Smart Reading

САММАРИ
& ЛУЧШИЕ ИДЕИ ИЗ КНИГ

ДЖАСТИН ХАТЧИНС

ЯЗЫК ОБМАНА

The Language of Deception: Weaponizing Next Generation AI
Justin Hutchens

Smart Reading

Язык обмана. Как ИИ нового поколения становится оружием. Джастин Хатчинс. Саммари

Серия «Smart Reading. Ценные идеи из лучших книг. Саммари»
Серия «Впервые на русском (Smart Reading)»

http://www.litres.ru/pages/biblio_book/?art=73903982
Язык обмана. Как ИИ нового поколения становится оружием. Джастин Хатчинс. Саммари:

Аннотация

Это саммари – сокращенная версия книги «Язык обмана. Как ИИ нового поколения становится оружием» Джастина Хатчинса. Только самые ценные мысли, идеи, кейсы, примеры.

Люди создают машины, учат их думать и в конце концов становятся их жертвами. Это не сценарий банального фантастического фильма, а точное описание того, что происходит

с человечеством сегодня. Эксперт по кибербезопасности и поведению людей в цифровой среде Джастин Хатч Хатчинс написал свою книгу для всех, кто хочет понимать, как возник и развивался ИИ и куда он пойдет в ближайшем будущем. Изложенные в этом саммари факты, наблюдения и принципы полезны каждому, кто хочет использовать ИИ с умом и не отдавать контроль силе, которая прекрасно умеет манипулировать, но совершенно не способна сострадать.

Содержание

Smart Reading
Язык обмана. Как ИИ нового поколения становится оружием. Джастин Хатчинс. Саммари

Оригинальное название:

The Language of Deception: Weaponizing Next Generation AI

Автор:

Justin Hutchens

На грани доверия: человек и его умные машины

Чем дальше, тем больше мы доверяем умным помощникам – от поисковых и рекомендательных алгоритмов до устройств, установленных в наших домах, и инструментов для решения рабочих задач.

Мы отмахиваемся от паникеров, опасающихся восстания машин, и увлеченно общаемся с ботами обо всем на свете – от бытовых вопросов до рассуждений о смысле жизни… Но всегда ли мы понимаем, что растущее доверие к машинам без должного контроля действительно может обернуться – и уже оборачивается – против нас?

Эксперт в вопросах кибербезопасности и социальной психологии Джастин Хатчинс предлагает читателю не просто обзор современных технологий, а комплексный взгляд на логику развития ИИ и структурированное предупреждение о рисках, которые нужно учитывать уже сегодня.

Эта книга – обязательное чтение для всех, кто интересуется ИИ, кибербезопасностью, философией сознания и будущим цифрового общества. Хатчинс объясняет, как работает социальная инженерия в исполнении ИИ, какие уязвимости существуют в архитектуре моделей-трансформеров, почему тест Тьюринга нерелевантен для современных умных

помощников и что люди могут и должны сделать в сфере контроля над ИИ, пока не стало слишком поздно.

Читайте это саммари, чтобы отличать реальность от фантастики даже там, где они практически слились.

Как языковые модели изменили нашу жизнь

Появление моделей, основанных на технологиях обработки естественного языка (большие языковые модели, Large Language Model – LLM), радикально изменило общение людей с окружающим миром.

Мы получили возможность не только мгновенно получать ответы на сложные вопросы и создавать программы, не зная, как писать код, но и вступать в диалог с машиной, а также использовать ИИ для общения с другими людьми.

> *Когда появился Chat GPT, пользователи Tinder начали активно применять его для общения с потенциальными партнерами.*

Эволюция языковых моделей

Системы NLP (Natural Language Processing), разработка которых началась еще в 1950-х годах, можно разделить на два типа:

1) первые работают по заранее заданным параметрам и созданным человеком правилам;

2) вторые – на основе машинного обучения, самостоятель-

но формируя правила из полученных данных.

Ранние системы обработки языка использовали условную логику: если встречается определенное слово, система дает заранее подготовленный ответ. Такие системы было сложно масштабировать и учить отвечать на сложные запросы.

> *В диагностической системе для выявления инфекций крови MYCIN, созданной в 1970-х годах, использовалось 500 логических правил. Она была успешна, но требовала очень больших вычислительных мощностей, что затрудняло ее широкое внедрение.*

С конца 1990-х годов разработчики начали фокусироваться на создании узкоспециализированных чат-ботов, которые могли поддерживать диалог в пределах конкретной темы. Для улучшения взаимодействия стали использовать сопоставление шаблонов: вместо прописывания всех возможных фраз создавались универсальные конструкции, позволяющие распознавать общие смыслы в репликах пользователей.

В 2000-е годы активно развивался анализ настроений : системы учились интерпретировать эмоции пользователя, например использовать капслок или ругательства как индикаторы гнева. Это должно было сделать взаимодействие более «человечным», но часто приводило к неестественным и фальшивым ответам. Параллельно добавлялись антропоморфные элементы: чат-ботам давали имена, их обучали ре-

агировать на эмоциональные фразы и выстраивать ответы с намеком на сочувствие или юмор. Все это усиливало иллюзию общения с живым собеседником.

В первом десятилетии XXI века началось внедрение более продвинутых методов предобработки текста, среди них – автоматическое исправление орфографических ошибок, нормализация диалектных форм и стемминг (выделение основы слова для облегчения анализа и сопоставления смыслов). Эти приемы позволяли системам лучше понимать разнообразные пользовательские формулировки. Если запрос оказывался непонятным, система переходила к заранее заготовленным универсальным ответам. ***Появились и первые попытки использовать память: краткосрочную – для учета контекста текущего диалога и долгосрочную – для запоминания предпочтений пользователя между сессиями.*** Это создавало ощущение более индивидуального (и человеческого) общения.

Наконец, в 2010-х появились статистические языковые модели (SLM), которые обучались на больших объемах текстов и предсказывали наиболее вероятное продолжение фразы на основе частотности словосочетаний. А затем – и нейросетевые модели, такие как LSTM (Long Short-Term Memory), которые позволили учитывать более длинный контекст и последовательно обрабатывать текст с сохранением связи между частями фразы.

Статистические и нейросетевые языковые модели актив-

но применяются в:

- распознавании речи – для повышения точности за счет учета контекста;
- машинном переводе – для анализа вероятностей соответствия слов в разных языках;
- предсказании текста – в пользовательских интерфейсах, когда система помогает продолжить предложение;
- автоматической вычитке – при анализе и корректировке больших объемов текста.

В 2014 году компания Amazon выпустила голосового ассистента Alexa вместе с первым устройством Amazon Echo. В 2016-м на конференции Google I/O был представлен Google Assistant, который сначала был встроен в мессенджер Allo и устройство Google Home, затем появился на смартфонах Pixel, а позже – на большинстве Android-устройств. Обе системы использовали достижения в области NLP[1]

[1] Раздел машинного обучения на стыке ИИ и лингвистики, посвященный распознаванию, обработке и генерации устной и письменной речи.

Конец ознакомительного фрагмента.