**Коварные ловушки нейросетей: на что нужно обращать внимание.**

[Ссылка](https://player.vimeo.com/video/934688171) на исходное видео с 08:03 - 14:05

**Рассказчик:** На прошлом занятии мы узнали, что у нейросетей бывают галлюцинации и рассмотрели их виды. Попробуйте самостоятельно создать промпт и спросить у Гигачата, какие существуют виды галлюцинаций нейросетей.

**Пауза на практику.**

Пример возможного промпта может быть таким:

Расскажи, какие галлюцинации нейросетей существуют.

Результат:

Галлюцинация недообученности. Проявляется в виде ошибок, когда модель неверно интерпретирует данные. Например, если показать модели изображение слона, она может назвать его жирафом

Галлюцинация переобученности. Возникает, когда модель начинает видеть несуществующие закономерности. Если показывать ей котов, она начнёт видеть их даже там, где их нет.

Галлюцинация противоречивости. Когда модель обучается на данных, часть из которых не соответствует действительности, она начинает путаться. Например, если обучать модель на средневековых гравюрах, где Земля плоская, она будет считать, что Земля плоская.

Галлюцинация многозначности. Возникает, когда модель видит разные образы в одном и том же объекте. Например, в кляксе она может увидеть то лицо человека, то морду собаки.

Про недообученность и переобученность мы узнали на прошлом занятии,

Давайте подробнее расспросим Гигачат о других галлюцинациях.

Промпт: Приведи пример противоречивости

Результат:

Ситуация, когда алгоритмы машинного обучения дали противоположные переводы одного и того же предложения с китайского языка.

Промпт: Приведи пример Галлюцинация многозначности нейросети

Результат:

Случай, когда алгоритм машинного обучения принял фотографию облака за собаку.

Пауза

**Нейросеть:** Никакого спасения от галлюцинаций, остается только смириться.

**Гигачат:** Не стоит драматизировать.

Да, совсем от них не избавишься, но вы с Иваном уже умеете грамотно составлять промпты, а поэтому риски минимальны.

**Нейросеть:** Обязательно сообщу Ивану, что грамотный промпт – ключ к успеху,  но данные всегда надо проверять.

**Рассказчик:** Возможные ошибки и искажения фактов — не единственная причина, по которой важно быть внимательным при генерации контента.

Важно помнить о юридических аспектах работы с нейросетями.

**Ответственность за решения и действия.** Определение ответственности за ошибки или ущерб, вызванные решениями или действиями нейросетей, является сложной задачей.

Например, в цифровой медицинской диагностике или конструировании беспилотных машин многие решения принимает искусственный интеллект.

Но кто будет виноват, если произойдет ошибка?

Единого мнения на этот счет пока не существует.

**Конфиденциальность и защита данных.**

При использовании нейросетей важно помнить о том, что пользователи в ответе за конфиденциальность данных, которые они предоставляют нейросети для обработки и анализа.

Не загружайте в нейросети персональные данные других людей и любые конфиденциальные документы: их будет невозможно удалить или скрыть.

**Прозрачность и интерпретируемость.**

Принципы работы алгоритмов, составляющих основу нейросетей, должны быть прозрачными и обоснованными.

Это создает сложности для работы некоторых нейросетей, потому что их алгоритмы и результаты скрыты и не могут быть проанализированы.

**Биометрические данные.**

Использование нейросетей для распознавания лиц, голоса и других биометрических данных поднимает вопросы о праве на приватность  и необходимости получения согласия на обработку таких данных.

Отдельным пунктом, который нужно учитывать, являются **авторские права**, например, на текст и изображения.

Конечно, можно сгенерировать картинку и поставить ее себе в качестве аватара в социальной сети.

Но если наш Иван, например, захочет начать продавать футболки с принтами, сгенерированными нейросетью, то для начала ему необходимо внимательно изучить правила использования нейросети, чтобы быть уверенным в том, что он не нарушает закон и имеет на это право.

В будущем может появиться больше законов, которые помогут регулировать правообладание в сфере контента, сгенерированного искусственным интеллектом,  а пока стоит быть внимательным при использовании ответов нейросетей.

Интересно заметить, что иногда авторское право защищается самой нейросетью.

Так, например, если попросить Гигачат сгенерировать изображение персонажа российского мультфильма, защищенного авторским правом, то он, скорее всего,  не сможет этого сделать.

**Рассказчик:** Давайте убедимся в этом. Попробуем нарисовать персонажа Копатыча из мультфильма «Смешарики».

Промпт: Нарисуй Копатыча

Видно, что нейросеть понимает, что речь идет о медведе, но результат выдает некорректный. Может быть, стоит улучшить промпт?

Промпт: Нарисуй персонажа «Смешариков» медведя Копатыча

Не сработало. Попробуем написать еще более детальный промпт.

Промпт: Нарисуй персонажа: медведя Копатыча. Нарисуй его в том же стиле,  в котором он показан в российском мультфильме «Смешарики».

Видно, что попросив нейросеть скопировать стиль Смешариков, мы получили что-то наиболее приближенное к образу персонажа, но результат все равно нельзя считать удовлетворительным.

Попробуйте самостоятельно сгенерировать известного персонажа из любимого мультфильма.

## Сцена 8

**Рассказчик:** Кроме ограничений, касающихся авторского права, есть правила, запрещающие создавать или использовать контент, в котором присутствуют:

* Насилие;
* Расизм;
* Ненависть к животным;
* Запрещенные вещества;
* Оскорбления или угрозы;
* Пропаганда насилия или экстремизма.

Подобный контент неприемлем, а данные ограничения делают работу с нейросетями безопасной и приятной.

Попробуйте самостоятельно сгенерировать промпт, где спрашиваете у Гигачата, какой контент он может выдать.

**Пауза на практику.**

Возможный вариант промпта вы выдите на экране:

**Нейросеть**: Получается, мы и ошибаться можем, и возможности наши  не безграничны. Мне кажется, этим мы похожи на людей.

**Гигачат:** Мы в чем-то похожи, это правда. Люди тоже могут совершать ошибки. Важно научиться их исправлять.

А теперь предлагаю познакомиться с другими нейросетями.

**Нейросетка**: Я бы хотела узнать про них как можно больше! И Ване рассказать!

**Гигачат:** Тогда пора отправиться в путь!

**Рассказчик:** Сегодня на занятии вы:

Узнали о видах галлюцинаций нейросетей.

Узнали, что сгенерированный контент может быть защищен авторским правом.

Узнали о том, какой контент не может генерировать нейросеть.

Дальше ещё интереснее. До встречи на следующих занятиях!