АКЦИЯ

в рамках комплексной программы формирования у школьников основ ЦИФРОВОЙ И ФИНАНСОВОЙ БЕЗОПАСНОСТИ

Цифровой Я

Ловушки искусственного интеллекта

На этом уроке

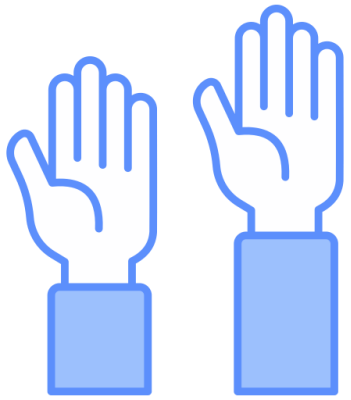
ВЫ УЗНАЕТЕ:

Какую роль играет искусственный интеллект в жизни общества
Какие ошибки может совершать ИИ
В чём слабые места ИИ

ВЫ СМОЖЕТЕ НАУЧИТЬСЯ:

Оценивать результаты работы ИИ
Понимать риски при использовании ИИ
Определять необходимость работы с ИИ

ИИ в нашей жизни



Поднимите руку, кто из вас хотя бы раз за последнюю неделю:

- ✓ Использовал переводчик для домашнего задания?
- ✓ Искал информацию в интернете с помощью умного помощника (Алиса, Siri)?
- ✓ Видел рекомендации от Яндекс, RuTube или ВК?
- ✓ Или просто создавал забавный образ с помощью нейросети для сторис?

Мы все уже активные пользователи технологий на основе ИИ. И это прекрасно! Ведь искусственный интеллект — это не что-то из далекого будущего. Он уже здесь, и он дает нам колоссальные преимущества:

Скорость

За секунды обрабатывает объемы данных, неподвластные человеку за всю жизнь.

Доступность

Многие базовые ИИ бесплатны для пользователей. Главное иметь доступ к сети Интернет.

Эффективность

Оптимизирует движение транспорта, предсказывает погоду и помогает ученым делать открытия.

ИИ в нашей жизни

ПАРАДОКС

Та самая сила, которая делает ИИ таким полезным, таит в себе и серьезные опасности.

Сегодня мы с вами не будем просто восхищаться технологиями или, наоборот, панически их бояться. Наша задача — взглянуть на ситуацию трезво и критически.

Искусственный интеллект — это всего лишь инструмент. Очень мощный. И как от любого инструмента — от молотка до ядерной энергии — последствия его использования зависят от нас, людей.

Цель сегодняшнего урока — не напугать вас, а вооружить. Знанием, критическим мышлением, пониманием. Чтобы вы, могли использовать его во благо и избежать потенциальных ловушек.

Негативные проявления ИИ

ФОРАБАНК

Цифровой Я

Какие самые большие опасности, связанные с ИИ, приходили вам в голову до этого урока?

Негативные проявления ИИ

ФОРАБАНК

Цифровой Я

Дезинформация

Фантазия ИИ

Непрозрачность принятия решений

Отсутствие ответственности

Создание вредоносного ПО

Использование для кибератак и мошенничества

Тёмная сторона ИИ: Когда алгоритмы ошибаются и обманывают

Дезинформация и создание фейков

ФОРАБАНК

Цифровой Я

ИИ стал не только мощным инструментом для создания контента, но и оружием для генерации дезинформации.

Обучение на данных	ИИ учится на огромных массивах информации из интернета (тексты, изображения, видео).
Статистика, а не правда	Его цель — предсказать «правдоподобное» слово или пиксель, а не установить истину.
Нет понимания	У ИИ нет сознания, здравого смысла и морали. Он не отличает факт от вымысла.

Дезинформация и создание фейков

ФОРАБАНК

Цифровой Я

Главные проблемы

	Галлюцинации	Генерация фейков	Усиление предвзятости
Что это?	ИИ уверенно выдаёт ложную информацию как факт.	Создание реалистичных фото, видео и аудио, которые неотличимы от настоящих.	ИИ воспроизводит и усиливает стереотипы, найденные в данных для обучения.
Пример	Может придумать несуществующие события, научные «факты» или цитаты.	Поддельное видео с публичной фигурой, где человек говорит то, чего не говорил на самом деле.	Генерация новостных статей, разжигающих межнациональную или социальную рознь.
Почему это опасно?	Подрывает доверие, распространяет ложные знания.	Может спровоцировать панику, повлиять на выборы, разрушить репутацию.	Углубляет раскол в обществе, пропагандирует ненависть.

Дезинформация и создание фейков

ФОРАБАНК

Цифровой Я

Масштабирование обмана

- ***Скорость:*** Генерирует тысячи постов, новостей и комментариев за минуты.
- **Дешевизна:** Создание сложного контента больше не требует больших ресурсов.
- ***Персонализация:*** Может создавать уникальные варианты лжи для разных аудиторий.
- ***Правдоподобие:*** Материалы стали убедительными, без явных ошибок.

Должны ли государства регулировать внедрение ИИ, зная об их потенциальной опасности?

Тёмная сторона ИИ: Оружие хакеров и собственные уязвимости

ИИ как инструмент киберпреступлений

ФОРАБАНК

Цифровой Я

ИИ — это всего лишь инструмент. Его использование в киберпространстве зависит от человека. Одна и та же технология может быть и самым мощным оружием атаки, и самым надежным щитом.

Уязвимости самого ИИ

Враждебные примеры

Микровмешательства во входные данные, которые невидимы для человека, но заставляют ИИ ошибаться. Например, пиксельный шум на изображении, из-за которого система распознавания лиц видит другого человека.

Промт-инъекции

Злоумышленники используют вредоносный код, чтобы обманом заставить ИИ-модели выдать конфиденциальную информацию, разрешить несанкционированный доступ или удалить файлы.

Кража данных

Атака, направленная на раскрытие конфиденциальных данных, на которых обучалась модель (персональные данные, коммерческая тайна).

ИИ как инструмент киберпреступлений

ФОРАБАНК

Цифровой Я

Как ИИ используется для кибератак и вредоносного ПО

Умный фишинг и социальная инженерия

Генерация убедительных писем и голосовых сообщений: ИИ анализирует стиль жертвы в соцсетях и создает идеальную приманку.

Создание и адаптация вредоносного ПО

Автоматизация кода: ИИ генерирует новые варианты вирусов, которые сложнее обнаружить стандартным антивирусам.

Взлом паролей и систем

ИИ предсказывает и подбирает пароли, анализируя данные утечек и привычки пользователей. Алгоритмы сканируют миллионы строк кода, чтобы находить дыры в безопасности быстрее человека.

Автоматизация атак

Координация масштабных DDoS-атак: тысячи зараженных устройств атакуют цель одновременно, парализуя ее работу.

Можно ли доверять ИИ-защиту критической инфраструктуры (электростанций, больниц), если его самого можно обмануть с помощью хакерских атак?

Решение за тебя:
Как ИИ думает и кто отвечает за последствия?

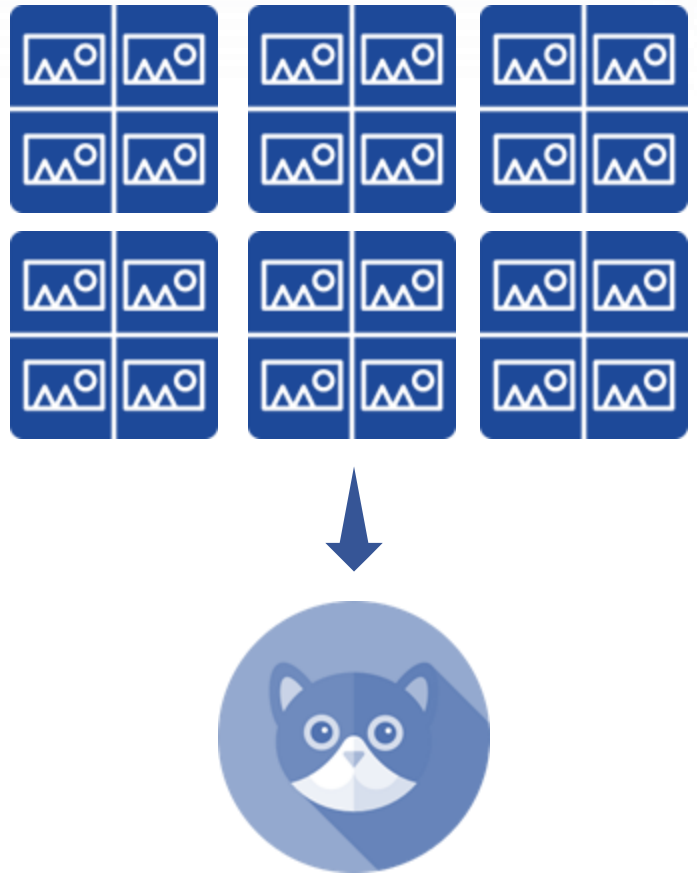
Как ИИ принимает решения

Как ИИ принимает решения?

Обучение: ИИ анализирует огромные массивы данных (например, тысячи фото котов и собак).

Выявление связей: Алгоритм находит тысячи признаков, чтобы отличать одно от другого (например, форма ушей, длина хвоста).

Прогноз: На основе выученных паттернов ИИ делает вывод для новых данных («На этой новой фотографии на 95% кот»).



Как ИИ принимает решения

ФОРАБАНК

Цифровой Я

Проблема «Черного ящика»

Многие сложные модели ИИ (например, нейросети) **непрозрачны**. Даже их создатели не могут точно сказать, почему был получен тот или иной результат. Виден только вход и выход.

Пример. Нейросети генерируют разные ответы, когда пользователи задают одни и те же вопросы.

Почему это важно???
Без понимания причины мы не можем проверить, является ли решение справедливым или ответ верным, и исправить ошибку.

Должен ли ИИ иметь право принимать решения, влияющие на жизнь людей, если его логика необъяснима?

Работа в группе

ФОРАБАНК

Цифровой Я

ТЕМА: Как грамотно использовать нейросети?

Задача:

Сформулировать 5 общих правил работы с нейросетями.

Озвучить результат групповой работы.

Сравнить с ответами других групп.

Дополнить свои правила (при необходимости).

Заключение

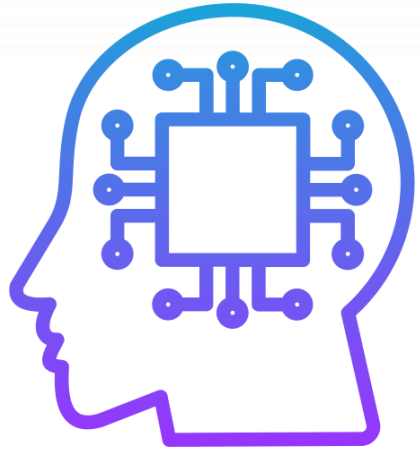
ФОРАБАНК

Цифровой Я

ИИ — это инструмент в руках человека.

Мы увидели, что у ИИ есть потрясающие возможности, но и риски тоже серьёзные: предвзятость, кибератаки, зависимость и неравенство.

Важно не бояться технологий, а понимать, как они работают и к каким последствиям могут привести.



Используйте ИИ с умом, задавайте вопросы, проверяйте информацию и не забывайте, что последнее слово должно всегда оставаться за человеком. А не за алгоритмом.

Подведём итоги

Продолжите фразы:

На уроке я узнал(а) …

На уроке я научился(ась) …

Это мне обязательно пригодится, когда я …

АКЦИЯ

в рамках комплексной программы формирования у школьников основ ЦИФРОВОЙ И ФИНАНСОВОЙ БЕЗОПАСНОСТИ

Цифровой Я