Редакция
ноябрь 2023

Рекомендации Комиссии по реализации Кодекса этики в сфере ИИ по теме: **«Прозрачность алгоритмов искусственного интеллекта и информационных систем на их основе»**

Содержание

РЕЗЮМЕ

Рекомендации по обеспечению прозрачности алгоритмов искусственного интеллекта и информационных систем на их основе (далее – Рекомендации) являются обобщением опыта и представлений членов Комиссии по реализации Кодекса этики в сфере искусственного интеллекта (далее – Комиссия), а также экспертов, принявших участие в обсуждении вопроса прозрачности алгоритмов искусственного интеллекта. Рекомендации предназначены для широкого круга разработчиков алгоритмов искусственного интеллекта, а также информационных систем и цифровых сервисов на их основе (далее – Разработчики, Алгоритмы).

С помощью Рекомендаций Разработчики смогут оценить комплексность и соразмерность предпринятых усилий по обеспечению прозрачности Алгоритмов, а также спланировать и подготовить следующие шаги по раскрытию информации об Алгоритмах.

ОСНОВНАЯ ПРОБЛЕМА ПРОЗРАЧНОСТИ

Обеспечить предсказуемость последствий использования Алгоритма и, как следствие, более высокий уровень доверия к нему со стороны существующих и потенциальных пользователей, а также со стороны третьих лиц, чьи права могут быть затронуты в результате легального использования Алгоритмов.

Как следствие, решение основной задачи прозрачности Алгоритмов позволяет обеспечить:

- реализацию потребителем своего права на осознанный и самостоятельный выбор Алгоритма среди аналогов;
- добросовестность действий Разработчика при раскрытии значимых фактов в отношении Алгоритма;
- накопление и систематизацию значимых для потребителя особенностей поведения Алгоритма как до проектирования информационных систем и сервисов с его использованием, так и в случае возникновения непредвиденных ситуаций;
- возможность распределения ответственности между Разработчиками в сложных информационных системах и сервисах, в которых различные Алгоритмы используются наряду с традиционными методами жесткого программирования поведения таких систем.

МЕТОД ВЫПОЛНЕНИЯ ЗАДАЧИ

Последовательное и системное раскрытие значимых фактов в отношении Алгоритмов Разработчиками, которое включает в себя:

- **первичное раскрытие информации** об Алгоритме, которое осуществляется Разработчиком при принятии решения о предоставлении широкой общественности информации о его разработке в любой момент времени, обусловленный привлечением инвестиций, грантового финансирования либо иной активности Разработчика;

- **дополнительное раскрытие информации** при появлении результатов, значимым образом влияющих на оценку возможных последствий использования Алгоритма;

- **раскрытие информации потенциальному клиенту,** предполагающее добросовестное информирование лиц, принимающих решение об использовании или внедрении Алгоритма в тех или иных сферах.

Раскрытие значимых фактов Разработчиками в отношении Алгоритмов может осуществляться в свободном формате, удобном для широкой общественности и отдельных категорий потребителей такой информации, включая форматированные полнотекстовые документы, веб-страницы на корпоративных сайтах, специализированные машиночитаемые форматы для последующего использования в базах данных и инструментах повышения индивидуальной производительности труда (включая цифровые персональные помощники). При этом Комиссия рекомендует обеспечить неизменность опубликованных документов, а в случае необходимости их изменения – надлежащее версионирование для сохранения доступа к документам, содержавшим ранее опубликованную информацию.

Комиссия полагает целесообразным со временем сформировать по отдельным значимым фактам надлежащую инфографику, классификаторы и иные способы повышения понятности документов для сокращения временных издержек пользователей при ознакомлении с информацией в отношении значимых фактов прозрачности Алгоритмов.

С учетом специфики деятельности Разработчиков, применение всех без исключения Рекомендаций остается возможной, но не обязательной мерой по обеспечению прозрачности Алгоритмов. При этом меры по обеспечению прозрачности Алгоритмов в отдельных случаях (например, до заключения соглашения с потенциальным заказчиком либо инвестором) могут быть отнесены как к законному интересу заказчика (инвестора), и, как следствие, могут использоваться в качестве одного из условий заключения соглашения. Равно как и отсутствие возможности раскрытия той или иной информации Разработчиком может быть отнесено к законному интересу такового, если соответствующее раскрытие может привести к нарушению коммерческой тайны, о чем потенциальные заказчики уведомляются по их требованию.

ПЕРВИЧНОЕ РАСКРЫТИЕ ИНФОРМАЦИИ

Первичное раскрытие информации об Алгоритме осуществляется Разработчиком добровольно, адресовано широкой общественности, общедоступно и содержит следующую информацию о значимых фактах:

- **цели обучения Алгоритма:** какие цели были поставлены перед Алгоритмом при его обучении (с использованием простой лексики);
- **метрики оценки эффективности Алгоритма:** какая функция от каких параметров оптимизировалась при проведении машинного обучения (с использованием профессиональной лексики, включая математические формулы);
- **состав обучающей выборки данных:** какие данные использовались для обучения, были ли это стандартизированные наборы данных, если нет – то указывается метод сбора данных и иные важные особенности, способные по мнению Разработчика повлиять на статистические свойства обучающей выборки по сравнению с генеральной выборкой;
- **использованные алгоритмы машинного обучения:** описание на профессиональном языке использованных стандартных алгоритмов и их комбинаций в формате, не нарушающем коммерческие права Разработчиков.

Кроме этого, на основании указанных значимых фактов при первичном раскрытии информации от Разработчиков ожидаются аргументированные суждения о следующих значимых фактах:

- **известные Разработчику недостатки обучающей выборки данных,** включая указание категорий либо комбинаций свойств объектов, заведомо либо вероятно в ней отсутствующих вследствие источника данных, либо особенностей их сбора;

- **известные Разработчику недостатки использованных Алгоритмов,** включая указание на недостатки, которые обусловлены их комбинацией;

- **результаты, полученные Разработчиком на собственных тестовых выборках данных,** включая индикативное информирование о доле ложно-положительных и ложно-отрицательных рекомендаций при их применимости, а также характеристиках объектов, на которых Алгоритм работает лучше или хуже, чем в среднем на тестовой выборке данных.

Комиссия полагает не этичным для Разработчиков намеренно затруднять идентификацию потребителями недостатков обучающей выборки данных или использованных при обучении алгоритмов.

РАСКРЫТИЕ ПРИ ПОЯВЛЕНИИ ИНФОРМАЦИИ

Предполагается, что перед выводом Алгоритма на рынок (включая реализацию первых пилотных внедрений), Разработчик в рамках механизма обеспечения прозрачности Алгоритма составляет **первичные рекомендации Разработчика по сфере применения Алгоритма.** Рекомендуется обеспечить соответствие таких рекомендаций следующим требованиям:

- должны отвечать на вопрос, для каких целей и в каких условиях Разработчик рекомендует, а для каких – не рекомендует применять Алгоритм;
- должны быть понятны широкой общественности, в идеале – исключать возможность двусмысленности либо возникновения неверного понимания последствий использования Алгоритма у значимой части пользователей;
- объясняют рекомендуемые Разработчиком ограничения на основе значимых фактов первичного раскрытия информации.

Кроме того, в состав сведений, раскрываемых при появлении информации, рекомендуется включать следующие значимые факты:

- **результаты, полученные Разработчиком на публичных выборках данных,** причем как в случаях, в которых эффективность Алгоритма подтверждена или даже оказалась выше ожиданий Разработчика, так и в случаях, если она оказалась ниже (считается неэтичным скрывать информацию о таких результатах);
- **эффективность Алгоритма по отношению к отраслевым стандартам (benchmarks),** в особенности, если такая эффективность продемонстрирована в публичных конкурсах, состязаниях и т.п.;

- **эффективность Алгоритма по сравнению с эффективностью людей,** выполняющих сходные интеллектуальные операции, при этом рекомендуется информировать пользователей об источниках информации и условиях оценки эффективности людей, включая ссылки на соответствующие исследования или научные публикации;

- **случаи, когда Алгоритм работает хуже, чем в среднем:** установленные Разработчиком либо его клиентами особенности эффективности Алгоритма в тех или иных специальных условиях – Разработчику рекомендуется информировать широкую общественность об условиях, при которых пользователям не следует ожидать достижение тех показателей эффективности, которые содержатся в общих рекламно-информационных материалах Разработчика.

- **применимые специальные условия использования Алгоритма,** к которым следует относить условия использования персональных данных или объектов авторского права или интеллектуальной собственности третьих лиц, если таковые применялись при обучении Алгоритма или содержались в обучающей либо тестовой выборке, наличие отраслевой сертификации, особенно если таковая является обязательным требованием использования Алгоритма в определенных отраслях и т.п.

РАСКРЫТИЕ ПРИ ОЦЕНКЕ ПРИМЕНИМОСТИ У ПОТЕНЦИАЛЬНОГО ЗАКАЗЧИКА

Разработчикам, добросовестно применяющим Рекомендации для повышения прозрачности Алгоритма при взаимодействии с потенциальным заказчиком, для поддержки принятия осознанных и самостоятельных решений рекомендуется предоставлять следующую информацию:

- **случаи внедрения Алгоритма с доказанной эффективностью,** при этом рекомендуется использовать разумно обезличенные либо (в случае его согласия) персонализированные сведения о предыдущем заказчике с указанием тех метрик эффективности, которые были достигнуты в условиях реального использования Алгоритма;

- **рекомендации по сфере применения,** с учетом ожидаемых потенциальным заказчиком вариантов использования Алгоритма, в том числе информируя такого заказчика, в какой мере случаи использования Алгоритма, известные и проанализированные Разработчиком, покрывают все множество вариантов его использования, заявленных потенциальным заказчиком;

- **оценка безопасности применения** (также по отношению к вариантам использования Алгоритма потенциальным заказчиком), в частности, в случаях, когда Разработчику известны обстоятельства применения Алгоритма, которые в условиях потенциального заказчика не приведут к достижению необходимых последнему параметров эффективности, Разработчику рекомендуется включать такие случаи в перечень не рекомендуемых вариантов использования Алгоритма.

Раскрытие значимых для потенциального заказчика фактов должно осуществляться на взаимовыгодной основе. Как следствие, Разработчик вправе самостоятельно определить возможность такого раскрытия информации до заключения возмездного соглашения, либо

предусмотреть такую возможность в качестве первого этапа работ по внедрению Алгоритма, которая оплачивается потенциальным заказчиком как самостоятельная услуга либо работа.

Вместе с тем, считается неэтичным составление соглашений, которые обязывают потенциального заказчика оплачивать штрафы или неустойки в пользу Разработчика по результатам расторжения соглашения на основании результатов первого этапа работ по внедрению Алгоритма, предусматривавшего предоставление персонализированных рекомендаций Разработчика, если такие результаты продемонстрировали заказчику высокие риски для безопасности применения Алгоритма либо рекомендации не применять Алгоритм в условиях заказчика.