Конспект лекции

С.С. Жумажанова, П.С. Ложников

Этика и правовые проблемы искусственного интеллекта

Этические аспекты доверия к ИИ

ЭТИЧЕСКИЕ АСПЕКТЫ ДОВЕРИЯ К ИИ

**1 Понятие доверенного ИИ**

*Ethics guidelines for trustworthy AI*: Доверенный ИИ состоит из трех компонентов, которые должны соблюдаться на протяжении всего жизненного цикла системы:

(1) он должен быть законным, соответствовать всем применимым законам и постановлениям;

(2) он должен быть этичным, обеспечивать соблюдение этических принципов и ценностей;

(3) он должен быть надежным (робастным) как с технической, так и с социальной точки зрения, поскольку даже при благих намерениях системы ИИ могут причинить непреднамеренный вред. Каждый компонент сам по себе необходим, но недостаточен для достижения ИИ, заслуживающего доверия. В идеале все три компонента работают согласованно и частично совпадают в своей работе. Если на практике между этими компонентами возникает несогласование, общество должно стремиться их согласовать.

*ГОСТ Р 59 276−2020 «Системы искусственного интеллекта. Способы обеспечения доверия. Общие положения»:*

*Доверие к системе искусственного интеллекта*: Уверенность потребителя, и при необходимости, организаций, ответственных за регулирование вопросов создания и применения систем искусственного интеллекта, и иных заинтересованных сторон в том. что система способна выполнять возложенные на нее задачи с требуемым качеством.

*Доверенная система искусственного интеллекта*: Система искусственного интеллекта, в отношении которой потребитель и. при необходимости, организации, ответственные за регулирование вопросов создания и применения систем искусственного интеллекта, проявляют доверие.

**2 Этические аспекты доверия к ИИ**

*Этика ИИ* - это подраздел прикладной этики, в котором основное внимание уделяется этическим вопросам, возникающим при разработке, развертывании и использовании ИИ. Его основная задача - определить, как ИИ может способствовать или вызывать озабоченность в отношении хорошей жизни людей, с точки зрения качества жизни или человеческой автономии и свободы, необходимых для демократического общества. Доверенный ИИ может улучшить индивидуальное процветание и коллективное благополучие, обеспечивая процветание, создание ценности и максимизацию благосостояния. Он может способствовать построению справедливого общества, помогая укреплять здоровье и благополучие граждан таким образом, чтобы способствовать равенству в распределении экономических, социальных и политических возможностей.

Уважение основных прав в рамках демократии и верховенства закона является наиболее многообещающей основой для определения абстрактных этических принципов и ценностей, которые могут быть реализованы в контексте ИИ.

* *Уважение человеческого достоинства:* каждый человек обладает «внутренней ценностью», которая никогда не должна уменьшаться, компрометироваться или подавляться другими - или новыми технологиями, такими как системы ИИ. В этом контексте уважение человеческого достоинства предполагает, что ко всем людям необходимо относиться с уважением как к моральным субъектам, а не просто как к объектам, относительно которых совершает действие система ИИ. Следовательно, системы искусственного интеллекта должны разрабатываться таким образом, чтобы уважать, обслуживать и защищать физическую и психическую целостность людей, личную и культурную самобытность и удовлетворение их основных потребностей.
* *Свобода личности:* люди должны оставаться свободными в принятии жизненных решений за себя. Это влечет за собой свободу от суверенного вторжения, но также требует вмешательства со стороны правительства и неправительственных организаций для обеспечения того, чтобы отдельные лица или люди, находящиеся под угрозой исключения, имели равный доступ к преимуществам и возможностям ИИ. Фактически, свобода личности означает обязательство дать людям возможность еще больше контролировать свою жизнь, включая (среди других прав) защиту свободы ведения бизнеса, свободы искусства и науки, свободы слова, право на частную жизнь и неприкосновенность частной жизни, а также свободу собраний и ассоциаций.
* *Уважение к демократии, справедливости и верховенству закона*: вся правительственная власть в конституционных демократиях должна быть санкционирована и ограничена законом. Системы искусственного интеллекта должны служить для поддержания и развития демократических процессов и уважения множественности ценностей и жизненного выбора людей. Системы искусственного интеллекта не должны подрывать демократические процессы, человеческое мышление или демократические системы голосования. Системы искусственного интеллекта также не должны подрывать фундаментальные обязательства, на которых основывается верховенство закона, обязательные законы и постановления, а также обеспечивать надлежащую правовую процедуру и равенство перед законом.
* *Равенство, недискриминация и солидарность* - включая права лиц, находящихся под угрозой исключения. Необходимо обеспечить равное уважение моральной ценности и достоинства всех людей. Это выходит за рамки недискриминации, которая допускает проведение различий между разнородными ситуациями на основе объективных оправданий. В контексте ИИ равенство влечет за собой, что операции системы не могут генерировать несправедливо предвзятые результаты (например, данные, используемые для обучения систем ИИ, должны быть как можно более всеобъемлющими и представлять разные группы населения).
* *Гражданские права:* системы искусственного интеллекта обладают значительным потенциалом для увеличения масштабов и эффективности правительства в предоставлении общественных благ и услуг обществу. В то же время права граждан также могут подвергаться негативному воздействию со стороны систем искусственного интеллекта, и их следует защищать.

Системы искусственного интеллекта должны улучшать индивидуальное и коллективное благополучие. Существует четыре этических принципа, которые необходимо соблюдать, чтобы гарантировать, что системы ИИ разрабатываются, развертываются и используются доверенным образом:

* *Принцип уважения автономии человека:* люди, взаимодействующие с системами ИИ, должны иметь возможность сохранять полное и эффективное самоопределение над собой и иметь возможность участвовать в демократическом процессе. Системы искусственного интеллекта не должны необоснованно подчинять, принуждать, обманывать, манипулировать людьми. Напротив, они должны быть разработаны для расширения, дополнения возможностей когнитивных, социальных и культурных навыков человека. Распределение функций между людьми и системами ИИ должно соответствовать принципам проектирования, ориентированным на человека, и оставлять значимую возможность для человеческого выбора. ИИ должен поддерживать людей в рабочей среде и стремиться к созданию значимой работы.
* *Принцип предотвращения вреда:* системы искусственного интеллекта не должны ни причинять, ни усугублять вред или иным образом отрицательно влиять на людей. Системы искусственного интеллекта и среда, в которой они работают, должны быть безопасными и надежными. Они должны быть технически надежными и не допускать злонамеренного использования. Уязвимым группам людей следует уделять больше внимания и привлекать их к разработке, развертыванию и использованию систем искусственного интеллекта. Предотвращение вреда также влечет за собой учет окружающей среды и всех живых существ.
* *Принцип справедливости:* разработка, развертывание и использование систем искусственного интеллекта должны быть справедливыми. Приверженность: обеспечению равного и справедливого распределения как выгод, так и затрат и обеспечение того, чтобы отдельные лица и группы были свободны от несправедливой предвзятости, дискриминации и стигматизации. Для этого субъект, ответственный за принятое решение, должен быть идентифицируемым, а процессы принятия решений - объяснимыми.
* *Принцип объяснимости*: расширяемость имеет решающее значение для создания и поддержания доверия пользователей к системам искусственного интеллекта. Это означает, что процессы должны быть прозрачными, возможности и цели систем ИИ должны быть открытыми, а решения - насколько это возможно - объяснимыми для тех, кто прямо или косвенно затронут. Без такой информации решение не может быть оспорено должным образом. Объяснение того, почему модель сгенерировала конкретный результат или решение (и какая комбинация входных факторов способствовала этому), не всегда возможно. Эти случаи называются алгоритмами «черного ящика» и требуют особого внимания. В этих обстоятельствах могут потребоваться другие меры по объяснению (например, отслеживаемость, возможность аудита и прозрачное сообщение о возможностях системы) при условии, что система в целом соблюдает основные права. Степень, в которой требуется объяснение, сильно зависит от контекста и серьезности последствий, если этот вывод ошибочен или неточен по иным причинам.

**3 Российские и мировые стандарты, связанные с доверенными системами ИИ**

Национальные и международные стандарты по доверенному ИИ:

* ГОСТ Р 59276-2020. Национальный Стандарт Российской Федерации «Системы искусственного интеллекта. Способы обеспечения доверия. Общие положения»
* Ethics guidelines for trustworthy AI
* ГОСТ Р ИСО/МЭК 19784-1–2007 Автоматическая идентификация. Идентификация биометрическая. Биометрический программный интерфейс. Ч. 1. Спецификация биометрического программного интерфейса.
* ГОСТ Р 58293-2018. (ИСО/МЭК 19785-1:2015) Национальный Стандарт Российской Федерации. Информационные технологии. Биометрия. Единая структура форматов обмена биометрическими данными. Часть 1. Спецификация элементов данных.

**Список источников**

1. Ethics guidelines for trustworthy ai [electronic resource]. URL: <https://www.aepd.es/sites/default/files/2019-12/ai-ethics-guidelines.pdf>.
2. ГОСТ Р ИСО/МЭК 9126-93 Информационная технология (ИТ). Оценка программной продукции. Характеристики качества и руководства по их применению от 28 декабря 1993 - docs.cntd.ru [electronic resource]. URL: <https://docs.cntd.ru/document/1200009076>.
3. ГОСТ Р 59276-2020 Системы искусственного интеллекта. Способы обеспечения доверия. Общие положения от 23 декабря 2020 - docs.cntd.ru [electronic resource]. URL: <https://docs.cntd.ru/document/1200177291>.
4. ГОСТ Р ИСО/МЭК 19784-1-2007 Автоматическая идентификация. Идентификация биометрическая. Биометрический программный интерфейс. Часть 1. Спецификация биометрического программного интерфейса (с изменением n 1) от 25 декабря 2007 - docs.cntd.ru [electronic resource]. URL: <https://docs.cntd.ru/document/1200067412>.
5. ГОСТ Р 58293-2018 (ИСО/МЭК 19785-1:2015) Информационные технологии (ИТ). Биометрия. Единая структура форматов обмена биометрическими данными. Часть 1. Спецификация элементов данных от 05 декабря 2018 - docs.cntd.ru [electronic resource]. URL: <https://docs.cntd.ru/document/1200161701>.