можно отнести:

1. На этапе разработки требований к системе и ее проектировании:

– некорректное определение самих требований, что может привести к созданию неэффективной или ненадежной системы;

– закрытость требований к системе, перечня анализируемой информации, критериев принятия решения, что может привести к отсутствию доверия
к системе;

– ошибки в спецификации, которые могут вызвать проблемы в процессе разработки и эксплуатации системы;

– недостаточное вниманием к обеспечению кибербезопасности, что может привести к уязвимостям в системе;

– утечка критически важной информации о системе, которая может быть использована для создания атак на систему.

2. На этапе сбора данных и обучении моделей ИИ:

– ошибки, допущенные при составлении обучающей выборки и

прежде всего при ее балансировке, то есть если данные собираются только от определенных групп людей, это может привести к тому, что система может работать
не корректно для других групп;

– ошибки допущенные при формализации цели системы при определении функции потерь, что ведет к необъективности рейтинга;

– выборе низкоинтерпретируемого алгоритма ИИ, лежащего в основе системы, что снижает доверие граждан к системе;

– утечке наборов данных, которые используются для обучения системы;

– преднамеренные манипуляции данными с целью искажения результатов работы и влияния на принимаемые в последствии решения.

3. На этапе тестирования и оптимизации источником рисков кибербезопасности как правило являются сниженные требования к безопасности на начальных этапах проекта, а так же спешка и требование ускоренного выход на этап тестирования до реализации всех необходимых мер киберзащиты.

4. На этапе внедрения системы и ее сопровождения имеют место стандартные риски кибербезопасности информационных систем, связанные с возможностью взлома системы, или ее заражения вредоносным программным обеспечением, утечками персональных данных, нарушения ее функционирования посредствам атак типа «отказ в обслуживании».

Для минимизации рисков, вызванных противоречивостью предъявляемых требований к системе видится целесообразным придерживаться следующих принципов при ее разработке и внедрении:

– организация тщательного анализа требований к информационной системе социального рейтинга, с учетом потребностей всех категорий граждан;

– минимизация перечня анализируемой информации;

– максимизация требований кибербезопасности;

– проведение широких общественных слушаний проекта;

– формирование монополии государства на формирование обезличенных выборок данных для обучения моделей ИИ;

– отказ от искусственных нейронных сетей ввиду их низкой интерпретируемости; выбор в пользу детерминированных алгоритмов, таких как деревья решений;

– проведение пилотных проектов для проверки концепции, только после реализации всех необходимых мер киберзащиты.

Анализ платформы социального рейтинга «Мы»

В РФ разработка системы социального рейтинга проводится в стенах Российского государственного социального университета (РГСУ). Проект платформы социального рейтинга «Мы» РГСУ представили в сентябре 2022 года. Авторы предлагают связать социальный рейтинг с обширным перечнем персональных данных (таблица 1). В РГСУ хотят получить доступ к такой информации, как СНИЛС, ИНН, паспорт, номер телефона, образование, наличие детей, источники дохода, льготы, данные о финансовых задолженностях, судимости, болезни, участие в общественной жизни, наличие государственных наград и так далее.

В 2022 году в своём докладе первый проректор вуза Джомарт Фазылович Алиев пояснил, что систему планируют предложить коммерческим банкам, губернаторам и государству [7].

Таблица 1.

Перечень персональных данных, обрабатываемой платформой социального рейтинга «Мы»

№	Наименование	Анкета	Импорт
1	ФИО полностью и идентификатор субъекта; может быть паспорт, ИНН, СНИЛС или иной ID на его усмотрение	—	—
2	Возраст, пол, уровень совокупных доходов как персональный, так и доступный как члена семьи	—	ДС, Ан
3	Семейное положение и реальный маритальный статус, как фактически сложился; без деталей и причин	да	ДС
4	Образование, профессия, квалификация; все имеющиеся с выделением тех, что активны и применяются в жизни	да	ДС
5	Место, условия, характер и правовая рамка проживания; с метриками; адрес может указываться опционально	част	Ан
6	Место работы, уровень или должность; при наличии множества вариантов - на усмотрение субъекта	част	МД
7	Дети и иные иждивенцы с их описательными краткими метриками без любых персональных данных	част	ДС
8	Долговой статус субъекта без расшифровок, без изъятий, без данных третьих лиц; включая алименты, налоги и т.п.	част	Ан
9	Медицинский статус субъекта и его иждивенцев с указанием фактов диагнозов без их раскрытия (опция)	да	ДС, Ан
10	Имущественный статус: все долговременные и крупные активы с оценкой и указанием локаций (не адресов)	част	Ан
11	Отношения с правоохранительной системой: приговоры, ограничения, открытые и/или активные состояния	да	Ан
12	Данные об уже используемых субъектом мерах социальной	да	МД

№	Наименование	Анкета	Импорт
	поддержки и/или ранее использовавшихся		
13	Опционально: знание языков, награды и звания, политический профиль, общественные роли	да	—

***«Анкета»**: прочерк означает, что система не даёт оцениваемому никаких подсказок в процессе заполнения, кроме структур (граф); «да» означает, что вплывающие словари выбора и справочники подсказа даются по всему разделу анкеты; «част» означает, что подсказки предлагаются частично, для некоторых граф.*

***«Импорт»**: прочерк означает, что для процессинга данных импорт какой-либо фактуры извне системы социального рейтинга не требуется (хотя и возможен); «Ан» означает, что в системе предусматривается импорт аналитических материалов; «ДС» означает импорт дата-сетов; «МД» означает импорт мастер-данных.*

Специалисты РГСУ ставят перед собой амбициозную задачу: «...видится целесообразным запустить опережающими темпами проект социального рейтинга, имея в виду создание надёжной, понятной, строгой и независимой от прикладных социальных ведомственных практик количественной метрики социальных статусов и уровней граждан посредством процедур социального оценивания. Выходные данные будут выпускаться в виде двухкомпонентного скоринг-кода (СК), где оба компонента будут двухразрядными числами, определяющими социальный статус человека (первый) и его социальный уровень (второй). Например 47-29. Социальный статус - отражение своего рода «заслуг» человека, а социальный уровень - оцифровка как-бы его «перспектив» [6]. При этом авторы инициативы отмечают, что оба названия компонентов, как и самого скоринг-кода, неудачные.

«Социальный статус» скоринг-кода авторы проекта определяют, как метрику накопленных качеств и особенностей человека, отражение его

совокупного опыта, образования, возраста, заболеваний, заслуг, социальных особенностей, наград, потребностей, подтверждённых прав и прочих нюансов, в совокупности характеризующих его значимость в социальном плане и градус социального внимания, проявление которого ожидается в отношении данного человека в моменте, но в связи с его прошлым [6].

«Социальный уровень» скоринг-кода авторы проекта определяют, как количественную характеристику достигнутых промежуточных результатов, специальности и профессии, должности, того же образования и возраста, амбиций и устремлений, достижений, мотивов и прочих особенностей, в совокупности характеризующих его потенциал в социальном плане и угол социальной перспективы, под которым ожидается проявление внимания к данному человеку в текущем моменте, но в связи с его будущим [6].

По словам авторов проекта, метрика «Социальный статус», например, многодетной матери-одиночки со средним образованием, проживающей в двухкомнатной квартире на условиях социального найма, воспитывающей ребёнка с ограниченными возможностями здоровья и на полставки работающей уборщицей в школе, легко и значительно может превышать социальный статус орденоносного академика, лауреата и директора института. «Социальный уровень» того же академика, по мнения авторов проекта, «вполне может уступать социальному уровню молодого неженатого мастера, работающего на машиностроительном заводе и производящего на продажу детские санки в своём гараже» [6].

Анализ публикаций о платформе «Мы» позволяет выделить положительные и отрицательные стороны проекта. К положительным следует отнести возможность оказания автоматической персонализированной социальной помощи, что является, конечно благом, но за эту возможность

придется заплатить отрицательными моментами, связанными с потенциальными рисками от внедрения системы. Более того, автоматическая персонализированная социальная помощь в том или ином виде реализована через интернет-портал Госуслуги.

Понятно, что это еще только проект и он находится в стадии проработки, но доступная информации о нем вызывает множество вопросов:

1. Прежде всего вызывает сомнение необходимость столь обширного набора критически важных персональных данных для предоставления заявленного перечня государственных услуг. Большое количество внешних связей с другими информационными системами для получения этих данных, несут риски кибербезопасности. К сожалению утечки персональных данных, а в проектируемой системе они действительно критически важные, происходят очень часто. Видится целесообразным минимизировать перечень персональных данных или внедрить механизмы гомоморфного шифрования, при которых модели ИИ обрабатывают шифрованные данные.

2. Сомнительна структура скоринг-кода. Авторы заявляют, что хотят создать «надёжную, понятную, строгую и независимую от прикладных социальных ведомственных практик количественную метрику социальных статусов и уровней граждан посредством процедур социального оценивания» [6] используя при этом только два измерения (социальный статус и социальный уровень). Вызывает большое сомнение возможность оптимальной, с точки зрения универсальности, прозрачности и не противоречивости, свертки того широкого перечня анкетных данных в два показателя. Очевидно показателей должно быть больше и с более осознанными названия, хотя бы потому, что социальный статус многодетной матери-одиночки со средним образованием, работающей уборщицей в школе, врядли может превышать социальный статус орденоносного академика, лауреата и директора института. В целом понятно, почему было выбрано только два показателя: для удобства восприятия специалистом МФЦ, но тогда еще более важно подобрать осмысленные

названия показателям.

3. Какие методы будут использованы для приведения в числовой вид таких анкетных данных, как например: «Медицинский статус субъекта и его иждивенцев с указанием фактов диагнозов» или «Имущественный статус: все долговременные и крупные активы с оценкой и указанием локаций»?

4. Как будут учтены корелляции в признаках? Для свертки анкетных данных в два итоговых показателя в РГСУ предлагают применить методы сегментирования и шкалирования, а после — взвешанное суммирование. При этом веса сегментов РГСУ предлагает не публиковать, поскольку они «составляют существенную часть ноу-хау». При этом очевидно, что сокрытие весов показателей не добавит доверия к системе. Пользователям системы точно захочется узнать на сколько государственная награда или должность директора института лучше (или хуже) звания матери-одиночки. Скорей всего эти веса будут очень спорными, поэтому авторы и не хотят их публиковать.

5. Какой порядок обжалования этого рейтинга, кто и в каком объеме будет нести ответственность за неправомерно назначенную или не назначенную социальную помощь? Разработчики системы или операторы?

На все эти вопросы, создающие социальную напряженность должны быть найдены ответы, как и на вопросы обеспечения безопасности персональных данных и киберзащиты самой автоматизированной системы.

Применение ИИ представляет собой инновационный подход к совершенствованию государственной и муниципальной системы управления. С одной стороны он позволяет существенно повысить эффективность решения ряда прикладных задач, но в то же время он может нести определенные риски, связанные с возможными ошибками в работе алгоритмов, нарушением конфиденциальности и безопасности данных, а так же возможностью манипуляции системой ИИ с целью нарушения прав граждан. Для минимизации этих рисков необходимо

разрабатывать и внедрять меры по обеспечению безопасности данных, проводить регулярный аудит систем и контролировать их работу, а также обеспечивать прозрачность и открытость процесса принятия решений с использованием ИИ.

В статье были рассмотрены некоторые вопросы реализации систем социального рейтинга на основе технологий искусственного интеллекта в государственном управлении. Сформулированы отдельные принципы их проектирования и внедрения. Кратко рассмотрен проект платформы социального рейтинга «Мы», разрабатываемой в Российском государственном социальном университете. Выявлены возможные проблемы его реализации.

Использованные источники:

1. Указ Президента РФ от 10.10.2019 N 490 (ред. от 15.02.2024) "О развитии искусственного интеллекта в Российской Федерации" (вместе с "Национальной стратегией развития искусственного интеллекта на период до 2030 года") // СЗ РФ. - 14.10.2019. - № 41. - Ст. 5700.

2. Островский, А.С. Государственное регулирование использования искусственного интеллекта в государственном управлении [Электронный ресурс] / А.С. Островский - Текст : электронный // Аллея науки. - 2022. - № 2(65). - URL: https://www.elibrary.ru/download/elibrary_48614179_82736031.pdf (дата обращения: 14.03.2024).

3. Xu H. et al. Adversarial attacks and defenses in images, graphs and text: A review //International journal of automation and computing. – 2020. – Т. 17. – С. 151-178.

4. The Equal Credit Opportunity Act : [Электронный ресурс]. - URL: https://www.govinfo.gov/content/pkg/USCODE-2011-title15/html/USCODE-2011-title15-chap41-subchapIV.htm (дата обращения: 14.03.2024). - Текст : электронный.

5. General Data Protection Regulation : [Электронный ресурс]. - URL: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A02016R0679-20160504 (дата обращения: 14.03.2024). - Текст : электронный.

6. Проект платформы социального скоринга (материалы к Учёному совету РГСУ) : [Электронный ресурс]. - URL: https://rgsu.net/about/activities/press_centre/life/platform/rgsu-life_20.html (дата обращения: 14.03.2024). - Текст : электронный.

7. В сети выяснили, что РГСУ разрабатывает систему социального рейтинга под названием «Мы» : [Электронный ресурс]. - URL: https://habr.com/ru/news/765866/ (дата обращения: 14.03.2024). - Текст : электронный.