Конспект лекции

С.С. Жумажанова, П.С. Ложников

Этика и правовые проблемы искусственного интеллекта

Заключение

Заключение

**1 Слабый и сильный ИИ**

**Слабый ИИ (Narrow AI, ИИ узкого назначения)**

*Особенности:*

* Запрограммирован на выполнение одной задачи (мониторинг погоды, игра в шахматы, анализ данных, сбор кубика Рубика и др.)
* Может работать в режиме реального времени, но извлекает информацию лишь из ограниченного набора данных
* Не обладает человеческими чувствами и сознанием, а работает лишь в заранее заданном диапазоне
* Это ИИ узкого назначения, потому что они слишком далеки от человеческого интеллекта — не могут думать самостоятельно.
* Примеры слабого ИИ: Google Assistant, Google Translate, Siri и другие инструменты обработки естественного языка.

*Преимущества:*

* Способны обрабатывать данные и выполнять задачи значительно быстрее, чем человек (в частности, для помощи врачам — они делают медобслуживание лучше, быстрее и безопаснее).
* Слабый ИИ избавит общество от мелкой работы (например, заказ пиццы) и неприятных ситуаций (пробки) и предоставит больше свободного времени.
* Слабый служит строительным материалом для более фундаментальных систем, с которыми общество можем столкнуться в ближайшем будущем.

**Сильный ИИ (General AI, ИИ общего назначения)**

Особенности:

* Сильный ИИ, схож с человеческим интеллектом.
* Может успешно выполнять любые умственные задачи, которые под силу людям.
* Работу этих когнитивных механизмов сложнее всего понять, а значит, и сложнее всего воспроизвести.

*Преимущества:*

* Сильный ИИ сможет рассуждать, справляться с проблемами, выносить суждения в условиях неопределенности, планировать, учиться, интегрировать предыдущие знания в процесс принятия решений, а также предлагать новаторские идеи.
* Для достижения всех этих целей исследователи должны придумать, как наделить машины сознанием.

Значимые концептуальные теории по ИИ были разработаны Аланом Тьюрингом и Джоном Серлем. Джон Серл ввел термины «слабый» и «сильный» ИИ.

Тьюринг предложил тест по аналогии с игрой «Imitationgame» — имитационная игра. Человек взаимодействует с одним компьютером и одним человеком. На основании ответов на вопросы он должен определить, с кем он разговаривает: с человеком или компьютерной программой. Если судья не может сказать определенно, кто из собеседников является человеком, то считается, что машина прошла тест. Переписка должна производиться через контролируемые промежутки времени, чтобы судья не мог делать заключения, исходя из скорости ответов. Во времена Тьюринга компьютеры реагировали медленнее человека. Сейчас это правило необходимо, потому что они реагируют гораздо быстрее, чем человек. Задача программы – ввести человека в заблуждение, заставив сделать неверный выбор.

Премия Лёбнера (англ. Loebner prize) - премия, присуждаемая победителю ежегодного конкурса «AI Loebner» (проводится с 1990 г.), в котором соревнуются программы в прохождении теста Тьюринга. Самой «человечной» программе вручается премия в $2000. Если компьютер сможет обмануть как минимум 30 % собеседников, тест считается пройденным.

Аргументированную критику подхода Тьюринга выдвинул американский философ Джон Серл. Его примеры основаны на программе, разработанной Роджером Шенком в 1977 году. Цель программы — имитация человеческого умения понимать рассказы.

Для доказательства отсутствия Понимания у программы, Серл предложил концепцию «китайской комнаты». Истории рассказываются не на английском языке, а на китайском, а все команды для компьютерного алгоритма представлены набором инструкций (на английском языке) для работы с карточками, на которые нанесены китайские символы. Проводя мысленный эксперимент, Серл представлял, что он выполняет все манипуляции внутри запертой комнаты. Истории описываются последовательностью символов, и вопросы (на китайском) к ним подаются в комнату через небольшие прорези. Никакой другой информации извне не допускается. Когда все действия выполнены, последовательность, содержащая ответ, выдается из той же прорези наружу.

Все эти операции повторяют составляющие выполнения алгоритма по программе Шенка. При этом Серл не знает ни слова по-китайски, и поэтому не имеет ни малейшего представления о содержании рассказанных историй. Тем не менее, выполнив ряд действий, составляющих алгоритм Шенка (инструкции к которому были даны ему на английском языке), он справился бы с задачей не хуже китайца, способного без труда понять эти истории. Довод Серла заключается в том, что простое выполнение подходящего алгоритма не говорит о понимании.

**2 Перспективы и направления развития ИИ в России и мире**

**Нейронные сети.** Продолжается совершенствование алгоритмов обучения и классификации в масштабе реального времени, обработки естественных языков, распознавания изображений, речи, сигналов, а также создание моделей интеллектуального интерфейса, подстраивающегося под пользователя. Среди основных прикладных задач, решаемых с помощью нейронных сетей, - финансовое прогнозирование, раскопка данных, диагностика систем, контроль за деятельностью сетей, шифрование данных. В последние годы идет усиленный поиск эффективных методов синхронизации работы нейронных сетей на параллельных устройствах.

**Эволюционные вычисления.** ЭВ затрагивают практические проблемы самосборки, самоконфигурирования и самовосстановления систем, состоящих из множества одновременно функционирующих узлов. При этом удается применять научные достижения из области цифровых автоматов. Другой аспект ЭВ - использование для решения повседневных задач автономных агентов в качестве персональных секретарей, управляющих личными счетами, ассистентов, отбирающих нужные сведения в сетях с помощью поисковых алгоритмов третьего поколения, планировщиков работ, личных учителей, виртуальных продавцов и т. д. Сюда же относится робототехника и все связанные с ней области. Основные направления развития - выработка стандартов, открытых архитектур, интеллектуальных оболочек, языков сценариев/запросов, методологий эффективного взаимодействия программ и людей.

**Нечеткая логика.** Системы нечеткой логики активнее всего будут применяться преимущественно в гибридных управляющих системах.

**Обработка изображений.** Продолжится разработка способов представления и анализа изображений (сжатие, кодирование при передаче с использованием различных протоколов, обработка биометрических образов, снимков со спутников), независимых от устройств воспроизведения, оптимизации цветового представления на экране и при выводе на печать, распределенных методов получения изображений.

Дальнейшие развитие получат средства поиска, индексирования и анализа смысла изображений, согласования содержимого справочных каталогов при автоматической каталогизации, организации защиты от копирования, а также машинное зрение, алгоритмы распознавания и классификации образов.

* Автоматический анализ естественных языков (лексический, морфологический, терминологический, выявление незнакомых слов, распознавание национальных языков, перевод, коррекция ошибок, эффективное использование словарей).
* Высокопроизводительный OLAP-анализ и раскопка данных, способы визуального задания запросов.
* Медицинские системы, консультирующие врачей в экстренных ситуациях, роботы-манипуляторы для выполнения точных действий в ходе хирургических операций.
* Создание полностью автоматизированных киберзаводов, гибкие экономные производства, быстрое прототипирование, планирование работ, синхронизация цепочек снабжения, авторизации финансовых транзакций путем анализа профилей пользователей.

Небольшое число конференций посвящено выработке прикладных методов, направленных на решение конкретных задач промышленности в области финансов, медицины и математики. Традиционно высок интерес к ИИ в среде разработчиков игр и развлекательных программ (это отдельная тема). Среди новых направлений их исследований - моделирование социального поведения, общения, человеческих эмоций, творчества.

**Военные технологии.** Существует ряд нерешенных научных проблем, не позволяющих в ближайшие десятилетия разработать в этой сфере самостоятельно действующие армии самоходных машин-роботов и беспилотных самолетов. Прежде всего это недостатки систем автоматического распознавания, не способных правильно анализировать видеоинформацию в масштабе реального времени.

**3 Итоги курса**

В ходе изучения данной программы студенты были ознакомлены со следующими темами*:*

* Национальный кодекс этики в сфере искусственного интеллекта
* Принципы ИИ
* Вопросы социальной ответственности, связанные с участием и развитием общества при внедрении систем ИИ
* Создание и использование этичного и социально приемлемого ИИ в современном мире
* Технологии ИИ в задачах защиты информации
* Правовые аспекты участия человека в естественно-научных (биомедицинских) экспериментах для обучения ИИ
* Методы оценки рисков, возникающих из-за внедрения ИИ, а также передовые методы обнаружения и снижения воздействия ИИ
* Вопросы доверия к ИИ

**Перспективы и направления развития этического ИИ в России и мире:**

* регулирование использования искусственного интеллекта в условиях пандемии;
* новые вызовы и глобальные тренды развития информационного общества и систем на базе ИИ;
* прогноз роста рынка ИИ в ближайшие годы;
* новые специфические сферы развития ИИ;
* перспективы развития российского законодательства в области искусственного интеллекта.

**Список источников**

1. Searle, J. Chinese room argument: [англ.]. — The mit encyclopedia of the cognitive sciences. — MIT press, 2001. — с.115-116. — ISBN 0262731444.
2. Перспективы и тенденции развития искусственного интеллекта [electronic resource]. URL: <https://www.computer-museum.ru/frgnhist/aireview.htm>.
3. Рассел С., Норвиг П. Искусственный интеллект: современный подход = Artificial intelligence: a modern approach / пер. К. Птицын. — 2-е изд. — Вильямс, 2005. — 1408 p. — 2000 экз. — ISBN 978-5-8459-0887-2.