Этические рекомендации для этически обоснованного проектирования

Концепция приоритезации жизнедеятельности людей при взаимодействии с искусственным интеллектом и автономными системи

Резюме

Чтобы полностью использовать потенциал искусственного интеллекта и автономных систем (ИИ/АС), нужно мыслить шире, чтобы раскрыть потенциал большой вычислительной мощности или возможностей решения.

Мы должны убедиться, что эти технологии соответствуют человечеству с точки зрения моральных ценностей и этических принципов. ИИ/АС, помимо достижения функциональных целей и решения технических проблем, должны работать таким образом, чтобы быть удобными людям. Это обеспечит повышенный уровень доверия между человечеством и нашими технологиями, которые необходимы для плодотворного повсеместного использования ИИ/АС в нашей повседневной жизни.

Эвдаймония, как объяснял Аристотель, - это практика, которая определяет благополучие людей как высшей добродетели для общества. Примерно переведенная как «процветание», польза эудаймонии начинается с осознанного созерцания, где этические соображения помогают нам определить, как мы хотим жить.

Согласуя создание ИИ/АС с ценностями пользователей и общества, мы можем уделить приоритетное внимание возрастанию благополучия людей как нашего показателя прогресса в алгоритмической эпохе.

Онас

Глобальная инициатива IEEE по этическим рекомендациям для искусственного интеллекта и автономных систем («IEEE Global Initiative») является программой Института Электротехники и Электроники («IEEE»), крупнейшей в мире технической профессиональной организацией, посвященной продвижению технологий на благо человечества с более чем 400 000 членами в более чем 160 странах.

Глобальная инициатива IEEE обеспечивает возможность объединить многочисленные мнения сообществ искусственного интеллекта и автономных систем для выявления и нахождения консенсуса по актуальным вопросам.

IEEE разработает рекомендации для этически обоснованного проектирования (Ethically Aligned Design EAD), доступны под некоммерческой лицензией Соединенных Штатов Америки Creative Commons Attribution-Non-Commercial 3.0.

В соответствии с условиями этой лицензии организации или отдельные лица могут принять аспекты этой работы на их усмотрение в любое время. Также ожидается, что содержание и предмет проблемыЭтически обоснованного проектирования будут выбраны для ввода в формальные процессы IEEE, в том числе для разработки стандартов.

Глобальная инициатива IEEE и этические рекомендации способствуют значительным усилиям, предпринимаемым в IEEE для открытого, широкого и содержательного разговора об этике в технологиях, известной как программа IEEE TechEthics ™.

Миссия Глобальной инициативы IEEE

Каждый научно-технический работник должен быть обучен, подготовлен и мотивирован делать приоритетными этические соображения при разработке и развитии автономных и интеллектуальных систем.

Под «научно-техническим работником» мы подразумеваем любого, кто участвует в исследованиях, разработке, производстве или обмене сообщениями в области ИИ/АС, включая университеты, организации и корпорации, претворяющие эти технологии в реальность для общества.

Этот документ представляет собой коллективный труд более ста человек с мировым именем в областях искусственного интеллекта, права и этики, философии и политики в сфере научных сообществ, науки, правительственных и корпоративных секторов. Наша цель состоит в том, чтобы рекомендации для этически обоснованного проектирования могли предоставить информацию и предложения от этих специалистов, обеспечивающих ключевые компетенции для работы научно-технический сотрудников ИИ/АС в ближайшие годы. Для достижения этой цели в текущей версии рекомендаций для этически обоснованного проектирования (ЭОП версия1) мы определяем проблемы и рекомендации для специалистов в областях, охватывающих искусственный интеллект и автономные системы.

Вторая цель Глобальной инициативы IEEE - предоставлять рекомендации для стандартов IEEE, основанных на ЭОП. IEEE P7000 ™ - процесс моделирования для решения этических проблем в процессе проектирования системы был первымпроектом стандарта IEEE (утвержден и находится в разработке), вдохновленным инициативой. Два других проекта по стандартизации, IEEE P7001TM – информационная открытость автономных систем и IEEE P7002TM – процесс сохранения конфиденциальности данных, были одобрены, что демонстрирует общественное влияние инициативы на вопросы этики ИИ/АС.

Структура и содержание документа

ЭОП включает восемь разделов, каждый из которых касается конкретной темы, связанной с ИИ/АС, котораяподробно обсуждается конкретнойрабочей группой Глобальной инициативы IEEE. Вопросы и рекомендации для разработчиков, относящиеся к этим темам, перечислены в каждом разделе. Ниже приводится сводка основных направлений и вопросов, охватываемых в их разделах:

1 | Общие принципы

Рабочая группа по общим принципам сформулировала этические проблемы высокого уровня применительно ко всем типам ИИ/АС:

1. Воплощение высших идеалов прав человека.

2. Приоретизация максимальной выгоды для человечества и окружающей среды.

3. Смягчение рисков и негативных последствий в процессе эволюции ИИ/АС как социально-технических систем.

Цель рабочей группы состоит в том, чтобы сформулированные ею принципы, проблемы и рекомендации специалистов в конечном итоге послужили основой для разработки и поддержки будущих норм и стандартов в рамках новой системы этического управления для проектирования ИИ/АС.

Вопросы:

1. Как мы можем гарантировать, что ИИ/АС не нарушают права человека?
2. Как мы можем быть уверены, что ИИ/АС могут быть привлечены к ответственности?
3. Как мы можем обеспечить прозрачность ИИ/АС?
4. Как мы можем расширить преимущества и свести к минимуму риски злоупотребления технологией ИИ/АС?

2 | Интеграция ценностей в автономные интеллектуальные системы

Для развития успешных автономных интеллектуальных систем (АИС), которые принесут пользу обществу, важно, чтобы техническое сообщество осознало и смогло внедрить соответствующие человеческие нормы или ценности в свои системы. Рабочая группа по интеграции этических ценностей в автономные интеллектуальные системы взяла на себя более широкую задачу интеграции ценностей в АИС в качестве трехстороннего подхода, помогая разработчикам:

1. Определить нормы и ценности конкретного сообщества, на которое влияют АИС;
2. Внедрить нормы и ценности этого сообщества в рамках АИС;
3. Оценить соответствие и совместимость этих норм и ценностей между людьми и АИС в рамках этого сообщества.

Вопросы:

•Ценности, которые должны быть внедрены в АИС, не являются универсальными, но в значительной степени специфичны для сообществ пользователей и задач.

•Моральная перегрузка: АИС обычно подчиняются множеству норм и ценностей, которые могут конфликтовать друг с другом.

•АИС может иметь встроенные данные или алгоритмические ошибки, которые ставят в невыгодное положение членов определенных групп.

•После определения соответствующих наборов норм (определенной роли АИС в конкретном сообществе) неясно, как такие нормы должны быть встроены в вычислительную архитектуру.

•Нормы, внедренные в АИС, должны быть совместимы с нормами в соответствующем сообществе.

•Достижение достаточного уровня доверия между людьми и АИС.

•Сторонняя оценка гармонизации ценностей АИС.

3 | Методологии для руководства этическими исследованиями и разработками

Современное формирование ИИ/АС должно гарантировать, что благополучие людей, расширение прав и возможностей и свобод лежат в основе развития ИИ/АС. Для создания машин, которые могут достичь этих целей, в *методических указаниях для руководства рабочей группой по этическим исследованиям и проектам* были сформулированы вопросы и рекомендации специалистов для обеспечения того, чтобы человеческие ценности, такие как права человека, определенные во Всеобщей декларации прав человека, были оформлены как методологии проектирования систем. Ценностно-ориентированные методологии проектирования должны стать основным предметом для организаций ИИ/АС, ориентированных на развитие людей на основе этических принципов. Машины должны обслуживать людей, а не наоборот. Этот этически обоснованный подход обеспечит равновесие между сохранением экономической и социальной доступности ИИ как для бизнеса, так и для общества.

Вопросы:

•Этика не входит в программы обучения.

•Необходимы модели междисциплинарного и межкультурного образования для учета отдельных вопросов ИИ/АС.

•Необходимость дифференцировать самобытные культурные ценности, встроенные в разработку ИИ.

•Отсутствие основанной на ценностях этической культуры и практики для промышленности.

•Отсутствие главных ценностей.

•Отсутствие возможностей для поднятия этических проблем.

•Отсутствие заинтересованности или ответственности со стороны технического сообщества.

•Необходимость учета мнения заинтересованных сторон для лучшего применения ИИ/АС.

• Наличие плохой документации, которая препятствует разработке этики.

•Непоследовательность или отсутствие контроля над алгоритмами.

•Отсутствие независимой наблюдательной организации.

•Использование компонентов черного ящика.

4 | Безопасность и использование основного искусственного интеллекта (ОИИ) и искусственного супер-интеллекта (ИСИ)

Будущие высокоэффективные системы искусственного интеллекта (иногда называемые основным искусственным интеллектом или ОИИ) могут оказывать преобразующее воздействие на мир в масштабах сельскохозяйственных или промышленных революций, что может привести к беспрецедентным уровням глобального процветания. Рабочая группа по безопасности и преимуществам основного искусственного интеллекта и искусственного супер-интеллекта (СИИ) предоставила множество вопросов и рекомендаций специалистов с целью обеспечения положительных результатов этой трансформации благодаря согласованным усилиям сообщества ИИ.

Вопросы:

• Поскольку системы ИИ становятся более функциональными - что измеряется способностью оптимизировать более сложные объективные функции с большей автономностью в более широком диапазоне областей - непредвиденное или непреднамеренное поведение становится все более опасным.

• Повышение безопасности в будущих, более общедоступных системах ИИ может быть затруднено.

• Исследователи и разработчики будут сталкиваться со все более сложным набором этических и технических вопросов безопасности при разработке и внедрении все более автономных и функциональных систем ИИ.

• Будущие системы ИИ могут иметь потенциал для воздействия на мир в масштабах сельскохозяйственных или промышленных революций.

5 | Личные данные и индивидуальный контроль доступа

Ключевой этической дилеммой в отношении личной информации является *асимметрия данных*. Для решения проблемы асимметрии рабочая группа по персональным данным и индивидуальному контролю доступа выявила проблемы и предоставила рекомендации специалистов, демонстрирующие фундаментальную потребность людей в *определении, доступе* и *управлении* своими персональными данными. Рабочая группа признает, что нет идеальных решений и что любой цифровой инструмент может быть взломан. Тем не менее она рекомендует включить среду данных, в которой люди контролируют свое чувство собственного достоинства, в система ИИ/АС и предоставляет примеры инструментов и разработанных практик, которые могут искоренить асимметрию данных в будущем.

Вопросы:

•Как человек может определить и организовать свои персональные данные в алгоритмическую эпоху?

•Каково определение и объем личной информации?

•Каково определение контроля над персональными данными?

•Как мы можем переопределить доступ к персональным данным, чтобы не оскорбить человека?

• Как мы можем переопределить согласие в отношении обработки персональных данных, чтобы оно не оскорбило человека?

•Данные, которые являются тривиальными для распространения, могут быть использованы для предположений, которыми человек не хотел бы делиться.

•Как обработчики данных могут обеспечить информацию для индивида о последствиях (положительных и отрицательных) доступа и сбора данных, чтобы он мог дать действительно осознанное согласие?

•Может ли человек иметь персональный ИИ или алгоритмического помощника?

6 | Пересмотр автономных систем вооружения

Автономные системы, предназначенные для нанесения физического вреда, имеют дополнительные этические последствия по сравнению с традиционным оружием и автономными системами, которые не предназначены для нанесения вреда. Профессиональная этика на эту тему может и должна иметь более высокие стандарты, охватывающие более широкий круг проблем. В широком смысле, рабочая группа по реформированию систем автономного вооружения призывает технические организации признать значимость человеческого контроля над системами вооружения для общества. Контрольные журналы, гарантирующие подотчетность, должны обеспечивать понимание последствий работы тех, кто создает эти технологии. Профессиональные этические коды должны надлежащим образом применены к системам, которые предназначены для причинения вреда.

Вопросы:

•Профессиональные кодексы поведения организаций часто имеют большие лазейки, в результате чего упускается из виду работа членов организаций, создаваемые ими артефакты и агенты с теми же ценностями и стандартами, которыми придерживаются сами представители организаций.

•Путаница в отношении определений относительно важных концепций в искусственном интеллекте, автономных системах и системах автономного вооружения (САВ) останавливает более существенные обсуждения важнейших вопросов.

•САВ по умолчанию поддается скрытому и неприменимому использованию.

•Существует несколько способов, с помощью которых ответственность за действия САВ может быть снижена.

•CAB может быть непредсказуема (в зависимости от ее проектирования и эксплуатации). Системы обучения объединяют проблему предсказуемого использования.

•Легитимизация разработки CAB имеет прецеденты, которые в среднесрочной перспективе являются геополитически опасными.

•Исключение человеческого контроля над боевым пространством может слишком легко привести к непреднамеренному нарушению прав человека и эскалации напряженности.

•Разнообразие прямых и косвенных покупателей CAB приведет к распространению и злоупотреблению этими системами.

•По умолчанию тип автоматизации в CAB способствует быстрой эскалации конфликтов.

•Нет стандартов для проверки надежности CAB.

•Понимание этических границ работы над системами CAB и полуавтономного оружия может сбить с толку.

7 | Экономика/Гуманитарные вопросы

Технологии, методологии и системы, направленные на сокращение вмешательства человека в наши повседневные жизни, развиваются быстрыми темпами и готовы разными способами изменить жизнь людей. Целью рабочей группы по экономике и гуманитарным вопросам является выявление ключевых факторов, определяющих глобальную экосистему человека и технологий, а также решение экономических и гуманитарных проблем и предоставление ключевых возможностей для решений, которые могут быть реализованы путем разблокировки критических точек напряженности. Цель рекомендаций рабочей группы заключается в том, чтобы предложить прагматическое направление, связанное с центральными проблемами в отношении людей, их объединений и возникающих информационных технологий, чтобы содействовать междисциплинарному, межсекторному диалогу, который может быть в полной мере информирован экспертным, направленным и ориентированным на современников мышлением по этим вопросам.

Вопросы:

•Неправильная интерпретация ИИ/АС в средствах массовой информации сбивает с толку общественность.

•Автоматизация обычно не рассматривается только в рыночных условиях.

•Сложность занятости игнорируется в отношении робототехники/ИИ.

•Технологические изменения происходят слишком быстро для существующих методов (пере) подготовки рабочей силы.

•Любая политика ИИ может замедлить внедрение инноваций.

•ИИ и автономные технологии имеют разную доступность в мире.

•Отсутствует доступ и понимание личной информации.

•Необходимость увеличить активное представительство развивающихся стран в глобальной инициативе IEEE.

• Появление ИИ и автономных систем может усугубить экономические и силовые структурные различия между и внутри развитых и развивающихся стран.

8 | Закон

Ранняя разработка ИИ/АС породила множество сложных этических проблем. Эти этические вопросы почти всегда прямо переходят в конкретные юридические проблемы - или они порождают сложные юридические проблемы обеспечения. Юридическая рабочая группа считает, что для юристов в этой области много работы. До сих пор было привлечено очень мало практиков и ученых, несмотря на то, что они были необходимы. Юристы должны быть частью рабочей группы обсуждений по регулированию, управлению, внутреннему и международному законодательству в этих областях, так как огромные выгоды, доступные человечеству и нашей планете от ИИ/АС направлены на будущее.

Вопросы:

• Как мы можем улучшить подотчетность и проверяемость автономных и интеллектуальных систем?

• Как мы можем обеспечить прозрачность ИИ и уважение индивидуальных прав?

Например, международные, национальные и местные органы власти используют ИИ, затрагивающие права граждан, которые должны быть в состоянии доверять правительству и, следовательно, ИИ, для защиты своих интересов.

• Каким образом системы ИИ могут быть разработаны, чтобы гарантировать юридическую ответственность в случае причинения этими системами вреда?

• Как можно создавать и развертывать автономные и интеллектуальные системы таким образом, чтобы обеспечить целостность персональных данных?

Наши новые рабочие группы и их текущая работа описаны в конце рекомендаций для этически обоснованного проектирования.