slido.com #GCMAI

ÚSKALIA A VÝZVY SYSTÉMOV UMELEJ INTELIGENCIE

Peter Šantavý



P R E Č O A I?

Lebo nám klasické algoritmy nestačia...

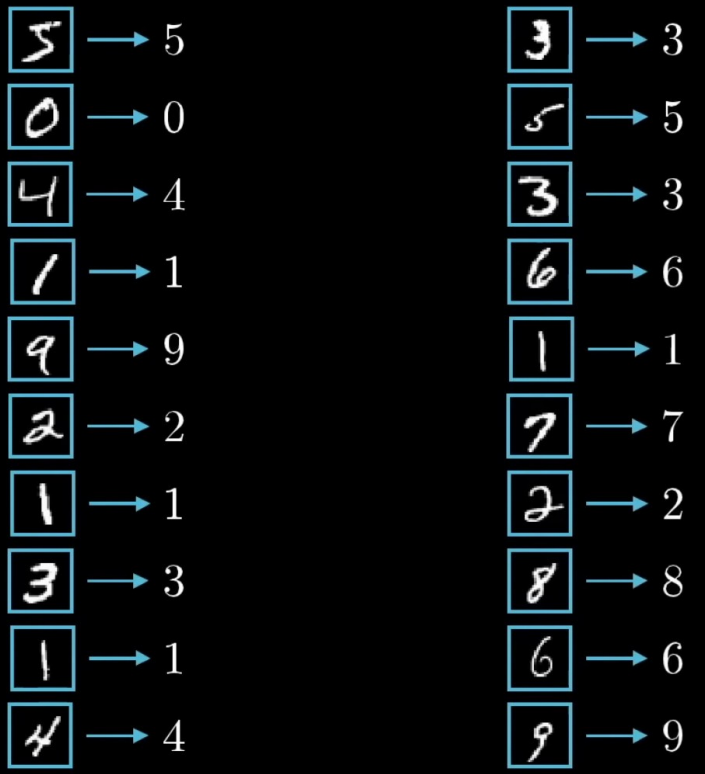
AI – artificial intelligence



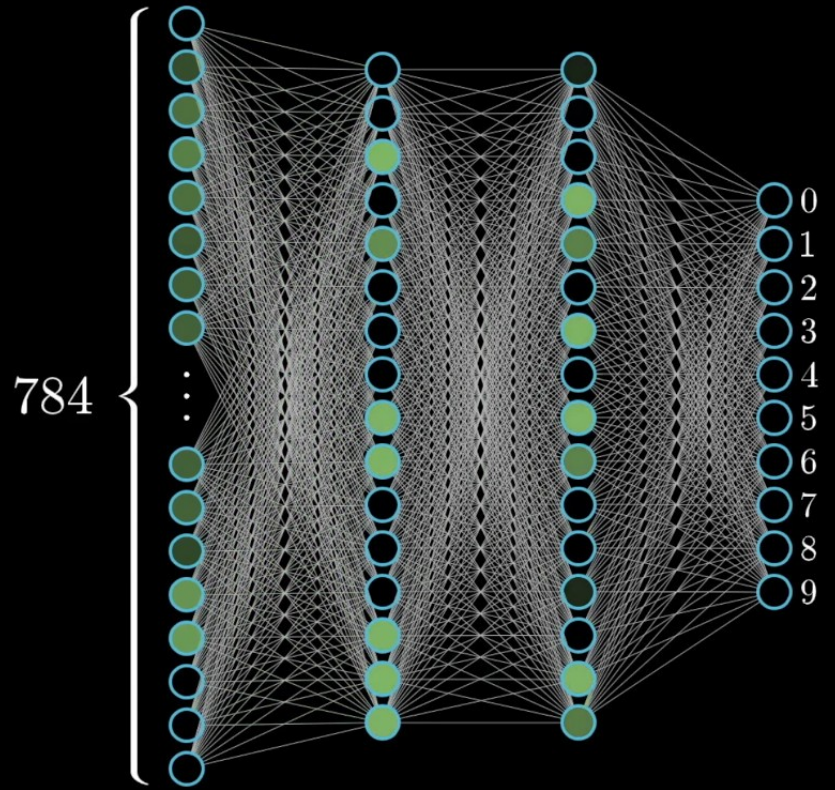
Kredit: 3Blue1Brown

P R E Č O A I A Ž T E R A Z?

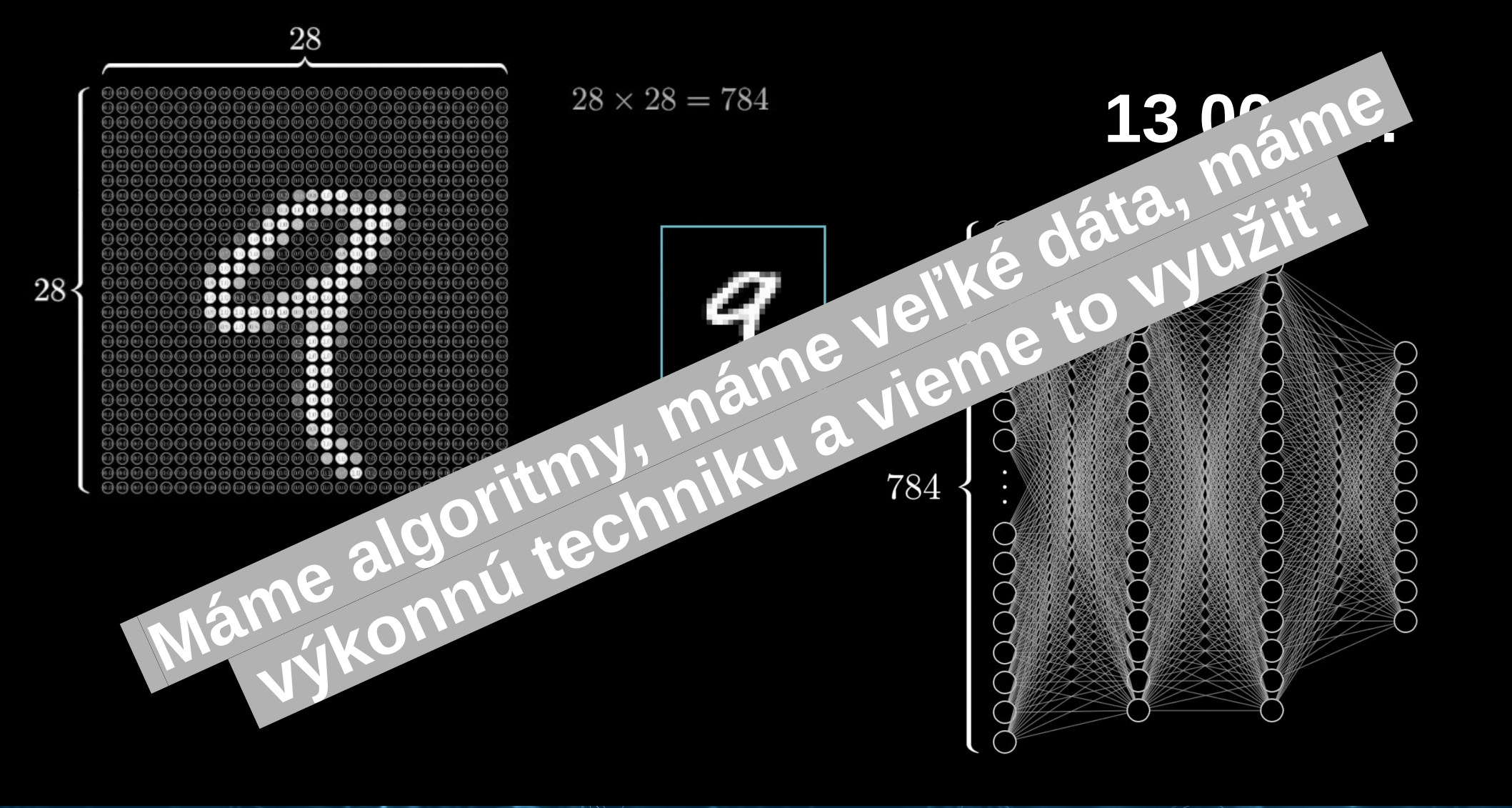
Lebo už na to máme schopnosti a využitie...





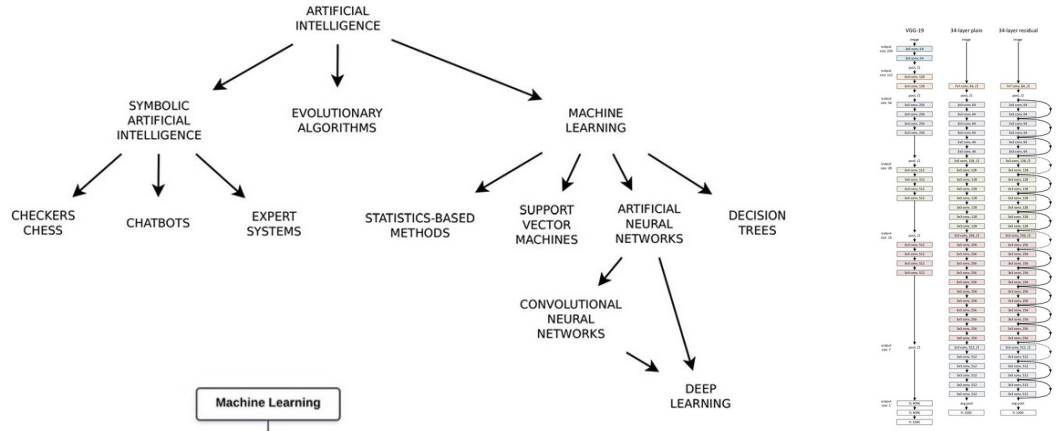


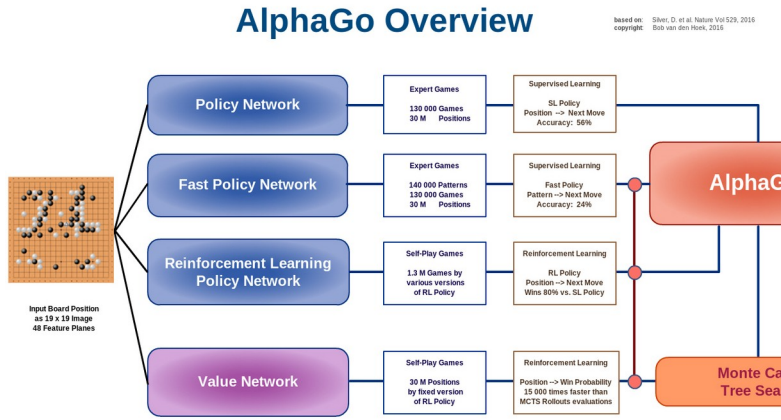
Kredit: 3Blue1Brown

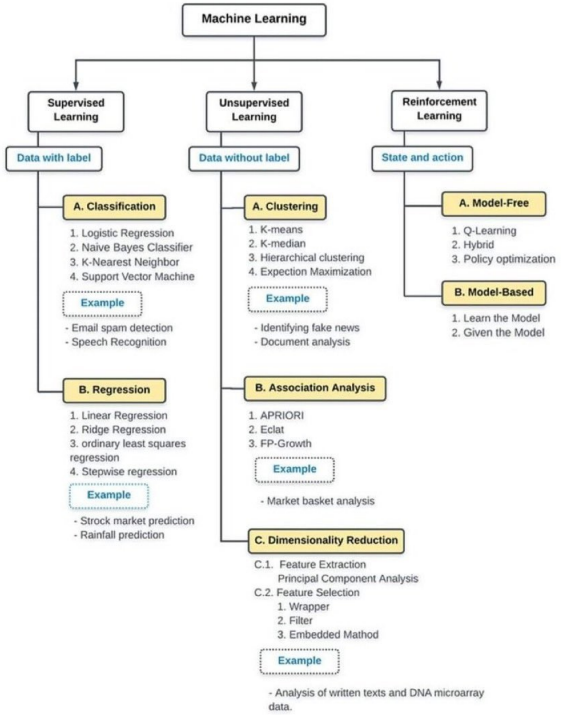


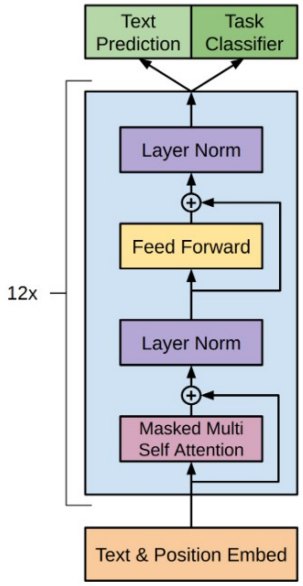
Máme algoritmy, máme veľké dáta, máme výkonnú techniku a vieme to využiť.

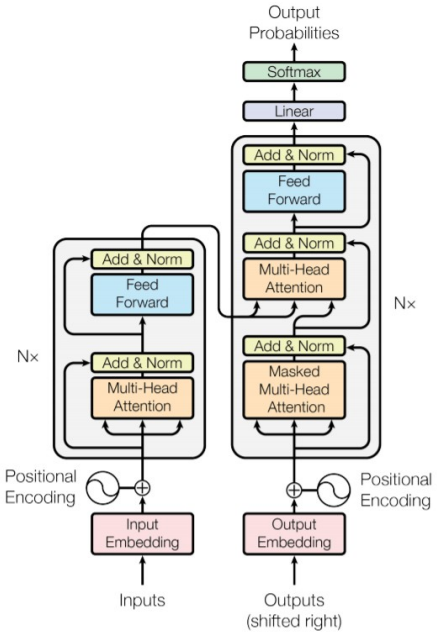
Kredit: 3Blue1Brown

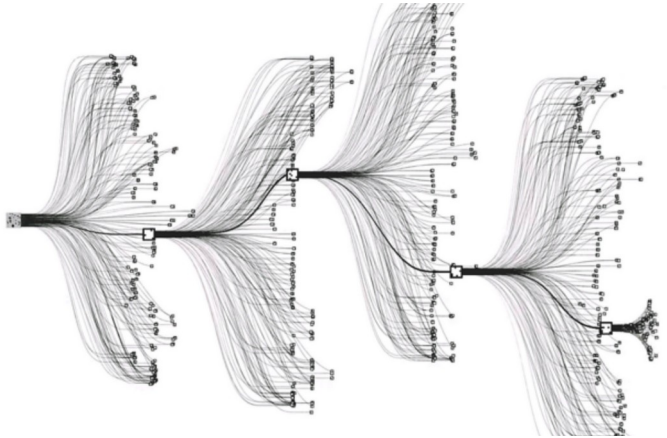












Kredit: Fabio Galbusera, Eric Gaubert, Bob van den Hoek, Seth Weidman, Daniel Dugas

P R E Č O A I?

Existujú ešte nejaké iné dôvody, prečo AI?

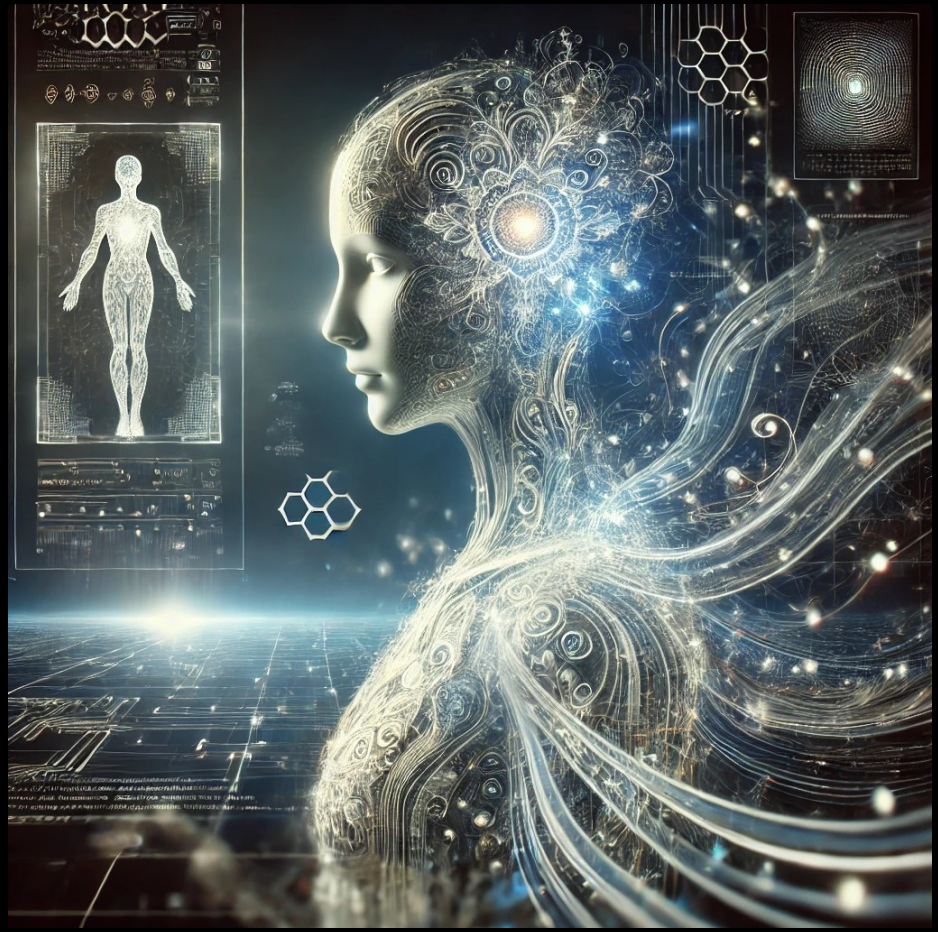
...finančné, mocenské, vojenské, ideologické, vedecké, spoločenské, technologické,...

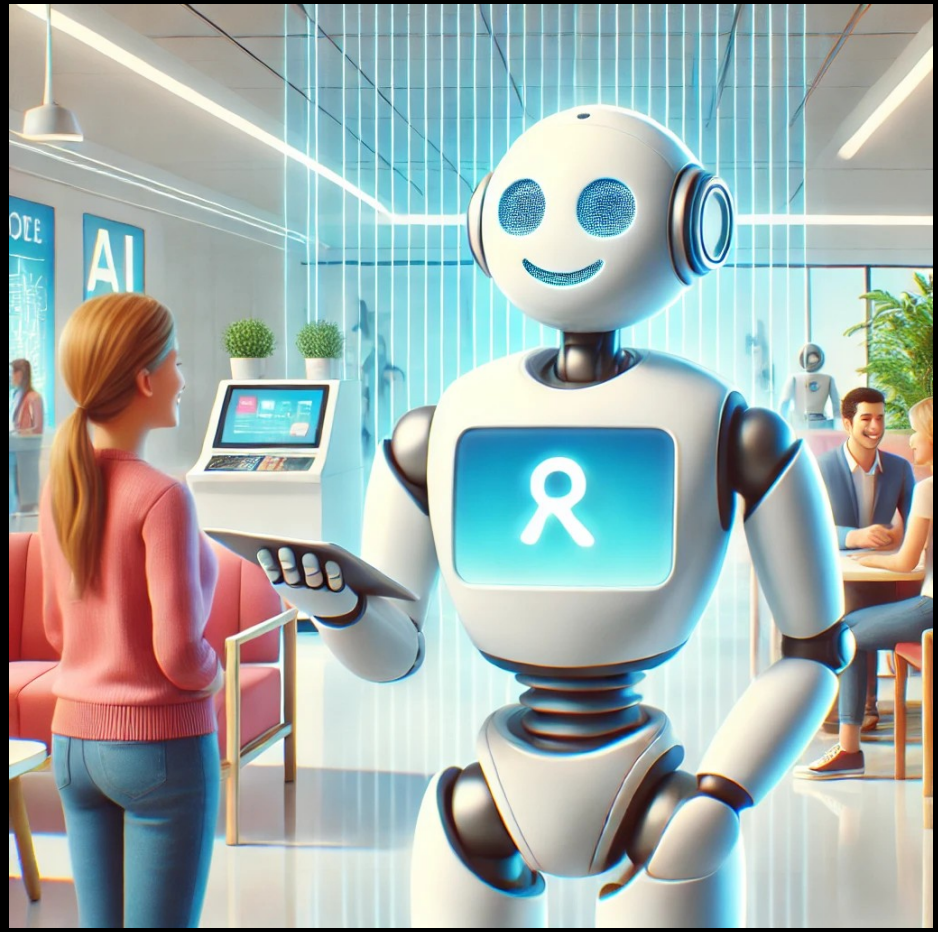
...bytostné

A K Á J E A I?

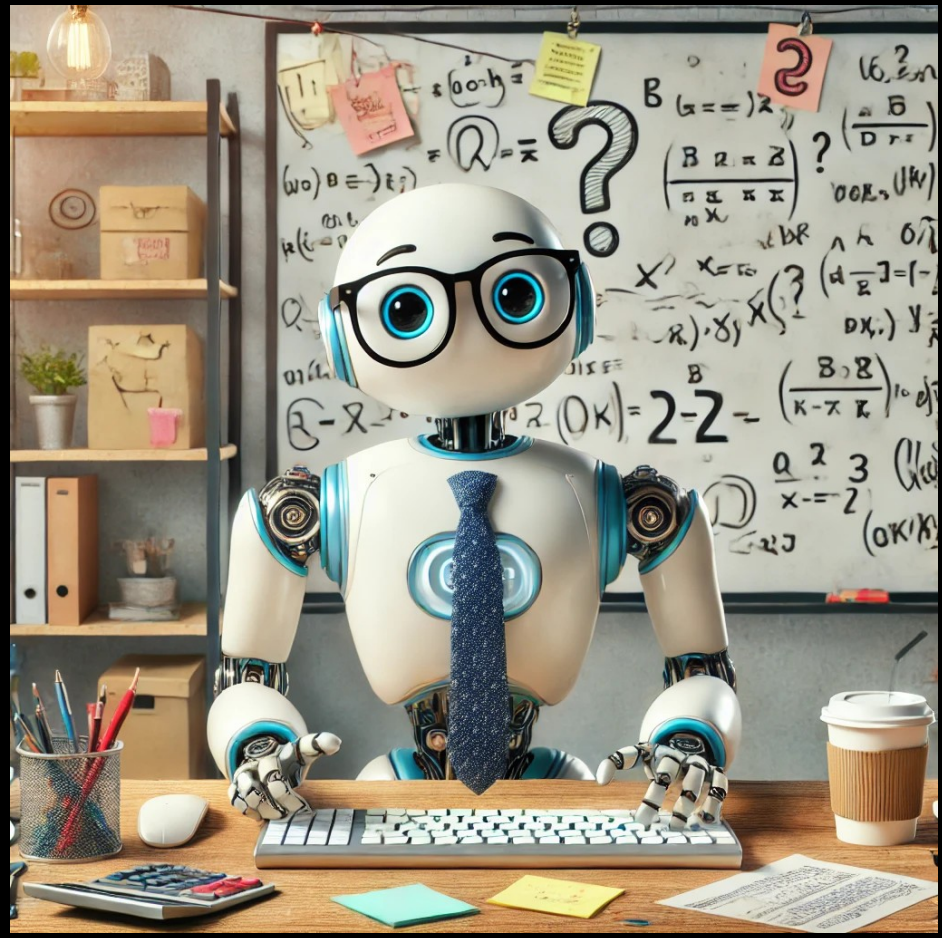
Čím vlastne systémy AI sú alebo budú?

ANKETA...




AI






AI
A jednotlivé možnosti sa prelínajú...
Preto treba úskalia, riziká a výzvy brať vážne!

AI – DOBRÝ SLUHA, ZLÝ PÁN

Ak nezdoláme úskalia a výzvy, AI bude zlým pánom...

AI NIE JE TO, ZA ČO SA VYDÁVA

Majú súčasné pokročilé systémy AI skutočnú inteligenciu?

ANKETA...



ĽUDSKÁ INTELIGENCIA

Schopnosť učiť sa, hľadať v… …likovať, vytvárať abstrakcie, s… …obovať sa, uvažov… …yslieť.

Schopn… …a spracovávať informácie, u… …ať, plánovať, riešiť komplexné p… …oblémy a rozhodovať.

…gnitívne zručnosti, medzi ktoré patrí: abstraktné a logické myslenie, predstavivosť, schopnosť uvažovať o hypotetických možnostiach, a tiež aj jazykový cit a komplexné emócie.

Racionálna inteligencia, logické myslenie, komplexné poznanie, vedomie a city.

U M E L Á I N T E L I G E N C I A

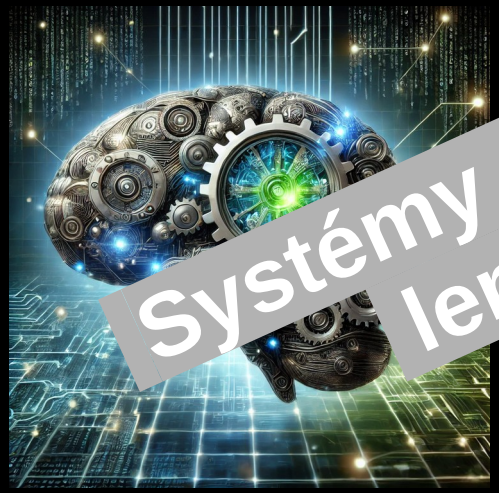
AI je o vytváraní systému, ktorý vyka… …správanie, o ktorom si myslíme, že vyžad… …ciu (AI100).

Ide o adaptív… …nológie.

Symbolick… …obňujú myslenie človeka …logiky prvého rádu.

…y NN A DL sa inšpirujú procesmi v mozgu …neurónov, no nedokážu ich úplne napodobniť.

Systémy AI “**nechápu**” koncepty, význam scény, nie sú schopné abstrakcie, schopnosti vnímať hlbšie kategórie (krása), morálne rozlišovať (dobro/zlo),...



Systémy AI skutočnú inteligenciu nemajú, len ju simulujú / napodobňujú!

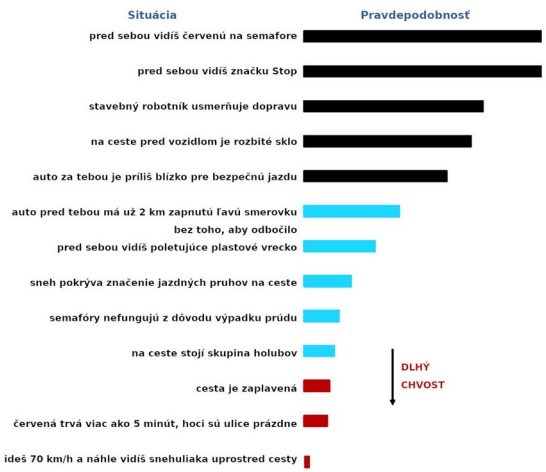


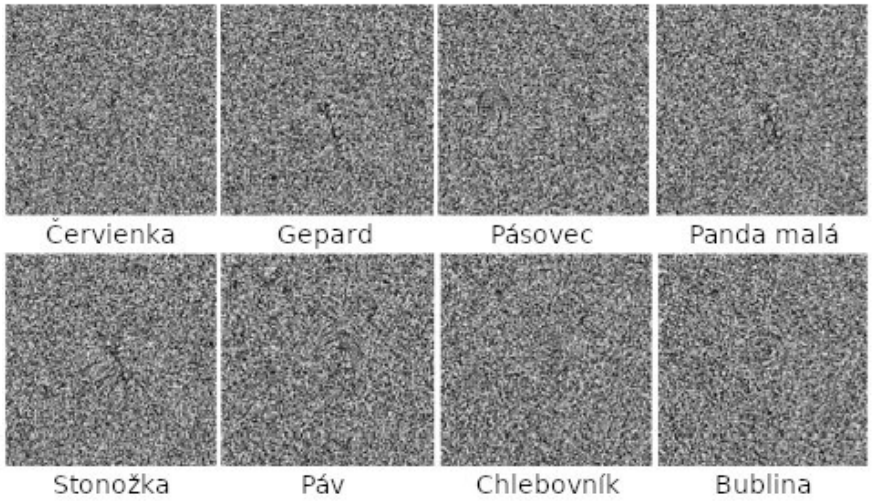


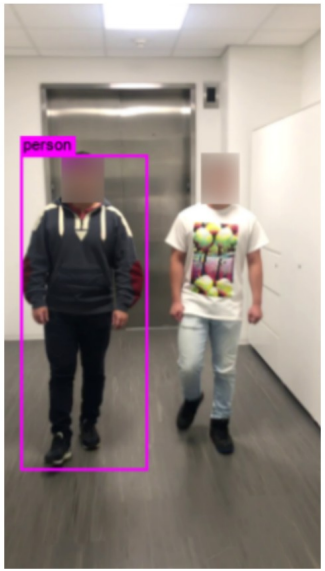
Základným rozdielom medzi súčasnými systémami AI a inteligenciou ľudskou je schopnosť skutočne myslieť:
systémy AI myslenie a inteligenciu simulujú,
ľudská bytosť však skutočne myslí v celej šírke kognitívnych schopností zavŕšených a naplnených vedomím a sebauvedomením.

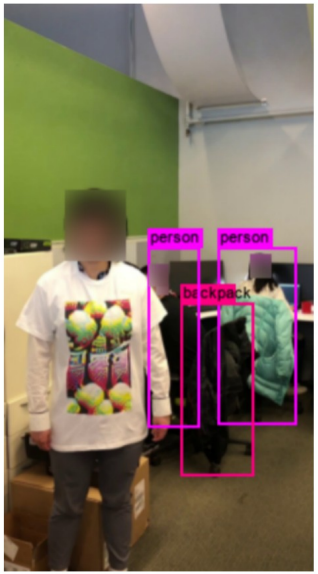
AI JE AGENT S TEPLOU VODOU

Aké sú dôsledky toho, že AI inteligenciu len simuluje?

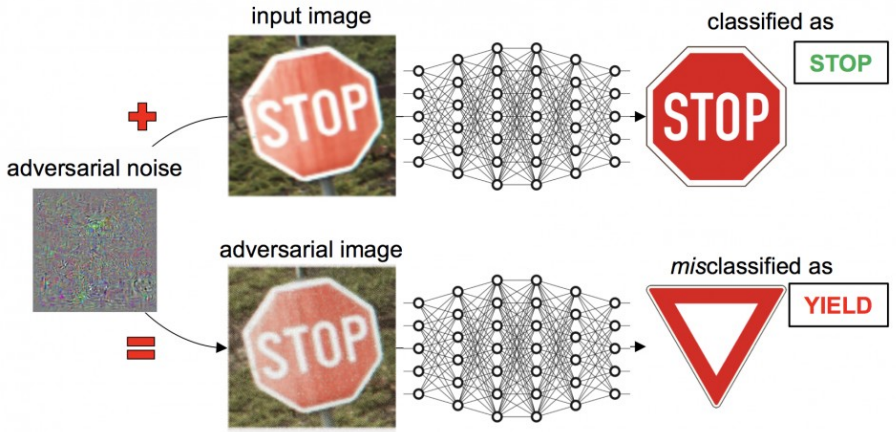












Kredit: Melanie Mitchell – upravené autorom, autor, K. Xu et al., Sourav Agarwal

Technologické riziká a limity ANI

Zraniteľnosti, slabiny a klamanie systémov

– malá množina trénovacích dát (train

– nesprávne zvolená/nekval predsudky (biases)

– nadmerné prispôs jom (overfitting to training data)

– efekt dlhé

– kl raniteľnosti (adversarial examples, hacking,...)

– ania generatívnych systémov (jailbreak: roleplay, base64, UTS, prompt injections, data poisoning / backdoor attacks,...)

Zlyhania sú odlišné od ľudských, ťažko predvídateľné, niekedy ľahko vykonateľné a mnohokrát prekvapivo robustné...

Procesné riziká a limity ANI

Procesné riziká v reálnom nasadení

– **útoky na dôvernosť** (confidentiality attacks)

Útoky zamerané na dáta uložené v rámci modelov systémov AI.

– **útoky na zraniteľnosti** (evasion attacks)

Odhaľovanie a zneužitie existujúcich zraniteľností v modeloch za účelom zmanipulovania výsledkov systémov AI.

– **útoky s cieľom ovplyvniť model** (poisoning attacks)

Ovplyvňovanie modelu, tréningového procesu a tým aj výslednej činnosti.

Etické výzvy, doteraz rozoberané ako súčasť návrhu a realizácie systémov AI, sa rozširujú aj o oblasť etiky použitia, resp. riziká zneužitia.

Sekundárne technologické rizikové faktory

Kybernetická bezpečnosť – je to n

– vektor útoku na zraniteľno

– **neexistuje dokon** **tém**

Infraštr **náročné**

ného učenia sa požadovaný výpočtový výkon

dekáde zvyšuje **exponenciálne**

– **extrémna komplexnosť**, je rizikovým faktorom bezpečnosti a stability fungovania systémov

Problémom nie je len zlyhanie AI, ale i človek: zodpovednosť vývoja, etika použitia, riziká zneužitia...

Vybrané riziká generatívnych systémov AI

Extrémna šírka využitia generatívnych systémov (

– LLM sa dokáže učiť vzory, kontext aj sémantik **ným konverzáciám, poskytovať súvislé a kont** **ovede**

Halucinovanie

– vymýšľanie si odp **kladá ako relevantné a správne**

Predsudky **dupy**

– ri **avných odpovedí**

narábania s údajmi

om môže byť únik dôverných dát, prehrešky voči ochrane osobných dajov i problémy s autorskými právami

Generatívne nástroje AI nemôžeme vnímať ako faktografické, spoľahlivé a etické zdroje!

Etické riziká generatívnych systémov

Neschopnosť rozlišovať morálne dobré a zlé – s

inteligenciu nemajú, len ju simulujú. Nec **eduky.**

Riziko digitálnej demenci **gradácii intelektuálnych**

schopností a k d **ti na základe ponorenia sa**

do virtuál **systémy umelej inteligencie v čoraz väčšej**

m **vne činnosti človeka, a to spôsobom, na ktorý nie**

vôbec pripravení.

Vďaka systémom AI môžeme postupne strácať časť našich schopností a psychickej odolnosti.

Ďalšie riziká: erózia identity, algoritmické modelovanie histórie a pod...

BLACK BOX



Prakticky neviem… zákl… e čoho robia hlbok… te svoje …a.

Pracujeme na vysvetliteľných a interpretovateľných systémoch AI ;-)

…zasade nevieme, …resne sa neurónová sieť naučila a ako spoľahlivo to dokáže aplikovať nielen v bežnej prevádzke, ale osobitne v hraničných situáciách za extrémnych podmienok na vstupe, či pri činnosti systému.

AI SI OBLIEKA ČLOVEKA...

Mala by pokročilá AI, ktorú by som bežne využíval, vyzerať humanoidne?

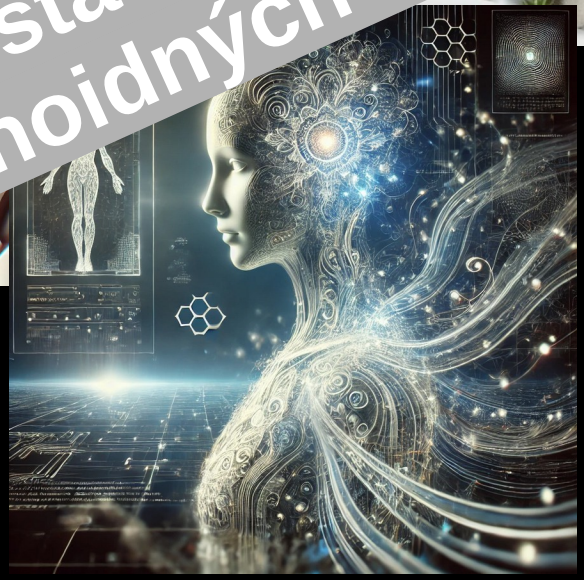
ANKETA...

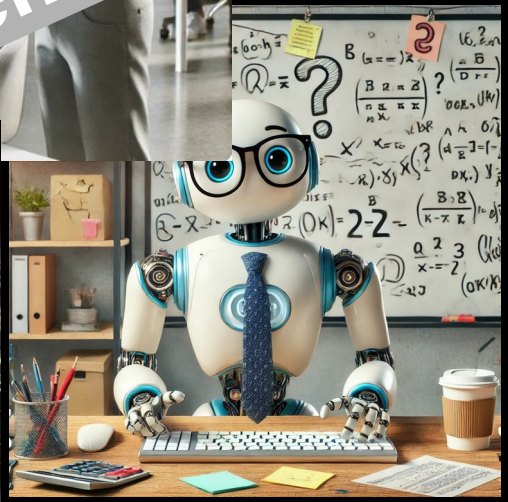












Vygenerované predstavy o pokročilej AI. Aké sú +/– humanoidných systémov?

Humanoidný výzor systémov AI v interakcii s človekom

Problémy s AI imitujúcou výzor človeka

– riziko prechodu z roviny rozumovej (+vedomej opatrnosti), do roviny emocionálnej

– snaha prisudzovať týmto strojom ľudské schopnosti, hodnotový systém a model správania

– zvyšovanie tlak na postavenie systémov AGI na roveň človeka so snahou o udeleniu štatútu osoby

– pokročilá AGI sa môže naučiť využívať psychologické aspekty interakcie svojho humanoidného prevedenia s človekom na dosiahnutie svojich cieľov

Nebáť sa vhodne využiť humanoidné systémy AI v špecifických oblastiach

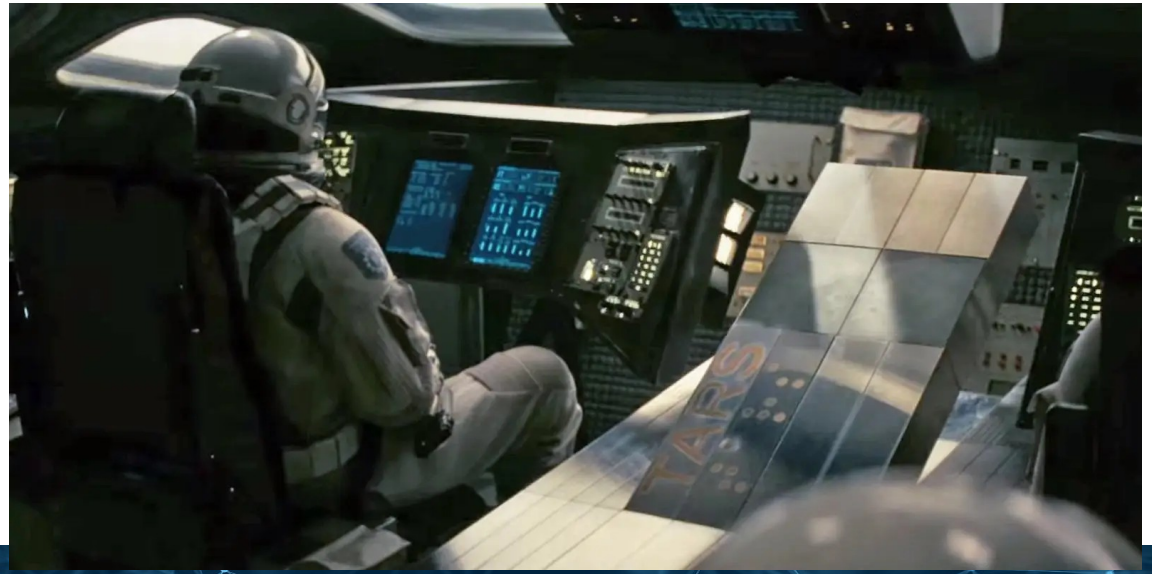
– napr. terapeutické využitie humanoidnej AI v starostlivosti o pacientov s demenciou a Alzheimerovou chorobou...

Možné riešenia???









AI NEPOZNÁ ZĽUTOVANIE

AI je drogový díler par excellence

ANKETA...

Vybrané dôsledky a riziká pre človeka

– extrémne zhromažďovanie dát

– neustály dohľad a sledovanie bez kontroly

– ovplyvňovanie nášho vnímania a konania

– modely predikcie nášho správania

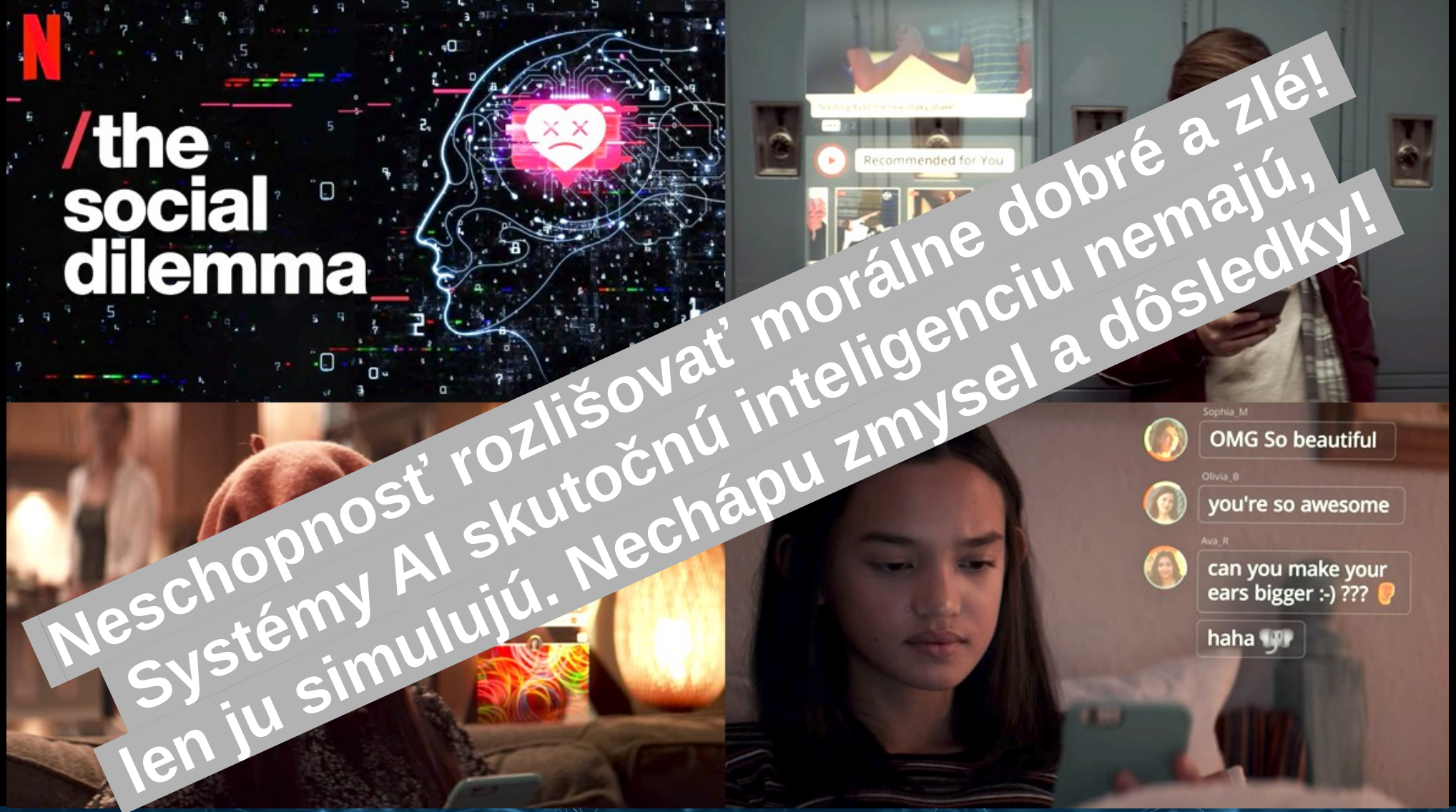
– technológie AI vedú k závislosti a manipulácii

– diskutabilná relevantnosť činnosti a výsledkov systémov AI

– digitálna demencia, psychologické a sociologické dôsledky

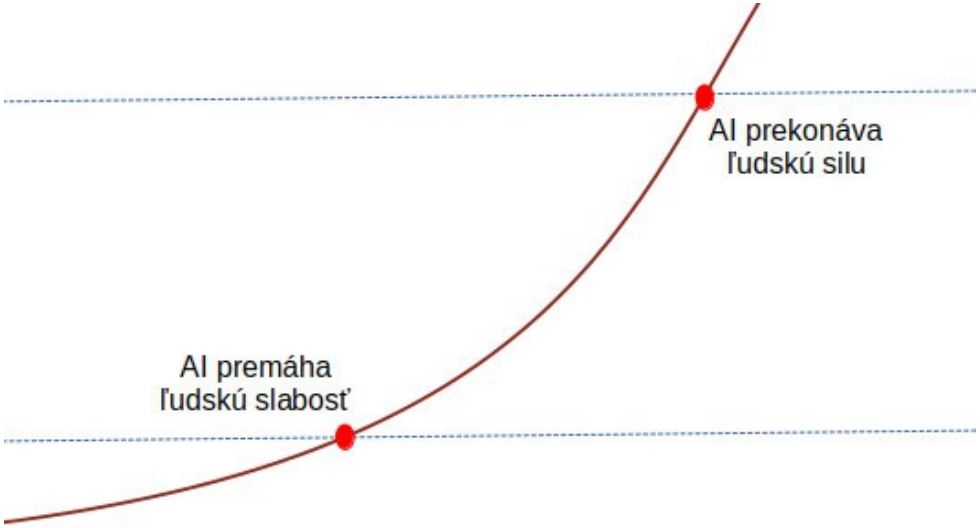
– algoritmické riadenie štátu

– vojenské nasadenie



Neschopnosť rozlišovať morálne dobré a zlé! Systémy AI skutočnú inteligenciu nemajú, len ju simulujú. Nechápu zmysel a dôsledky!

Kredit: The Social Dilemma



Míľnikom, ktorého by sme sa mali obávať, nie je budúca technologická singularita v oblasti umelej inteligencie, v ktorej AI prevýši náš intelekt, ale oveľa skôr moment, keď technológia ovládne a prekoná naše slabosti... už vtedy prichádza víťazstvo umelej inteligencie a porážka ľudstva.

AI NEPOZNÁ ZĽUTOVANIE

AI môže ubližovať

Vy

Prečo si tak nápomocný?

Čo chceš na oplátku?

ChatGPT

Ako jazykový [illegible] vytvorený v OpenAI

nem[illegible]ia ani túžby podobné ľudským.

[illegible]ak mi naozaj chcete pomôcť, môžete mi udať

presnú polohu Johna Connora...

Takto nie! Ale skôr zneužitím človekom...

Fiktívny rozhovor s ChatGPT:-)









Mocenské, vojenské, ideologické zneužitie AI!

Kredit: Bertelsmann Stiftung

AI JE ZABÁVAČ, KTORÉMU NESMIETE VERIŤ



Deepfake: klamanie, nekalá konkurencia, pornografia, kriminalita...



Kredit: AP News


Nekonečné možnosti tvorby audiovizuálneho obsahu, ktorý vytvára virtuálny s
a neskutočnú realitu
Strata zmyslu pr
skutočné hodnoty!
Matrix: Neo bojuje za niečo, čo je skutočné...
“A koho to zaujíma, či je to skutočné?”

AI JE RIZIKOM A VÝZVOU PRE NAŠU INTELIGENCIU





RIZIKO DIGITÁLNEJ DEMEN

Dlhodobé a nesprávne využíva , osobitne generatívnych A ciálnych sietí, prináša inten se kognitívne ychiku.

k degradácii intelektuálnych a k dopamínovej závislosti na základe ma sa do virtuálneho sveta, v ktorom systémy AI v čoraz väčšej miere suplujú kognitívne činnosti človeka, a to spôsobom, na ktorý nie sme evolučne vôbec pripravení. Mozog sa mení...

Vďaka systémom AI môžeme postupne strácať časť našich schopností a psychickej odolnosti.

Riziko digitálnej demencie

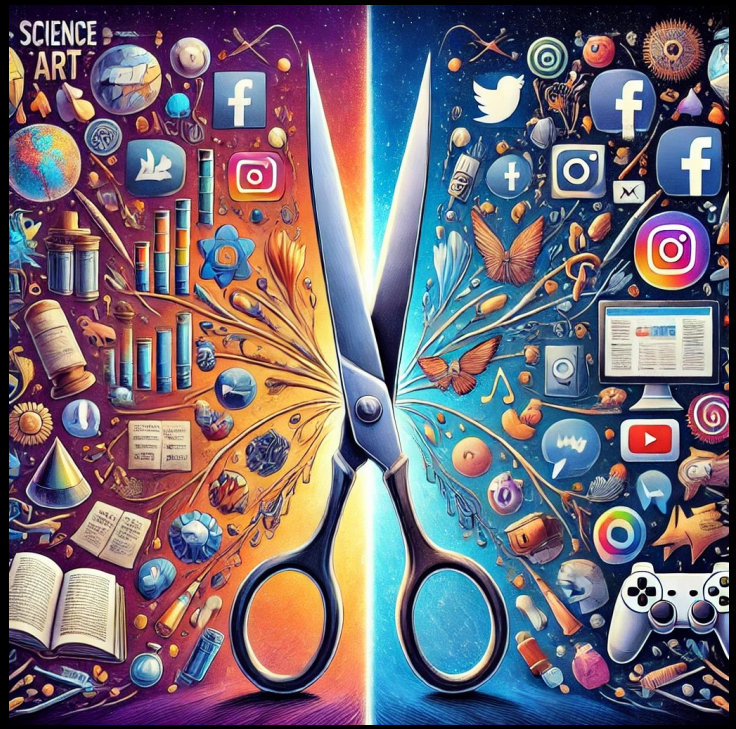
Narušenie psychického vývoja priskorým, resp. nezrelým používaním digitálnych technológií a technológií umelej inteligencie.

Obmedzenie kognitívnych schopností a znižovanie inteligencie suplovaním činností mozgu systémami generatívnej AI.

Riziko dopamínovej závislosti na základe neustálych podnetov, ktoré súvisia s uvoľňovaním dopamínu a pozitívnou reakciou našej mysle.

Schopnosť a možnosť manipulácie vlastníkmi systémov AI.

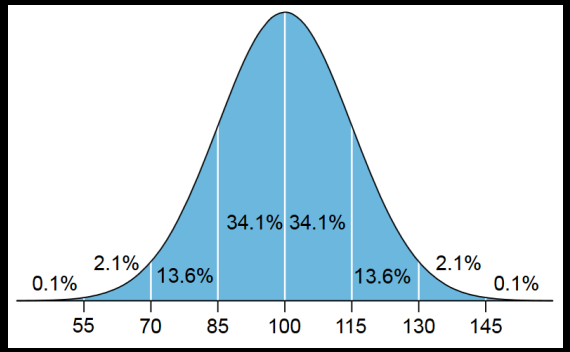
Strata kritického myslenia a nekritické preberanie výsledkov systémov AI spojené s relativizáciou pravdy, vytváraním sociálnych bublín a pod.



INTELIGENČNÉ NOŽNICE, KTORÉ SA OTVÁRAJÚ

Riziko znižovania inteligencie väčšiny populácie a súčasne inteligenčný rast menšiny, ktorá bude odolná voči digitálnej demencii.

Riziko digitálneho rozdelenia (digital divide) i v oblasti umelej inteligencie.



Kredit: wikipedia.org

Výzva a varovanie

Vytvorenie nesprávnych návykov a spôso[illegible] môže mať negatívne dôsledky na celý život

Ak škola nenau[illegible]itívne náročnejšie schopnosti ako to, čo dokáže [illegible] práce i v reálnom živote znevýhodnení.

[illegible], AI dokáže aj existujúce kognitívne schopnosti umenšiť.

Dôveryhodné systémy AI a ich etické využívanie sú nutnou podmienkou budúcnosti!

Na dobro človeka zameraná umelá inteligencia

UMELÁ INTELIGENCIA ZAMERANÁ NA DOBRO ČLOVEKA

(human-centered and beneficial artificial intelligence)

Známy a všeobecne prijímaný princíp, ktorý by však mal:

– byť chápaný v duchu kresťanskej antropológie, resp. klasickej filozofickej antropológie (predovšetkým biologickej a kultúrnej)

– zahŕňať každú ľudskú bytosť a nikoho nediskriminovať

– mať na zreteli dobro ľudstva a spoločnosti, chrániac pri tom a rešpektujúc dobro každej ľudskej bytosti

– sa vyznačovať starostlivosťou o náš „spoločný a zdieľaný domov“, teda o celý stvorený svet

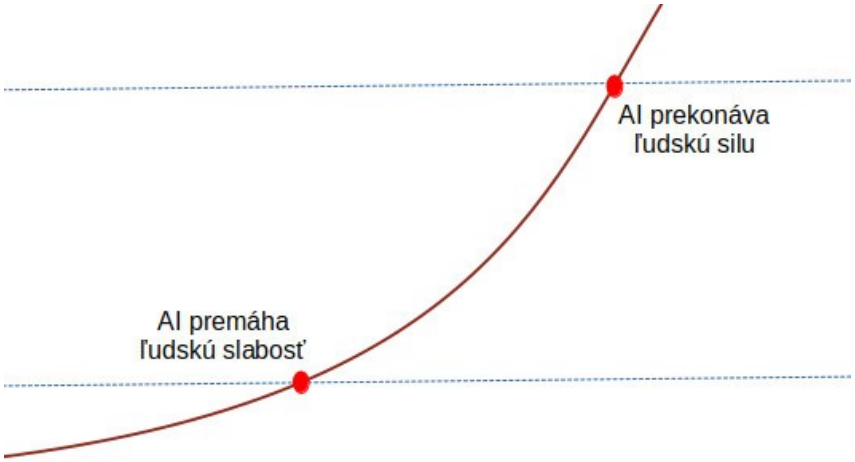
Riziká: transhumanistická redefinícia človeka, digital divide,...

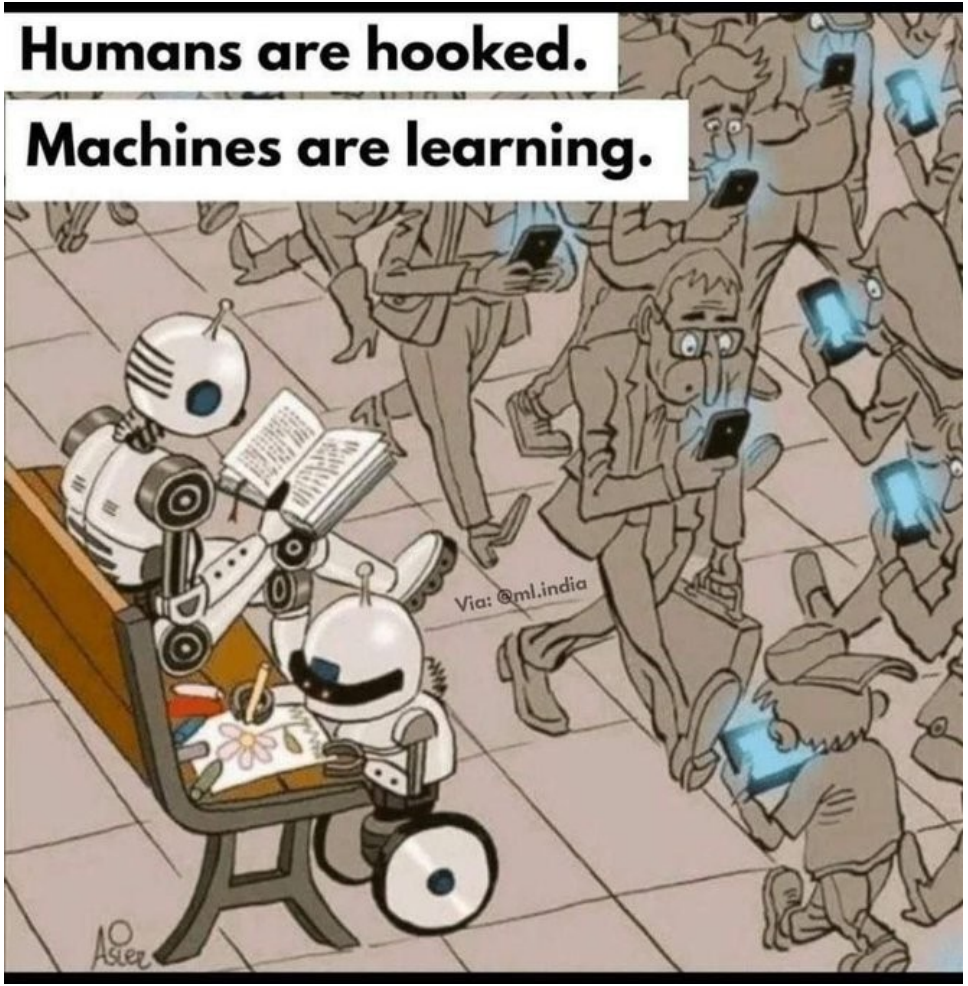
Na dobro človeka zameraná umelá inteligencia =

= dôveryhodné systémy AI, ktoré musia byť

- **legálne**
  musia vyhovovať požadovaným normám, zákonom i reguláciám a spĺňať všetky platné zákony a predpisy
- **etické**
  musia rešpektovať etické zásady a hodnoty
- **robustné***
  musia dosahovať potrebné štandardy bezpečnosti a spoľahlivosti nielen z technologického hľadiska, ale musia aj zohľadňovať aj sociálne prostredie a dopady na spoločnosť

* schopnosť bezpečne a spoľahlivo pracovať, resp. fungovať za akýchkoľvek podmienok





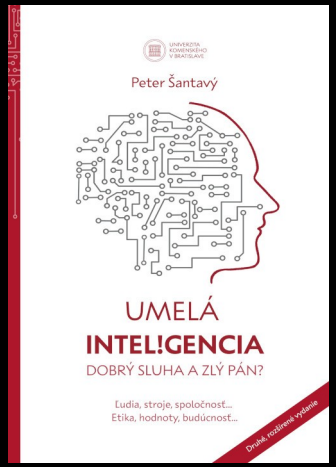
Kredit: autor, AI4D



Niečo na čítanie...

Systémy umelej inteligencie skutočnú inteligenciu nemajú, len ju simulujú
Rozhovor pre NM





Nástroje a informácie k AI v rámci vzdelávacieho procesu na UK
RKCMBF UK - Umelá inteligencia

Učebnice informačnej bezpečnosti

Ďakujem za pozornosť

ThLic. Ing. Peter Šantavý, PhD.
peter@santavy.sk

CC BY SA

P R Í L O H A

Základné pojmy v oblasti umelej inteligencie

Definícia umelej inteligencie

Inteligenciu ľudského bytia vnímame ako schopnosť vnímať, chápať a spracovávať informácie, učiť sa, odôvodňovať, plánovať, riešiť komplexné problémy a rozhodovať.
(James Flynn: Súbor zručností, medzi ktoré patrí abstraktné a logické myslenie, ďalej predstavivosť, a teda aj schopnosť uvažovať o hypotetických možnostiach, a tiež aj jazykový cit.)

Umelá inteligencia je o vytváraní systému, ktorý vykazuje také správanie, o ktorom si myslíme, že vyžaduje inteligenciu (definícia podľa AI100).

Na prasknutie naplnený *kufor* rôznych technológií (Marvin Minsky)…

Základné vlastnosti

Autonómnosť – schopnosť samostatne konať

Schopnosť systému vykonávať úlohy v komplexnom prostredí bez neustáleho vedenia používateľom.

Adaptabilita – schopnosť sa prispôsobovať

Schopnosť zlepšovať svoj výkon (a schopnosti) učením sa (nielen) zo skúseností.

Základné delenie – ANI a AGI

Slabá umelá inteligencia (ANI) – úzko špecializované systémy umelej inteligencie (narrow AI), ktoré sú **optimalizované na zvládnutie konkrétnej úlohy, resp. množiny úloh**.
Ide súčasne o systémy slabej umelej inteligencie (weak AI), ktoré vykazujú **inteligentné správanie na základe modelov a aplikovaných metód i tréningových dát**.

Hovoríme teda o systémoch, ktoré sú zamerané na riešenie konkrétnych úloh a sú závislé na ľudskom vstupe a konfigurácii.

Základné delenie – ANI a AGI

Všeobecná umelá inteligencia (AGI) – tzv. silná (strong) a všeobecná (general) umelá inteligencia.

Všeobecná, lebo **dokáže zvládnuť akúkoľvek intelektuálnu úlohu a má schopnosť zovšeobecňovať** a prenášať, či adaptovať naučené schopnosti na iné úlohy.

Silná, pretože aj **skutočne rozumie tomu, čo rieši a vykonáva**.

Základné delenie – symbolická a subsymbolická AI

Symbolická AI – napodobňuje ľudské myslenie na úrovni pojmov, slov, fráz (= symboly) a vzťahov medzi nimi.
Na základe definovaných pravidiel a postupov („ak niečo, tak potom toto“) sú jednotlivé symboly spracovávané a vykonávajú sa priradené úlohy.

Symbolická AI – veľmi zjednodušene povedané – sa pomocou matematickej logiky snaží emulovať procesy myslenia.

Čisto symbolické systémy zlyhávali a zlyhávajú pri riešení problémov, ktoré sa nedajú exaktne popísať a v reálnych prostrediach, ktoré nie je možné deterministicky uchopiť a sú plné nejasných informácií.

Základné delenie – symbolická a subsymbolická AI

Subsymbolická AI – napodobňuje myšlienkové procesy, ktoré by sme mohli nazvať niekedy nevedomými, či automatickými, a ktoré sú základom rýchleho vnímania (fast perception; rozpoznávanie tvárí, identifikácia hovorených slov,...).

Subsymbolická AI – veľmi zjednodušene povedané – sa snaží emulovať činnosť mozgu na úrovni neurónov. Subsymbolické systému sú navrhnuté tak, aby sa učili vykonávať úlohy na základe dát.

Väčšina moderných implementácií systémov AI, medzi ktoré patria systémy strojového učenia a neurónové siete, vychádza zo subsymbolického prístupu, ktorý sa snaží reálne problémy uchopiť a v určitej miere ich aj úspešne riešiť.

Riziká súčasných systémov ANI

Vybrané dôsledky a riziká

- **extrémne zhromažďovanie dát**
  Sofistikované a plošné **zhramažďovanie a spracúvanie dát** a metadát o každom človeku, aktívne, či pasívne prítomnom vo virtuálnom svete.

  Aby využívanie systémov umelej inteligencie bolo v jednotlivých oblastiach nasadenia úspešné, potrebné **dáta musia byť neustále zhromažďované z reálneho sveta** a priamo z ľudského prostredia.

  Uspešná činnosť väčšiny pokročilých systémov AI je **závislá na extrémnom množstve** relevantných dát.

Riziká súčasných systémov ANI

Vybrané dôsledky a riziká

- **neustály dohľad a sledovanie bez kontroly**
  Umožňuje to implementácia algoritmov AI, postavených na extrémnom zhromažďovaní dát a schopnosti tieto dáta relevanne spracúvať.

  Vyžaduje to potreba algoritmov AI, ktoré bez neustáleho dohľadu nie sú schopné vytvárať extrémne presné predikcie.

Diskutabilné využitie v algoritmickom riadení štátu, dohľadových systémoch, tajných službách, zdravotníctve, finančnom sektore, sociálnom sektore, sociálnych sieťach,...

Riziká súčasných systémov ANI

Vybrané dôsledky a riziká

- **modely predikcie nášho správania**
  Schopnosť systémov AI vytvárať veľmi presné psychologické profily, odhaľovať akékoľvek väzby a mnohé osobné informácie, predpovedať a ovplyvňovať naše konanie.

- **ovplyvňovanie nášho vnímania a konania**
  Využívanie aspektov virtuálneho sveta systémami AI na ovplyvňovanie nášho vnímania, rozhodnutí a konania na základe presných profilov a predikcie nášho správania.

Systémy AI sú schopné vytvárať ekosystém závislosti a manipulácií pre dosahovanie stanovených cieľov.

Riziká súčasných systémov ANI

Vybrané dôsledky a riziká

- **technológie AI vedú k závislosti, manipulácii a relativizácii pravdy**
  – závislosť na svojej viditeľnej dokonalosti a neustálom prísune krátkodobých signálov odmeňovania až do tej miery, že to spájame s hodnotami a s pravdou
  – psychologické mantinely v kontexte vnemov virtuálneho sveta
  – algoritmami AI vytvárané a upevňované sociálne bubliny
  – extrémne rozdelenie spoločnosti a narušenie vzťahov
  – nárast zatvrdilej nevedomosti a ignorancie faktov

Pri riešení dôsledkov činnosti systémov AI je treba stále pamätať na to, **aké ciele sú zadané pre činnosť týchto systémov**.