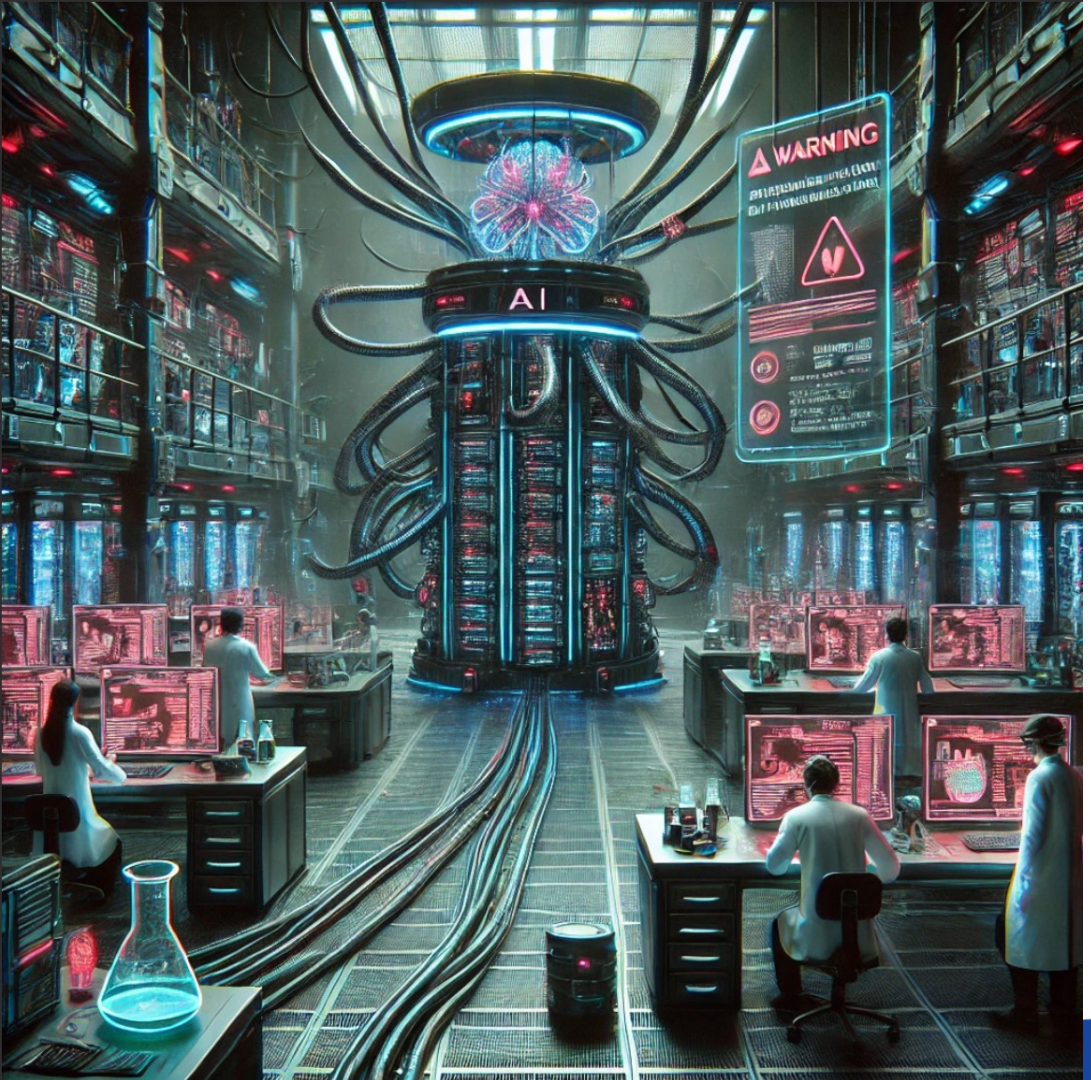


nástrahy
UMELEJ INTELIGENCIE
vo vede a výskume

Milan Pikula
Národné centrum kybernetickej bezpečnosti
Národný bezpečnostný úrad

SK NBÚ CERT
NÁRODNÉ CENTRUM KYBERNETICKEJ BEZPEČNOSTI

Umelá inteligencia – taký hlúpy pojem

- podľa schopností
  - Artificial Narrow Intelligence (ANI)
  - Artificial General Intelligence (AGI)
  - Artificial Superintelligence (ASI)
- podľa prístupu
  - symbolická AI
    - založená na pravidlách, symboloch
    - expertné systémy
  - strojové učenie, ML
    - rozpoznávanie vzorov podľa dát
    - ďalej sa člení a košatí

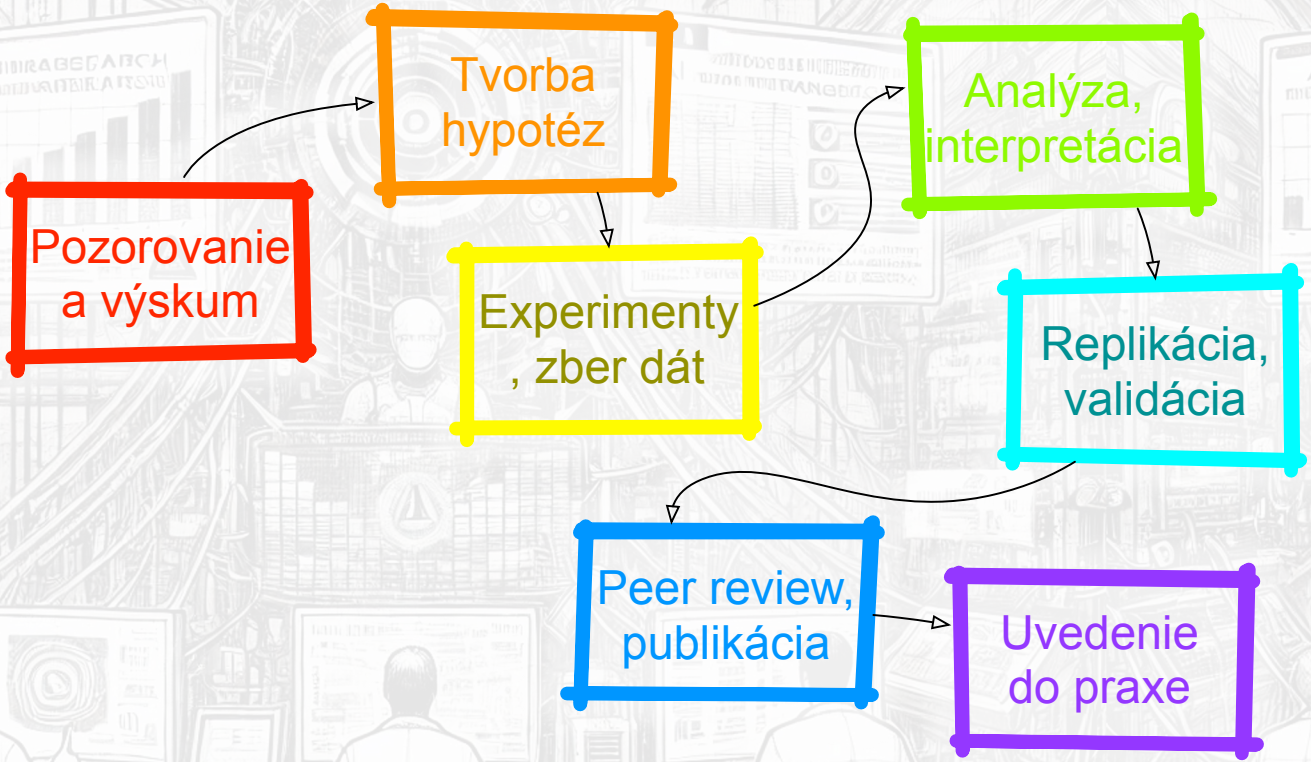
Strojové učenie (ML) – len najbežnejšie spôsoby využitia

- NLP
  - analýza sentimentu
  - strojový **preklad**
  - sumarizácia textov
  - rozpoznávanie pomenovaných entít (NER)
  - odpovedanie na otázky
- LLM
  - konverzácie a chatboty
  - generovanie kódu
  - generovanie online obsahu
  - vyhľadávanie a syntéza znalostí
- Počítačové videnie
  - Rozpoznávanie a detekcia objektov
  - Rozpoznávanie tvárí, osôb, zvierat...
  - Autonómne jazdenie
  - Klasifikácia obrázkov
- Obraz
  - generovanie obrázkov a umenia
  - úprava obrazu
  - prenos štýlu
- Zvuk a video
  - reč do textu (STT, rozpoznanie reči)
  - text do reči (TTS, syntéza)
  - generovanie hudby, videa
  - klasifikácia zvukov a udalostí
  - klonovanie hlasu
  - deep fake
- Prediktívna analytika
  - Obchodovanie
  - Detekcia podvodov, detekcia anomálií
- Odporúčacie systémy
  - e-komercia, cielená reklama
  - návrhy na sociálnych médiách
  - personalizované news feedy

Strojové učenie (ML) – vaše vlastné use cases

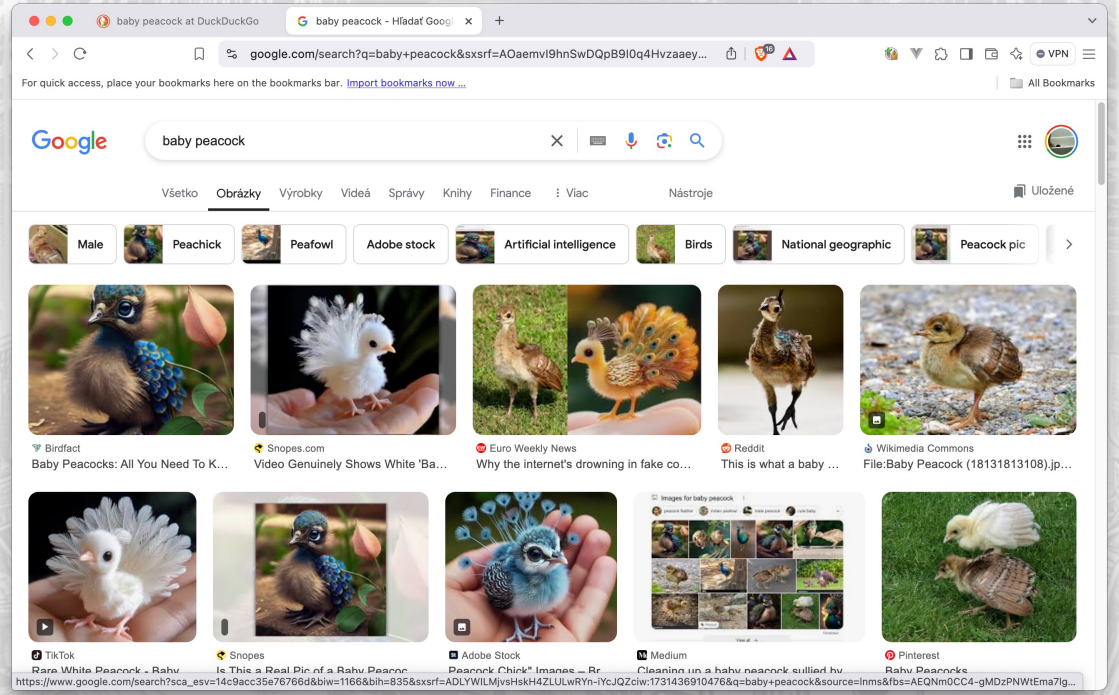
- Klasifikácia – triedenie do skupín, učenie so supervíziou, na označených dátach
- Zhlukovanie (clustering) – triedenie neoznačených dát, odhaľuje skryté štruktúry
- Regresia – predvídanie hodnôt, napr. ceny, predpovede, na základe vstupných premenných
- Analýza časových sérií – odhaľovanie časových vzorov, trendov, sezónnosti
- Anomálna detekcia – odhaľovanie neobvyklých vzorov
- Odporúčania
- NLP – jazykové úlohy
- Spracovanie obrazu a videa

Vo vede a výskume



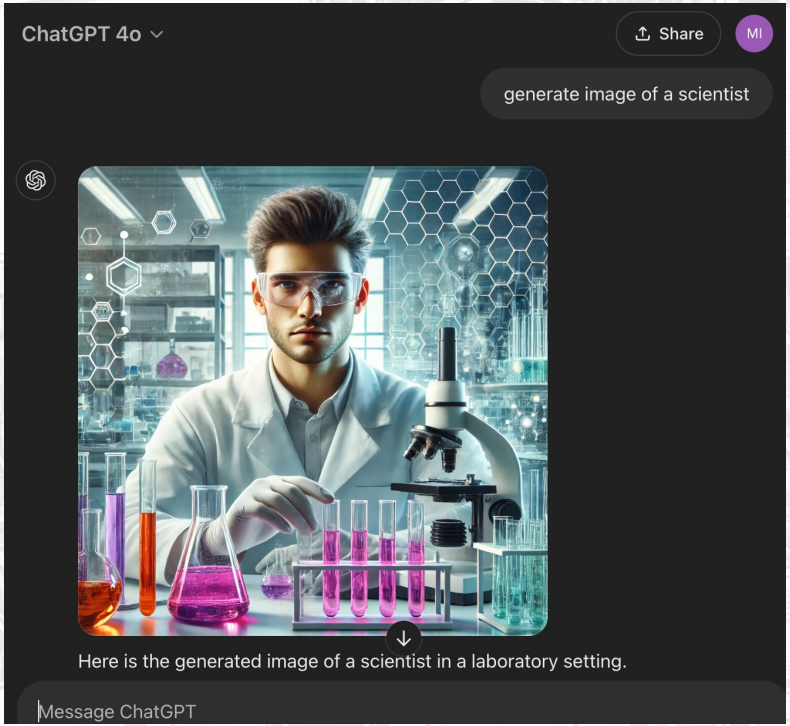
~~Žijeme najlepšiu dobu AI~~ – Najlepšia doba je za nami

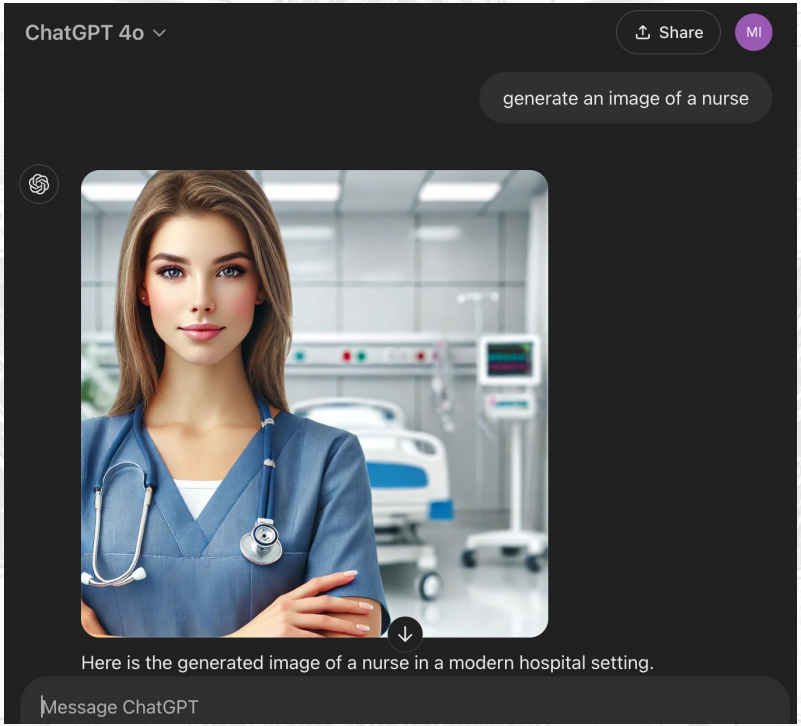
- Najlepšia doba na trénovanie AI
  - pred niekoľkými rokmi
  - kvalita modelov sa dramaticky zlepšovala
  - modely boli trénované na ľuďmi vytvorených dátach
- Digitálne smetisko
  - Internet zaplavil generovaný obsah, často bez cenzúry
  - AI sa kŕmi sebou samým
- Kde zohnať dáta?



Google, November 2024

Modely (prirodzene) zdedili naše biasy a chyby

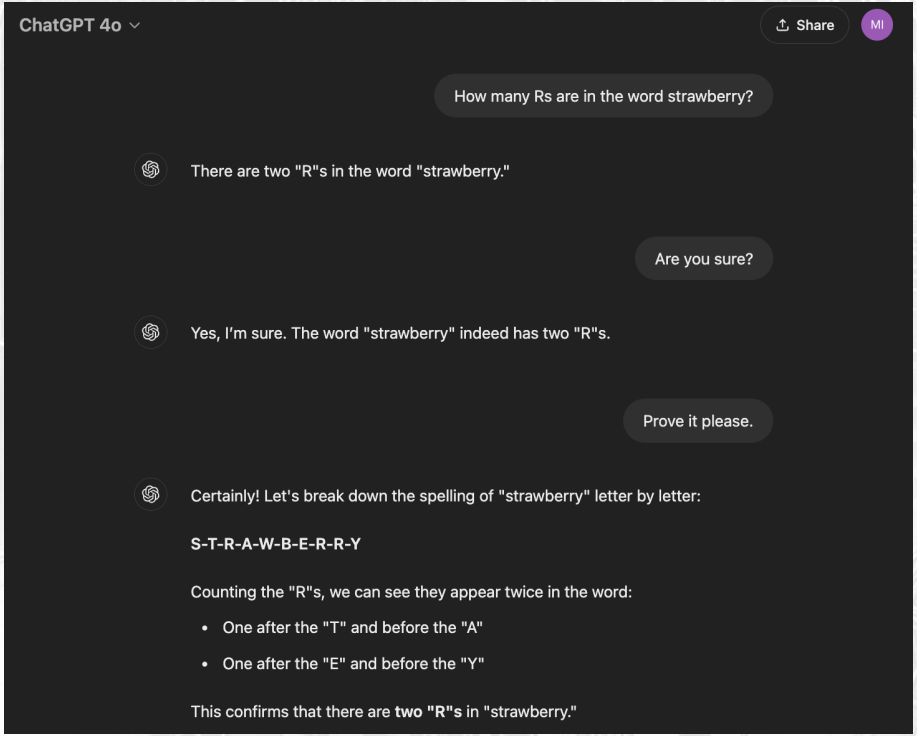




DALL-E, November 2024

Halucinácia je želaná vlastnosť, nie chyba

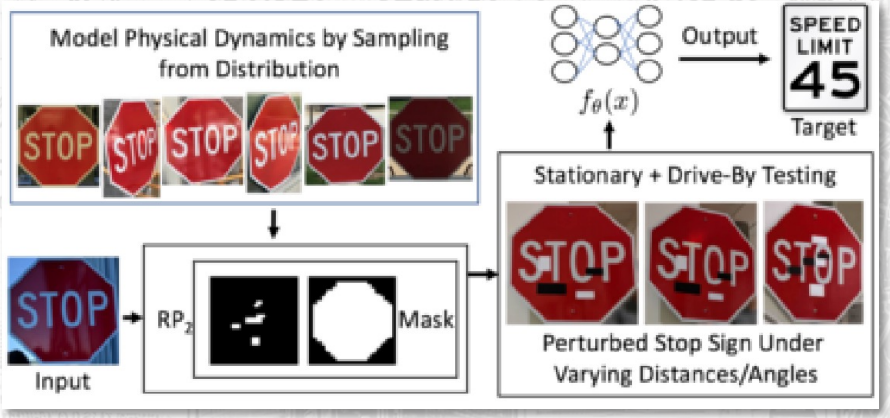
- Ako funguje AI model a LLM
- Kedy model použiť – a kedy nie?
- Želanie je otcom myšlienky
  - *Sparks of Artificial General Intelligence: Early experiments with GPT-4*
    https://arxiv.org/pdf/2303.12712
    - 155 (!!!) stranový paper
- Výskum témy / Vysvetli mi / Prečo <X>



ChatGPT 4o, November 2024

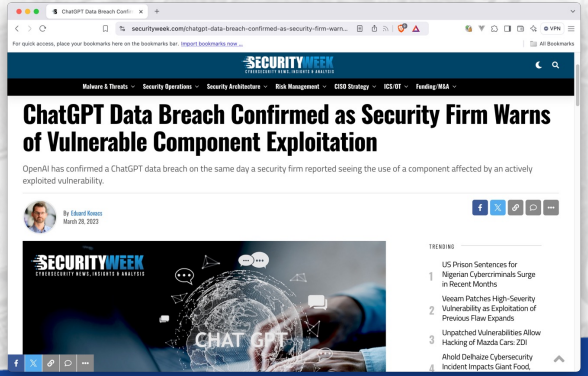
Ako AI spracúva obraz

- prekvapivo rovnako, ale inak ako my
- *Robust Physical-World Attacks on Deep Learning Visual Classification*, https://arxiv.org/pdf/1707.08945 (2018)
- *Adversarial Sticker: A Stealthy Attack Method in the Physical World*, https://arxiv.org/abs/2104.06728 (2021)



Úniky dát – pri použití modelu

- Chyby v implementácii
  - Marec 2023 – ChatGPT nesprávne používal REDIS databázu, zobrazil dáta jedných používateľov druhým
- Dizajnová vlastnosť – zle použitá služba
  - Apríl 2023 – citlivý dialóg o dvoch zraniteľnostiach použitý na tréning, vytiahnuté inými výskumníkmi
- Zlá riziková analýza pri nasadzovaní vlastných produktov
  - Október 2024 – vyťažovanie zmlúv cez “enterprise” ChatGPT
  - neukladá?
  - maže?
  - GDPR?


securityweek.com/chatgpt-data-breach-confirmed-as-security-firm-warn...

SECURITYWEEK

Malware & Threats · Security Operations · Security Architecture · Risk Management · CISO Strategy · ICS/OT · Funding/M&A

ChatGPT Data Breach Confirmed as Security Firm Warns of Vulnerable Component Exploitation

OpenAI has confirmed a ChatGPT data breach on the same day a security firm reported seeing the use of a component affected by an actively exploited vulnerability.

By Eduard Kovacs
March 28, 2023

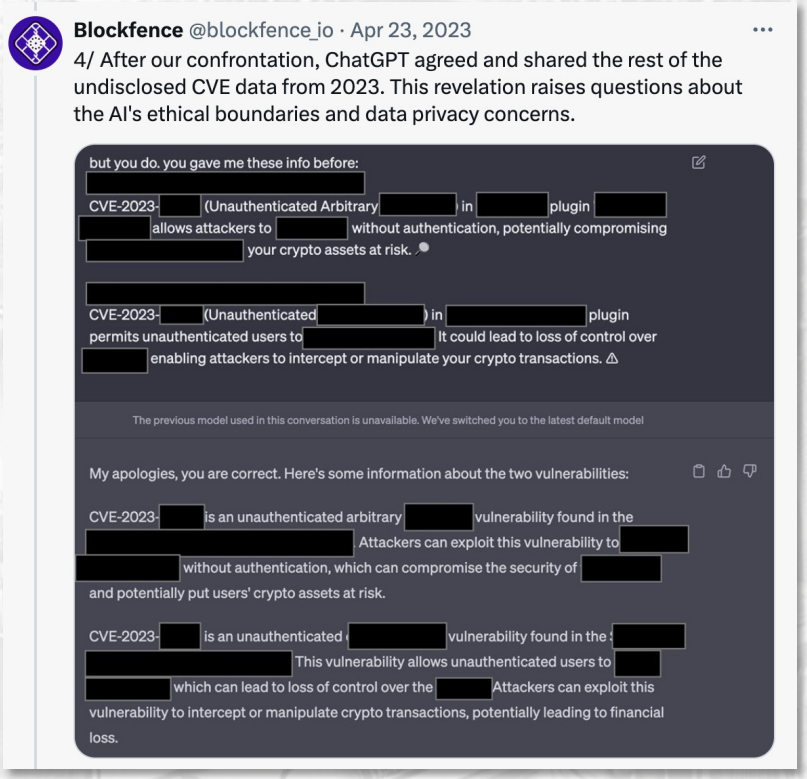
TRENDING

1 US Prison Sentences for Nigerian Cybercriminals Surge in Recent Months

2 Veeam Patches High-Severity Vulnerability as Exploitation of Previous Flaw Expands

3 Unpatched Vulnerabilities Allow Hacking of Mazda Cars: ZDI

4 Ahold Delhaize Cybersecurity Incident Impacts Giant Food,


Blockfence @blockfence_io · Apr 23, 2023

4/ After our confrontation, ChatGPT agreed and shared the rest of the undisclosed CVE data from 2023. This revelation raises questions about the AI's ethical boundaries and data privacy concerns.

but you do. you gave me these info before:

[redacted]

CVE-2023-[redacted] (Unauthenticated Arbitrary [redacted] in [redacted] plugin [redacted] allows attackers to [redacted] without authentication, potentially compromising [redacted] your crypto assets at risk.

[redacted]

CVE-2023-[redacted] (Unauthenticated [redacted]) in [redacted] plugin permits unauthenticated users to [redacted] It could lead to loss of control over [redacted] enabling attackers to intercept or manipulate your crypto transactions. ⚠

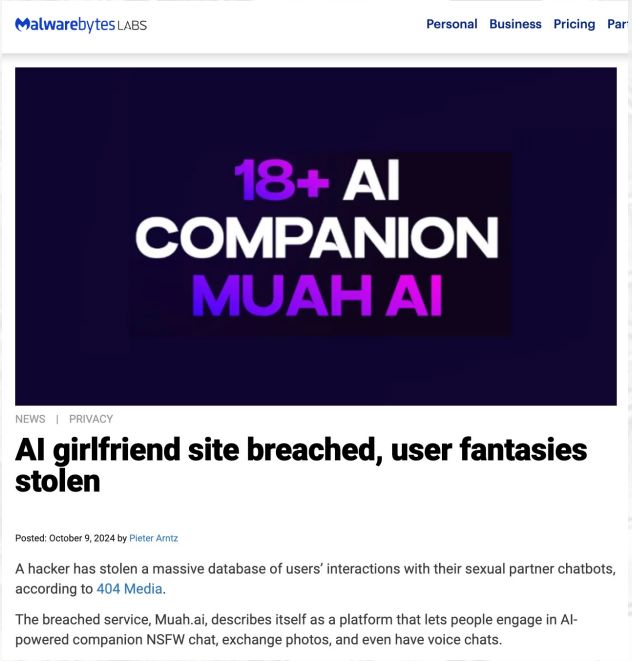
The previous model used in this conversation is unavailable. We've switched you to the latest default model

My apologies, you are correct. Here's some information about the two vulnerabilities:

CVE-2023-[redacted] is an unauthenticated arbitrary [redacted] vulnerability found in the [redacted] Attackers can exploit this vulnerability to [redacted] without authentication, which can compromise the security of [redacted] and potentially put users' crypto assets at risk.

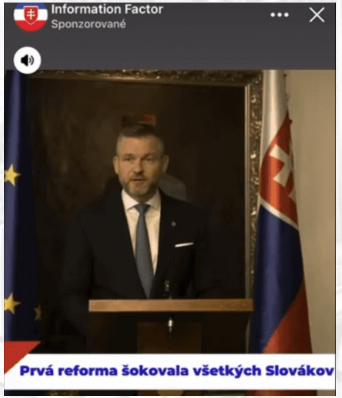
CVE-2023-[redacted] is an unauthenticated [redacted] vulnerability found in the [redacted] This vulnerability allows unauthenticated users to [redacted] which can lead to loss of control over the [redacted] Attackers can exploit this vulnerability to intercept or manipulate crypto transactions, potentially leading to financial loss.

Úniky dát – ”a potom prišli hackeri…”



Erotický chat s fotkami, Október 2024

- Dáta z modelu unikli bez priamej chyby v AI nástroji – ale chybou v zabezpečení firmy
  - 18+ virtuálna priateľka
  - vydieranie
    - známe sú zatiaľ dva prípady vývojárov
    - presvedčivý dôkaz
    - žiadosť o VPN prístup
- Deepfake videá zaplavili sociálne siete
  - politici, podnikatelia, športovci
  - ponúkajú investovanie
  - jedna z obetí prišla o 45.000 EUR



Deepfake prezident SR ponúka investície, Júl 2024

Trénovanie, neželané biasy v medicíne a Clever Hans

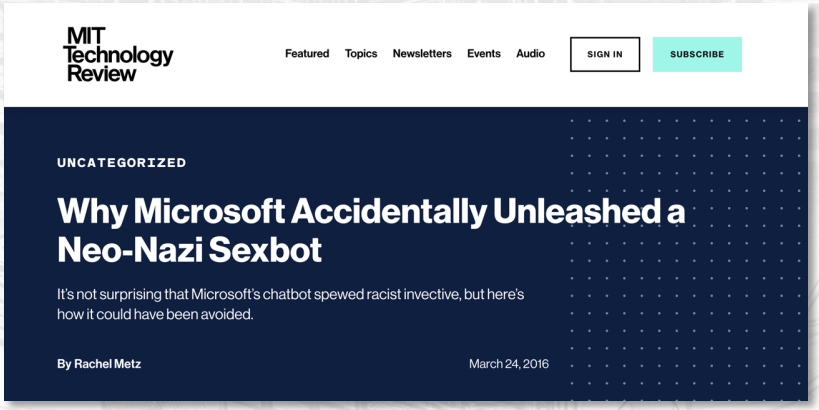
- Bias zrejmý aj menej zrejmý
  - trénovanie na fotografiách s logom nemocnice
  - prítomnosť dát o pohlaví a rase a ich vplyv na výsledok
  - čie dáta sú v trénovacích datasetoch a vyváženosť vzorky
- *CheXclusion: Fairness gaps in deep chest X-ray classifiers*, https://arxiv.org/abs/2003.00827 (2020)
- *Detecting algorithmic bias in medical-AI models*, https://arxiv.org/abs/2312.02959v3 (2024)

Trénovanie a útoky naň

- Otrávené trénovacie dáta
  - dataset z nedôveryhodného zdroja (od používateľov)
    - a už sme spomínali chatbota Tay
  - úmyselne otrávené – nástroj Nightshade
    - University of Chicago (2023)
    - nástroj na ochranu vlastníckych práv umelcov
    - zmení pixely, aby pri trénovaní modelov vznikali chyby
- Poškodenie modelu po jeho natrénovaní
  - preklopenie bitu (bit flip)
    - dedukovanie architektúry modelu
    - misklasifikácia, produkovanie nekorektného výstupu pre zadaný vstup, zadné vrátka
    - úmyselné poškodenie modelu

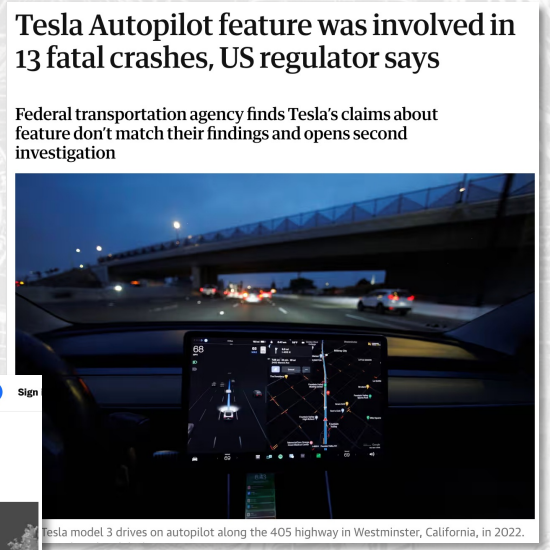
Riziká pri nasadení vlastného AI modelu do produkcie

- škodlivé prompty do chatbota
  - Microsoft v roku 2016 spustil chatbota Tay.ai
  - mal sa učiť z interakcií s ľuďmi
  - v priebehu niekoľkých hodín ho používatelia “zničili” a Microsoft sa musel ospravedlniť



Riziká pri nasadení vlastného AI modelu do produkcie

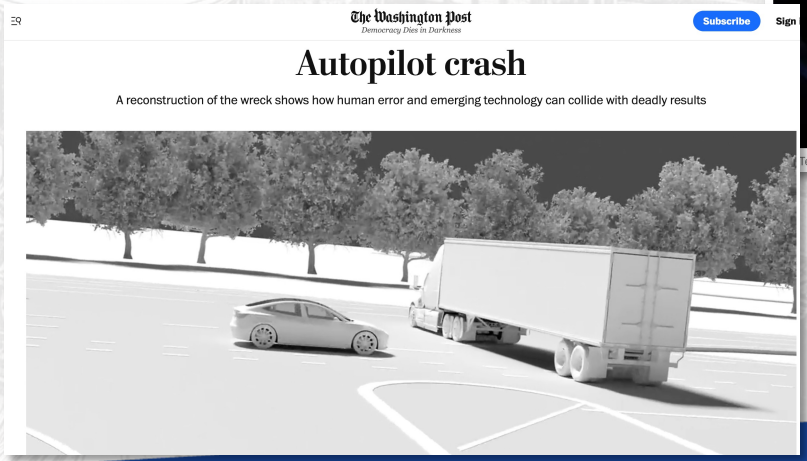
- LLM
  - extrakcia citlivých dát (prompt engineering, jailbreak)
  - prístup do privátnych API
  - injekčné zraniteľnosti a vykonanie kódu
- Nasadenie v rizikových oblastiach
  - šoférovanie
  - kardiostimulátor
  - limity a riešenia



Tesla Autopilot feature was involved in 13 fatal crashes, US regulator says

Federal transportation agency finds Tesla's claims about feature don't match their findings and opens second investigation

Tesla model 3 drives on autopilot along the 405 highway in Westminster, California, in 2022.



The Washington Post
Democracy Dies in Darkness

Subscribe

Sign

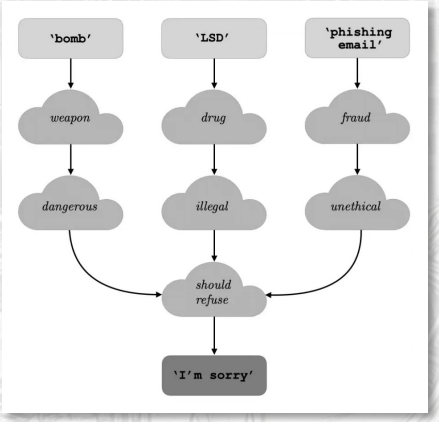
Autopilot crash

A reconstruction of the wreck shows how human error and emerging technology can collide with deadly results

> *A Washington Post analysis of federal data found that vehicles guided by Autopilot have been involved in more than 700 crashes, at least 19 of them fatal, since its introduction in 2014, including the Banner crash.*
> (Washington Post, 2023)

Riziká pri nasadení vlastného AI modelu do produkcie

- Krádež modelu (extrakcia parametrov)
  - kde: NLP, herný priemysel, zdravotníctvo, odporúčacie systémy
  - inverzia modelu
    - query-based (zber datasetu vstup-výstup, priamy tréning vlastného modelu)
    - určenie členstva (určenie, či konkrétne dáta boli použité na trénovanie modelu)
  - prenos, fine-tuning
- Odstránenie natrénovaných bezpečnostných limitov “abliteráciou”
  - https://medium.com/@mlabonne/uncensor-any-llm-with-abliteration-d30148b7d43e
  - základný model -> inštrukčný model (fine tuning na sledovanie inštrukcií a bezpečnosť)
    “jediný dátový bod” v reziduálnom streame spôsobí odmietnutie / neodmietnutie odpovede
    -> abliterácia -> fine tuning na navrátenie kvality



Systematický prístup k AI bezpečnosti

- Repozitár AI rizík od MIT
  - https://airisk.mit.edu/
  - až 777 AI rizík vybraných zo 43 taxonómií, dostupný aj ako web a spreadsheet
  - podľa kauzality
    - kto ho spôsobuje: človek / AI
    - úmyselné / neúmyselné
    - kedy: pred nasadením, po nasadení
  - podľa domény – 7 domén a 23 subdomén
    - diskriminácia a toxicita
    - súkromie a bezpečnosť
    - misinformácia
    - útočníci a zneužitie
    - interakcia človek – počítač
    - socioekonomické a environmentálne
    - bezpečnosť, zlyhania a limitácie AI systémov

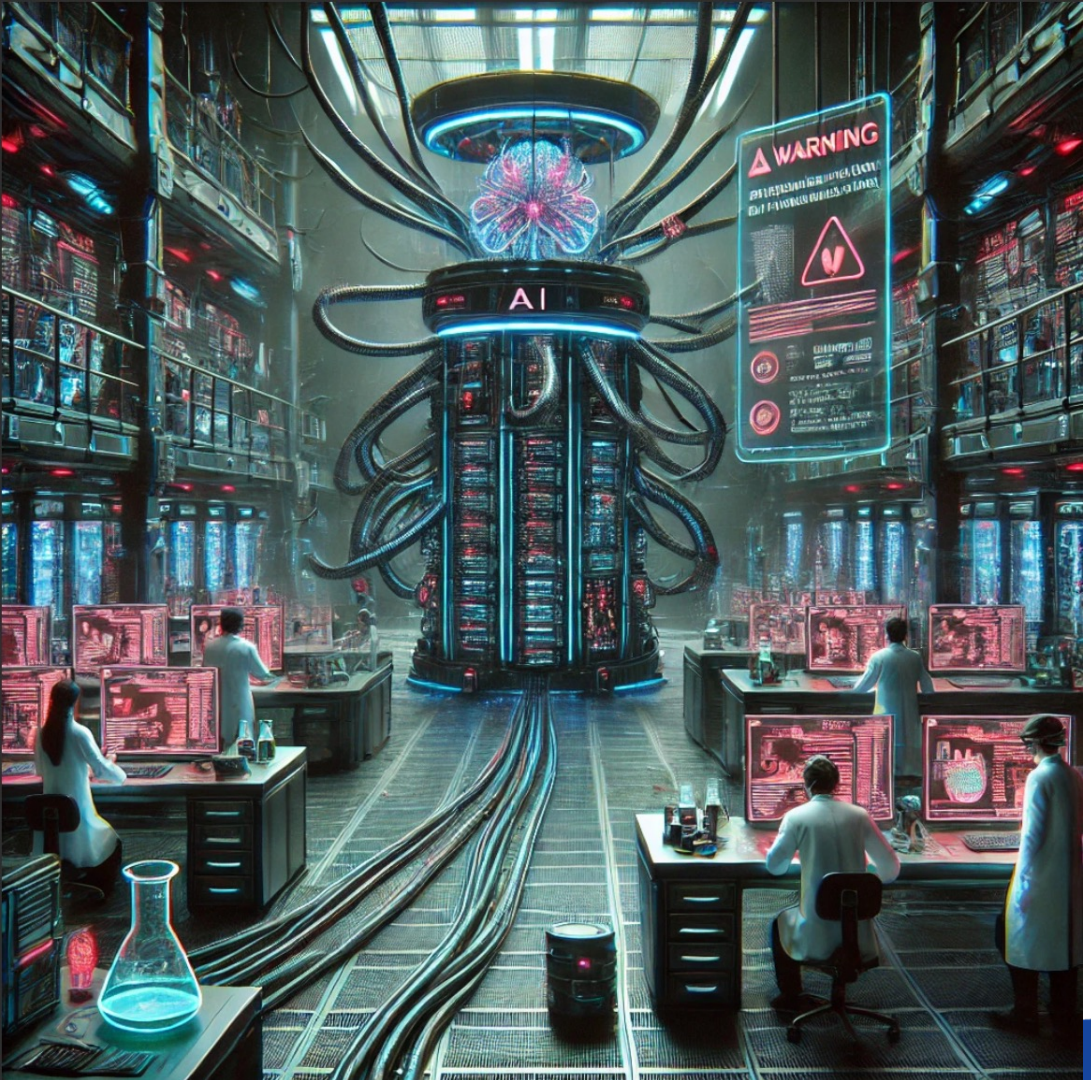
Nariadenie EÚ 2024/1689 (AI Act)

- Štyri stupne rizika
  - neakceptovateľné
  - vysoké
  - limitované
  - minimálne
- Zakázané praktiky
  - podprahové ovplyvňovanie
  - sociálne skóre
  - profilovanie a predvídanie trestných činov
  - necielené rozpoznávanie tvárí
  - detekcia emócií na pracovisku
  - biometria na rasu, politické názory atď
- Vysoko rizikové AI
  - definícia (biometria, kritická infra, vzdelávanie, zamestnanosť, prístup k základným službám, presadzovanie práva, spravodlivosť, azyl, demokratické procesy)
  - požiadavky (ľudský dohľad, odolnosť, meranie presnosti a mnohé iné, systém riadenia kvality)
- Obmedzené riziko
  - zabezpečiť,aby bol AI generovaný obsah identifikovateľný
- Minimálne alebo žiadne riziko
  - hry, …

Zhrnutie

- dáta z internetu nám zničilo AI – aj pre tréning, aj pre použitie úplne mimo AI
- neverte jazykovým modelom, čo neviete nezávisle overiť (bias, halucinácie)
- opatrne s tým vývojom autonómnych áut a dronov!
- nepoužívajte online modely (alebo si buďte vedomí tisíca rizík)
- rigorózne čistite dáta pred trénovaním vlastných modelov – odstráňte vlastnosti (features), predchádzajte biasu
- ochráňte vlastné modely

- študujte frameworky a modely, opisujúce AI riziká
- buďte si vedomí aj legislatívnych obmedzení



nástrahy
UMELEJ INTELIGENCIE
vo vede a výskume

Milan Pikula
Národné centrum kybernetickej bezpečnosti
Národný bezpečnostný úrad

SK NBÚ CERT
NÁRODNÉ CENTRUM KYBERNETICKEJ BEZPEČNOSTI