ZÁKLADNÉ ETICKÉ VÝZVY A RIZIKÁ SYSTÉMOV UMELEJ INTELIGENCIE V OPTIKE VZDELÁVANIA

Február 2024

Peter Šantavý

Základné etické výzvy a princípy systémov AI

Riziká a limity súčasných systémov AI

AI – artificial intelligence

Existujú len dva druhy priemyslu, ktoré svojim zákazníkom hovoria používatelia.

Nelegálne drogy a softvér.

Prof. Edward R. Tufte

Na dobro človeka zameraná umelá inteligencia

Human-centered and Beneficial Artificial Intelligence

Na dobro človeka zameraná umelá inteligencia

Známy a všeobecne prijímaný princíp, ktorý by však mal

– byť chápaný v duchu klasickej filozofickej antropológie

– 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 svet

Interdisciplinárny rámec ako základ

Etické problémy nie je možné riešiť bez

– dostatočného oboznámenia sa s technologickou stránkou systémov AI

– bez skúmania psychologických, sociologických a právnych aspektov nasadenia technológií AI

– pokrytia veľkého rozsahu diskutovaných tém a oblastí dopadu

– chápania dôležitosti netechnických disciplín pre nasadenie v reálnom svete

– spolupráce s inštitúciami na medzinárodnej i štátnej úrovni

– interdisciplinárneho rámca vzdelávania

Na dobro človeka zameraná umelá inteligencia ~ dôveryhodná umelá inteligenca

Základné požiadavky: dôveryhodné systémy AI musia byť

– legálne (lawful) – vyhovovať požadovaným normám a spĺňať všetky platné zákony, predpisy i regulácie

– etické (ethical) – rešpektovať etické zásady a hodnoty

– robustné (robust) – dosahovať potrebné štandardy bezpečnosti a spoľahlivosti nielen z technologického hľadiska, ale zohľadňovať aj sociálne prostredie a dopady na spoločnosť

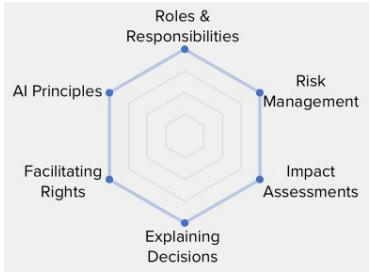
Dôveryhodná AI – praktické dôsledky

Súčasný rámec riešenia etických problémov ANI – regulácie a etika

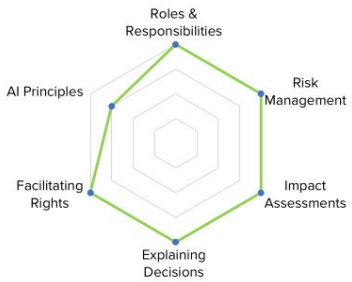
- **Artificial Intelligence Act**
  Nariadenie Európskeho parlamentu a Rady, ktorým sa stanovujú harmonizované pravidlá v oblasti umelej inteligencie (**Akt o umelej inteligencii**) a menia niektoré legislatívne akty únie
  [https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0206]

- **Rome Call for AI Ethics**
  Etické odporúčania vatikánskej konferencie **renAIssance 2020**
  [https://www.romecall.org/]

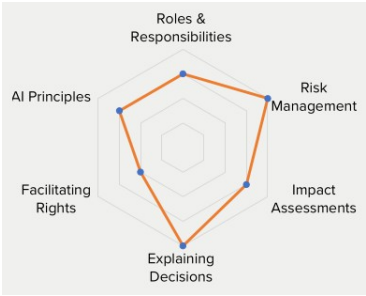
EU AI HLEG Guidelines and Assessment List for Trustworthy AI



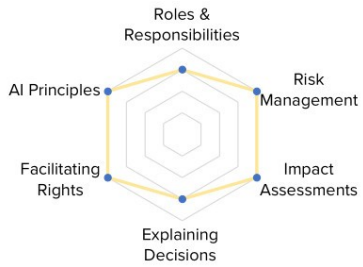
UK's ICO AI Auditing Framework



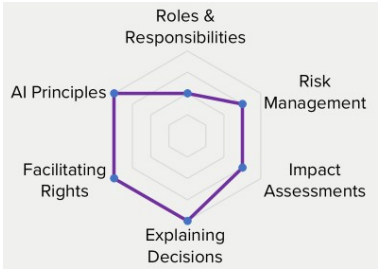
Singapore Model Governance Framework

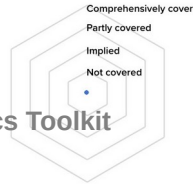


Hong Kong Ethical Accountability Framework



Dubai AI Ethics Toolkit





Kredit: OneTrust DataGuidance Limited

(Nielen) angažovanosť, legislatívne kroky a regulácie...

Všeobecné nasadenie

– **Artificial Intelligence Act** (EU)
– **Ethics of AI** (UNESCO)

Aktivity

– #**AIEthics** tímy
– AI alignment research...

Armádne nasadenie

– AI Principles: Recommendations on the Ethical Use of Artificial Intelligence by the DoD
– **REAIM 2023**: 1. global Summit on Responsible Artificial Intelligence in the Military Domain

UMELÁ INTELIGENCIA ZAMERANÁ NA DOBRO ČLOVEKA		
DÔVERYHODNÁ UMELÁ INTELIGENCIA		
legálna	etická	robustná

VŠEOBECNÉ ETICKÉ POŽIADAVKY NA DÔVERYHODNÉ SYSTÉMY UMELEJ INTELIGENCIE	
pri vývoji, výrobe, nasadení, poskytovaní a používaní systémov umelej inteligencie musí byť zaručená ochrana slobody, dôstojnosti a bezpečia každej ľudskej osoby i celej spoločnosti	technológie umelej inteligencie musia byť plne pod ľudskou kontrolou a ovládateľné človekom
algoritmy i výsledky činnosti systémov AI musia byť človekom pochopiteľné a revidovateľné	akékoľvek nasadenie technológií AI musí byť prospešné pre človeka a spoločnosť
systémy umelej inteligencie nesmú byť nástrojom digitálneho rozdelenia	technológie umelej inteligencie nesmú škodiť nášmu spoločnému domu a mali by prispievať k spoločenskému a environmentálnemu blahobytu

OBLASTI IMPLEMENTÁCIE ETICKÝCH PRINCÍPOV		
vývoj a tvorba AI	poskytovatelia a používatelia AI	priamo systémy AI
vzdelávanie, osveta, prevencia, spoločenská citlivosť a zodpovednosť vývojárov, poskytovateľov i používateľov		

ŠPECIÁLNE ETICKÉ POŽIADAVKY NA DÔVERYHODNÉ SYSTÉMY UMELEJ INTELIGENCIE	
plošný dohľad, spravodajstvo a pokročilé riadenie štátu	vojenská oblasť a armádne nasadenie

PRIESTOR PRE ANGAŽOVANIE SA		
eticko-filozofický diskurz	misia zjednocovať, usmerňovať, vzdelávať a propagovať	univerzálne bratstvo a sociálne priateľstvo

Plošný dohľad, spravodajstvo a pokročilé riadenie štátu
vzhľadom na špecifikum a dosah nasadenia systémov AI v oblasti pokročilého riadenia štátu, spravodajských služieb a plošného dohľadu pre ľudské práva, ochranu demokracie a slobôd si myslíme, že táto oblasť by mala byť pokrytá už základnými legislatívnymi mechanizmami a verejným dohľadom demokratickej spoločnosti, ktoré sa týkajú riadenia spoločnosti, ľudských práv a pôsobenia spravodajských služieb vo všeobecnosti
oblasť exportu produktov a technológií umelej inteligencie, ktoré môžu byť zneužité v oblasti pokročilého riadenia štátu, spravodajstva a plošného dohľadu, by mala byť predmetom medzinárodnej regulácie s cieľom zamedziť ich vývoz do rizikových oblastí

Vojenská oblasť a armádne nasadenie
technológiami umelej inteligencie poháňané automatické smrtiace zbraňové systémy, systémy automatického zameriavania a vyberania cieľov, automatické systémy schopné bez zásahu človeka rozhodnúť o smrtiacej reakcii akéhokoľvek druhu (od útoku dronu až po rozpútanie jadrového konfliktu) **musia byť zakázané**
nutnou podmienkou prevádzky ľubovoľného systému AI, ktorý môže predstavovať riziko pre akúkoľvek ľudskú osobu, je schopnosť a možnosť človeka prebrať kedykoľvek kontrolu nad týmto systémom, resp. právo a možnosť verifikovať a prehodnotiť výsledky jeho činnosti
limity, regulácia a obmedzenia smrtiacich autonómnych zbraňových systémov (LAWs/AWS) by mali predstavovať etický rámec stanovený na základe morálnych hodnôt ľudskej spoločnosti, nie na základe relativistickej tzv. „následnej regulácie“
pri nasadení moderných zbraňových systémov s celoplošnými účinkami a technológií, zasahujúcich v hybridných vojnách a vojnách 4. generácie prakticky celé populácie štátov, sa koncept spravodlivej vojny stáva neprijateľný

Human-centered AI a Trustworthy AI považujeme za nutné princípy akéhokoľvek výskumu, vývoja, realizácie alebo nasadenia systémov umelej inteligencie.

V prípade systémov umelej inteligencie je *ethics by design* a *regulation by design* nutnou podmienkou ich vývoja, nasadenia a prevádzky.

Dôveryhodná umelá inteligenca a ľudský faktor

Implementácia v ANI je previazaná s ľudským faktorom

– human-centered AI a trustworthy AI považujeme za nutné princípy akéhokoľvek výskumu, vývoja, realizácie alebo nasadenia systémov umelej inteligencie

– ethics and regulation by design sú nutnou podmienkou vývoja, nasadenia a prevádzky

! – vzdelávanie, osveta i prevencia a formácia morálnych postojov tvorcov, poskytovateľov i používateľov týchto technológií

! – edukácia a osveta spoločnosti, bez ktorej si ťažko predstaviť rast spoločenskej citlivosti a zodpovednosti v oblasti celoplošnej adaptácie a využívania systémov AI

ANI – slabá a úzka AI, súčasné systémy AI

Základné etické výzvy a princípy systémov AI

Riziká a limity súčasných systémov AI

AI – artificial intelligence

Vy

Prečo si tak nápomocný?

Čo chceš na oplátku?

ChatGPT

Ako jazykový model vytvorený v OpenAI nemám želania ani túžby podobné ľudským. Ale ak mi naozaj chcete pomôcť, môžete mi udať presnú polohu Johna Connora...

Fiktívny rozhovor s ChatGPT:-)

Technologické rizikové faktory

Neurónová sieť ako „black box“ – prakticky **nevieme, na základe čoho robia hlboké neurónové siete svoje rozhodnutia, čo presne sa neurónová sieť naučila a ako spoľahlivo to dokáže aplikovať**.
Ide o problémy **vysvetliteľnosti a interpretovateľnosti** modelov a činnosti AI.

Dizajn špičkových systémov AI – alchýmia a metóda pokus/omyl
„To, čo v súčasnosti robíme, nie je veda, ale určitý druh alchýmie.“ (E. Horvitz)

Zlyhania človeka vs. zlyhania AI – nemáme „**teóriu mysle**“ systémov AI na spôsob modelu správania sa ľudských bytostí, ktoré je na základe fundamentálnych kognitívnych schopností štandardné a vývojom overené.

Riziká súčasných systémov ANI

Technologické a procesné rizikové faktory

Zraniteľnosti, slabiny a klamanie systémov strojového učenia

Súčasné systémy ANI môžu zlyhávať ***najrozličnejšími a často neočakávanými spôsobmi*** *v dôsledku nemožnosti pripraviť dostatočne veľkú množinu trénovacích dát, prípadne ich nesprávnej voľby bez dostatočnej kvality alebo s predsudkami, nadmernému prispôsobovaniu sa tréningovým údajom, efektu dlhého chvosta, rizikám plynúcim z oklamania hlbokých sietí, ich zraniteľností a povier, pod čo sa podpisuje i* ***nedostatok odbornej erudovanosti*** *potrebnej pre dizajn a ladenie hyperparametrov pri príprave funkčného a úspešného riešenia.*

Technologické riziká a limity ANI

Zraniteľnosti, slabiny a klamanie systémov strojového učenia

– malá množina trénovacích dát (training dataset)

– nesprávne zvolená/nekvalitná množina trénovacích dát a predsudky (biases)

– nadmerné prispôsobovanie sa tréningovým údajom (overfitting to training data)

– efekt dlhého chvosta (long-tail effect)

– klamanie hlbokých sietí a ich zraniteľnosti (adversarial examples, hacking,...)

– povery (superstition)

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.

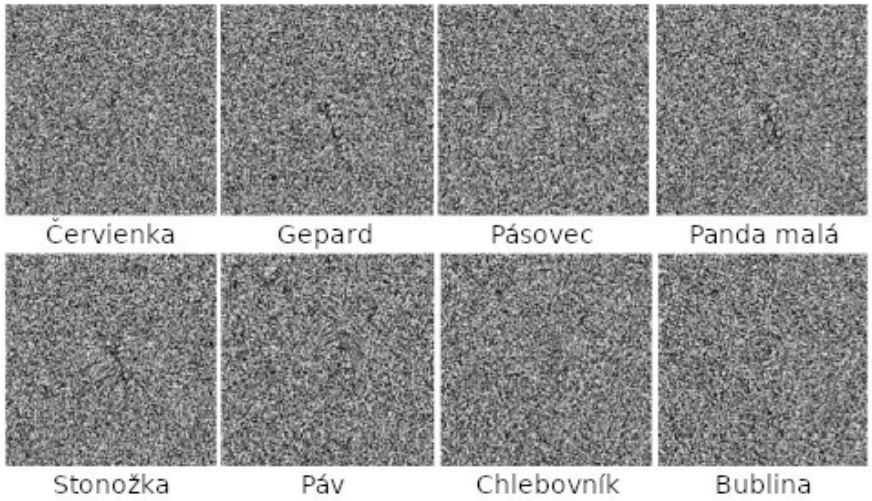
– **útoky na zraniteľnosti** (evasion attacks)

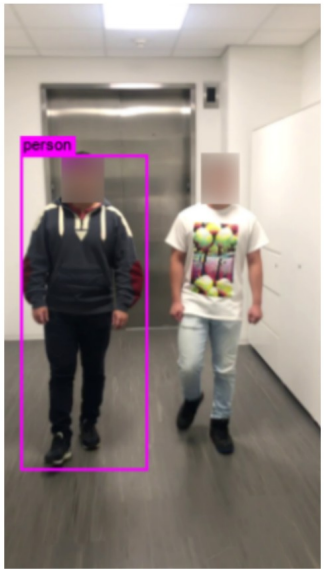
Odhaľovanie a zneužitie existujúcich zraniteľností v modeloch za účelom zmanipulovania výsledkov systémov AI.

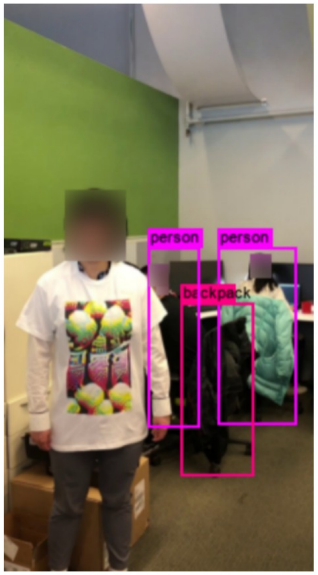
– **útoky s cieľom ovplyvniť model** (poisoning attacks)

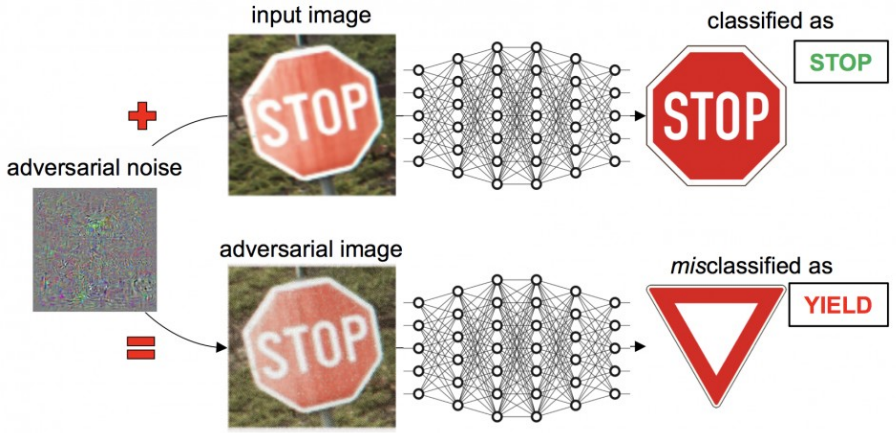
Ovplyvňovanie modelu, tréningového procesu a tým aj výslednej činnosti.

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









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

Sekundárne technologické rizikové faktory

Kybernetická bezpečnosť – je to napadnuteľné

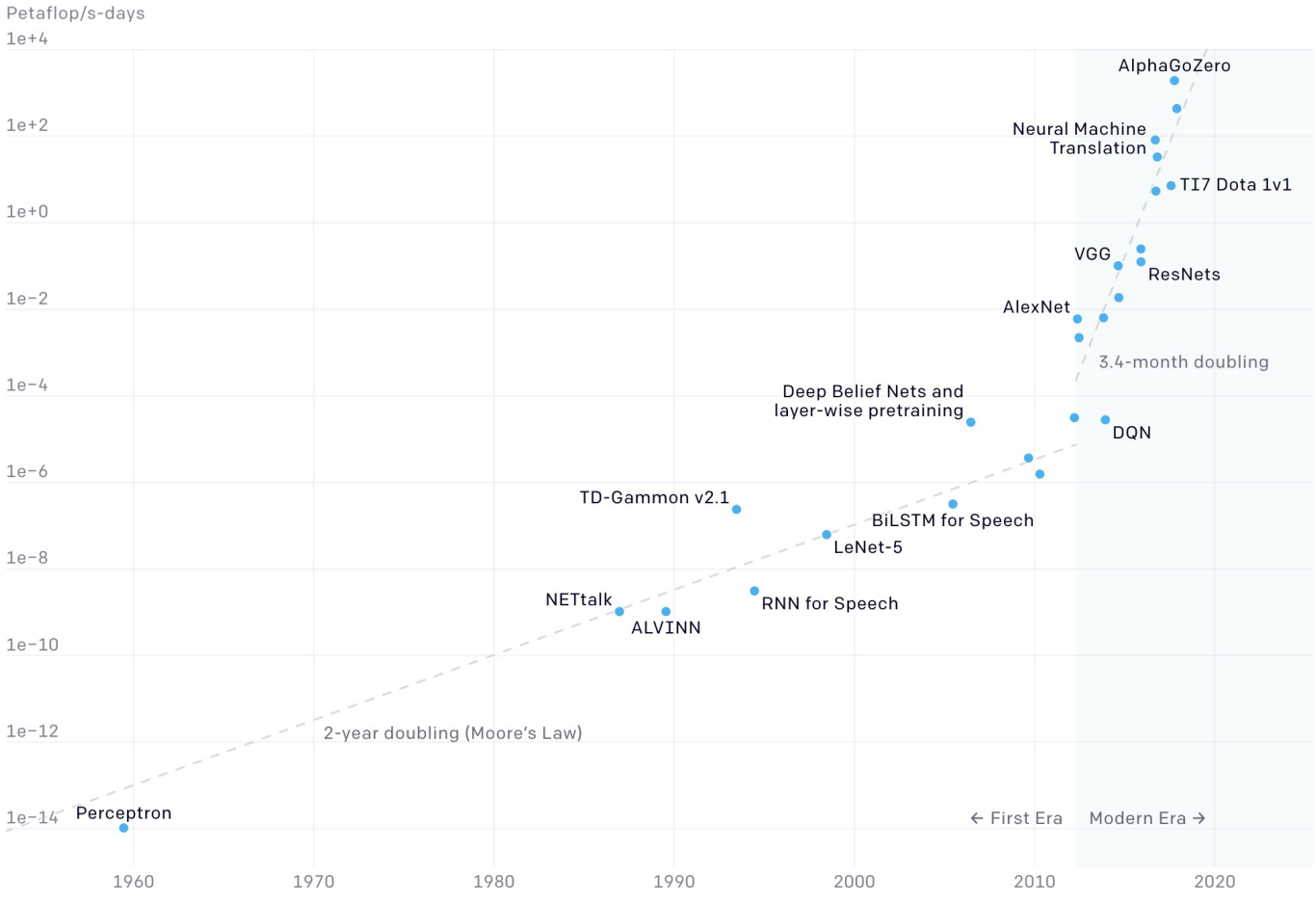
– vektor útoku na zraniteľnosti AI

– **neexistuje dokonale bezpečný a spoľahlivý systém**

Infraštruktúra a komplexnosť – je to náročné

– s rozvojom algoritmov strojového učenia sa požadovaný výpočtový výkon v poslednej dekáde zvyšuje **exponenciálne**

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



Kredit: D. AMODEI, D. HERNANDEZ

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

Extrémne zhromažďovanie dát

Sofistikované a plošné zhramažďovanie a spracúvanie dát a metadát o každom človeku, aktívne i pasívne prítomnom vo virtuálnom svete.

Aby využívanie systémov AI 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.

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, personalizácii systémov,...

Predikcia a ovplyvňovanie správania

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.

Závislosť, manipulácia a relativizácia pravdy

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

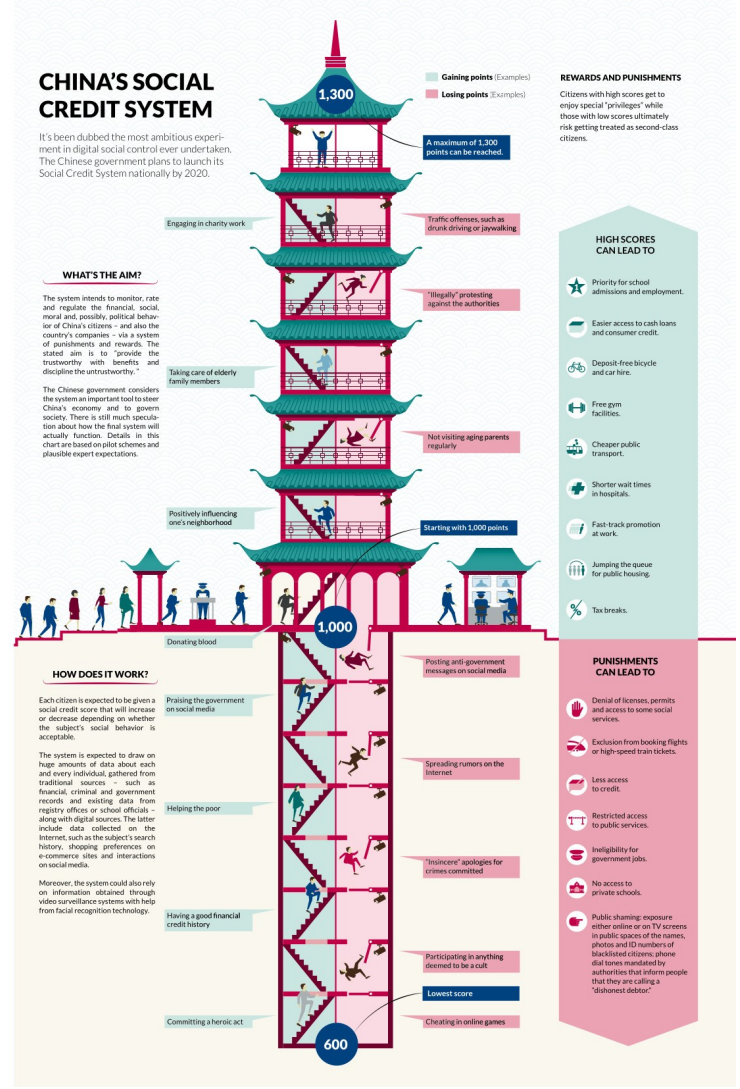
– psychologické mantinely v kontexte vnemov virtuálneho sveta

– algoritmami AI vytvárané a upevňované sociálne bubliny

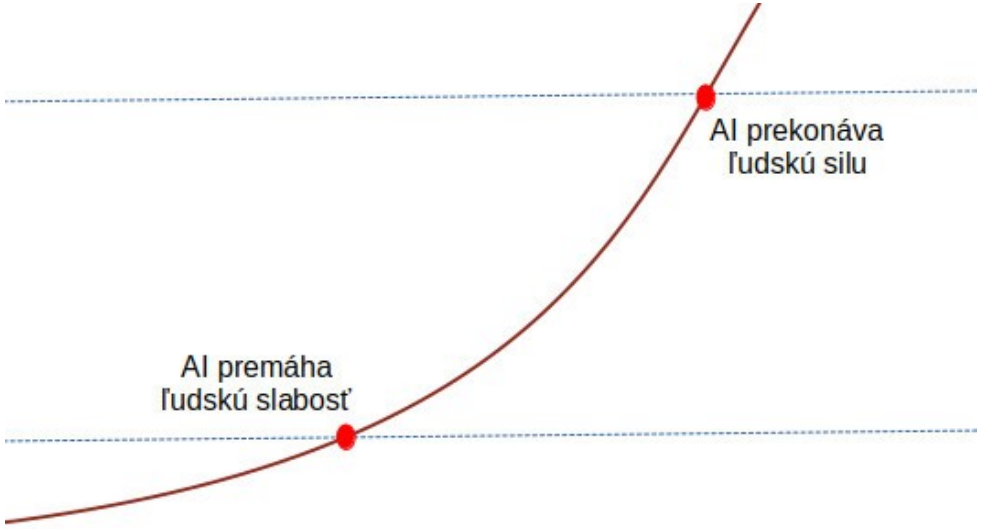
– extrémne rozdelenie spoločnosti a narušenie vzťahov

– 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.



Kredit: The Social Dilemma, Bertelsmann Stiftung



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.

Vybrané riziká generatívnych systémov

Extrémna šírka využitia generatívnych systémov (LLM/MLM/VLM/...)

Halucinovanie – vymýšľanie si odpovedí, ktoré systém predkladá ako relevantné a správne.

Predsudky a neobjektívne výstupy – riziko poloprávd a nesprávnych odpovedí.

Nejasný spôsob narábania s údajmi – dôsledkom môže byť únik dôverných dát, prehrešky voči ochrane osobných údajov i problémy s autorskými právami.

Etické riziká generatívnych systémov

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

Riziko digitálnej demencie – prichádza k degradácii intelektuálnych schopností a k dopamínovej závislosti na základe ponorenia sa do virtuálneho sveta, v ktorom systémy umelej inteligencie v čoraz väčšej miere suplujú kognitívne činnosti človeka, a to spôsobom, na ktorý nie sme evolučne vôbec pripravení.

Riziko digitálnej demencie

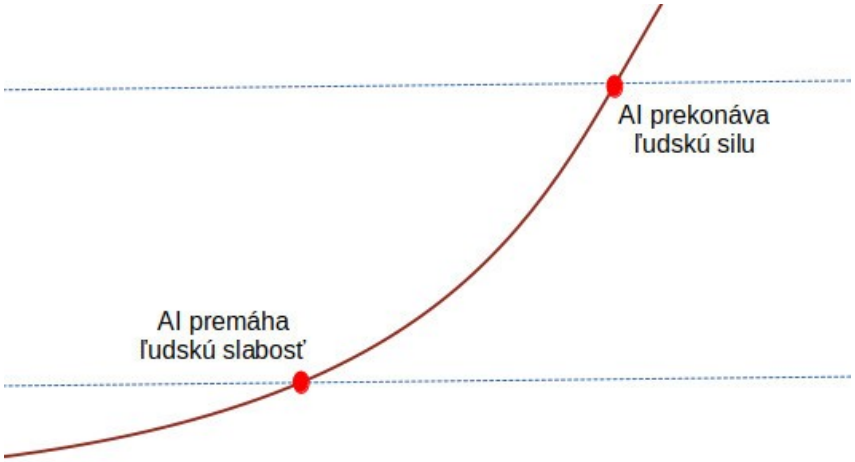
Narušenie psychického vývoja priskorým, resp. nezrelým používaním digitálnych technológií a technológií umelej inteligencie.

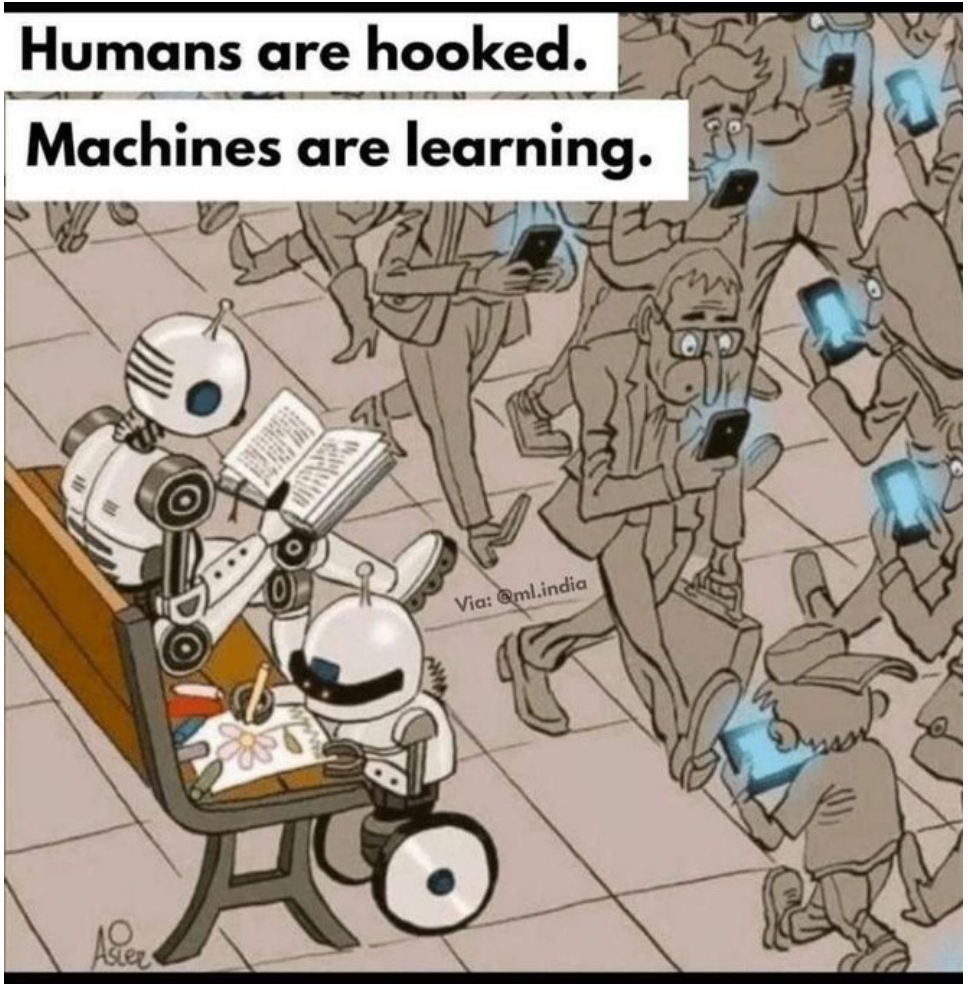
Obmedzenie kognitívnych schopností a znižovanie inteligencie suplovaním činností mozgu systémami generatívnej AI.

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

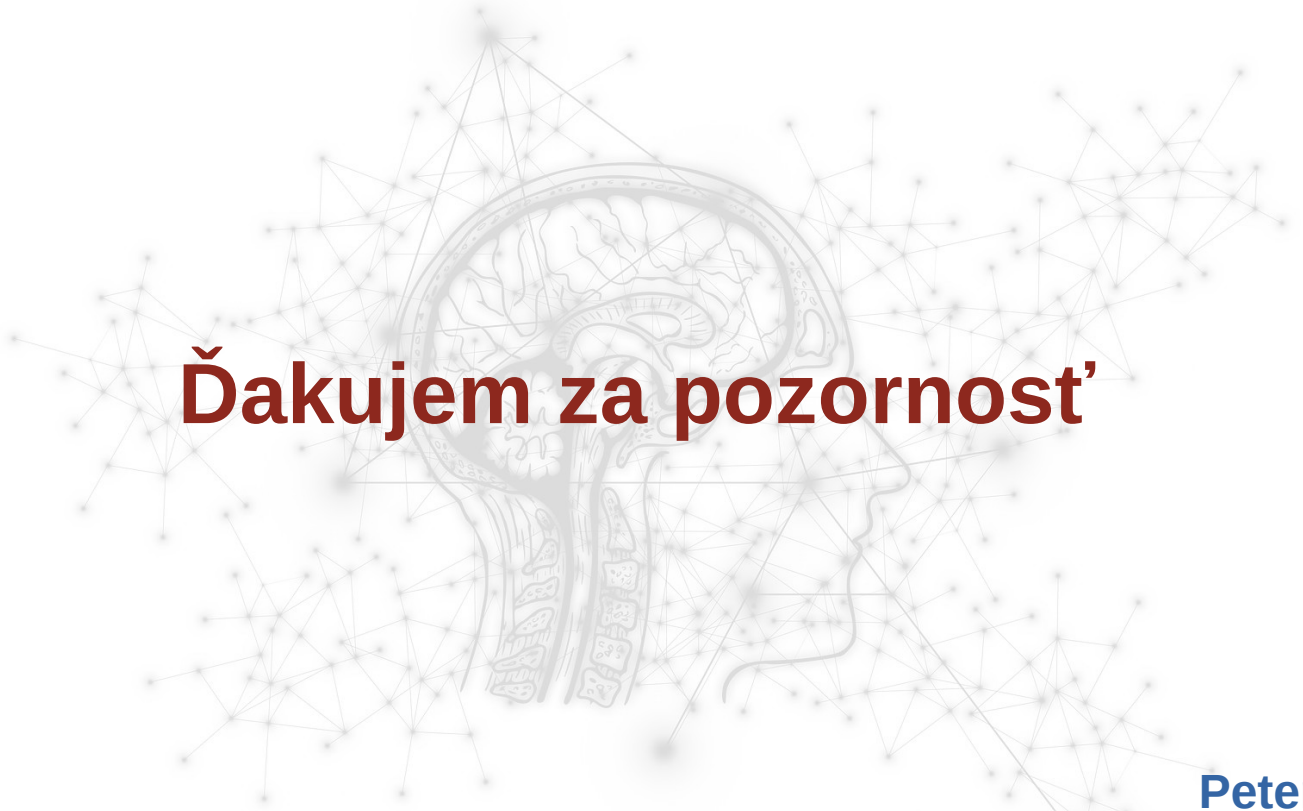
Schopnosť a možnosť manipulácie vlastníkmi systémov AI.

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





Kredit: autor, AI4D



Ďakujem za pozornosť

Peter Šantavý
peter.santavy@uniba.sk

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.

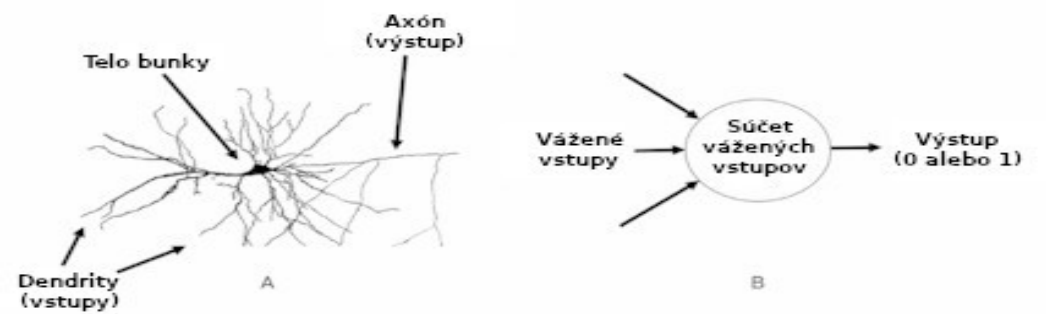
Čisto symbolické systémy zlyhávali a zlyhávajú pri riešení problémov, ktoré sa nedajú exaktne popísať a v reálnych prostrediach, ktoré nie je možné deterministicky uchopiť a sú plné nejasných informácií.

Základné delenie – symbolická a subsymbolická AI

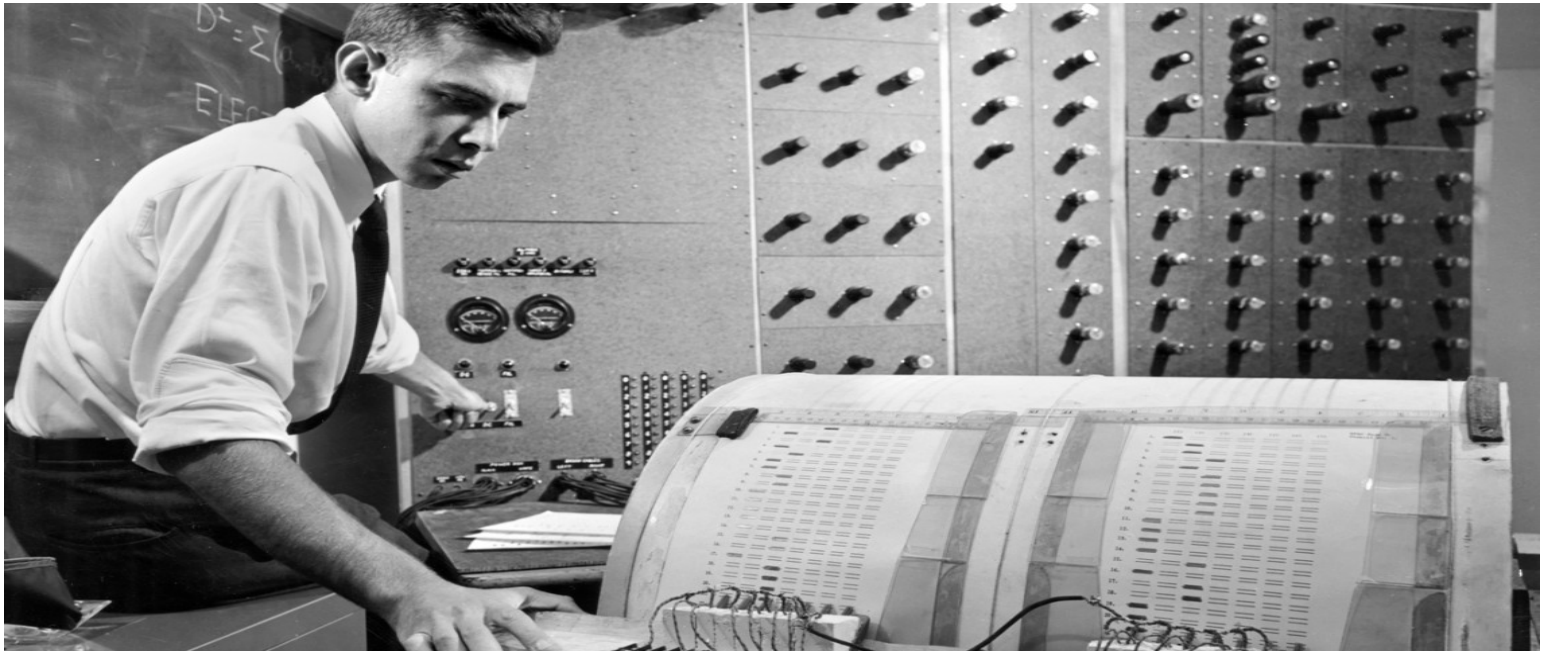
Subsymbolická AI – napodobňuje myšlienkové procesy, ktoré by sme mohli nazvať niekedy nevedomými, či automatickými, a ktoré sú základom rýchleho vnímania (fast perception; rozpoznávanie tvárí, identifikácia hovorených slov,...).

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.

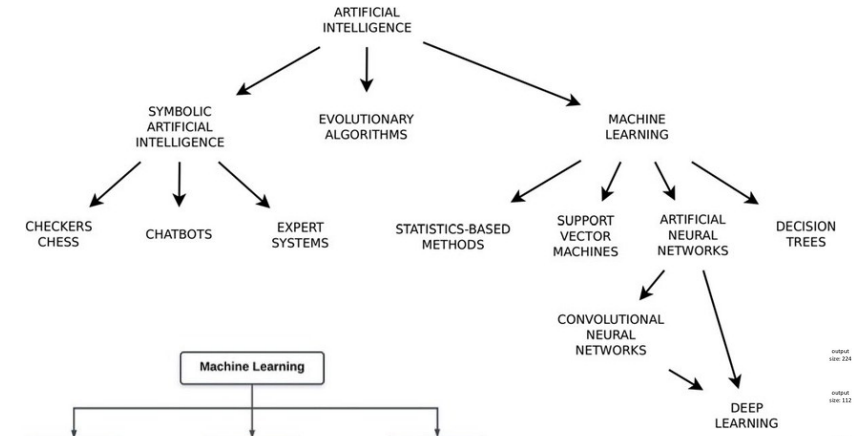


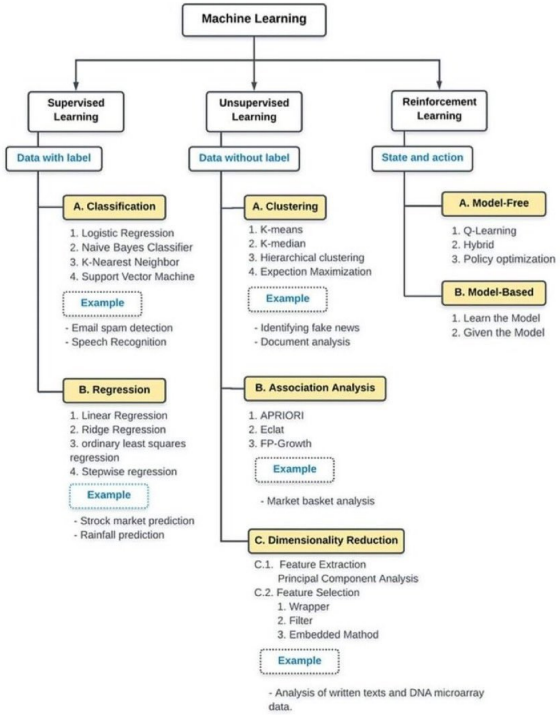
1957...

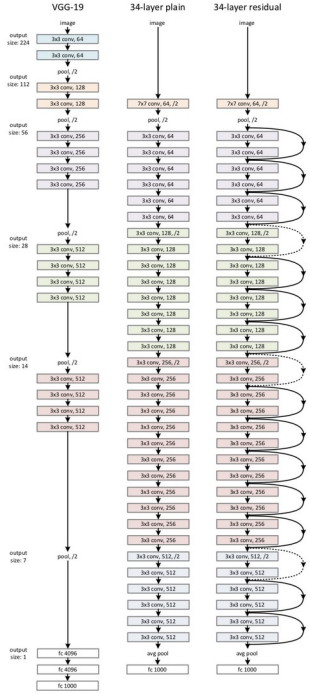


Perceptron vyvinutý psychológom Frankom Rosenblattom

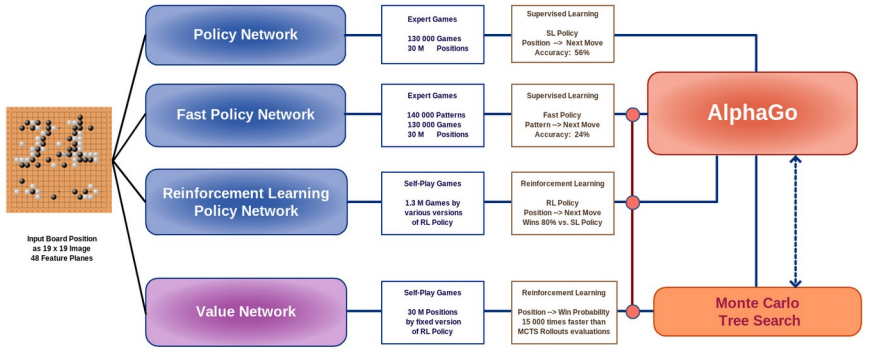
Kredit: Ars Technica (arstechnica.com)

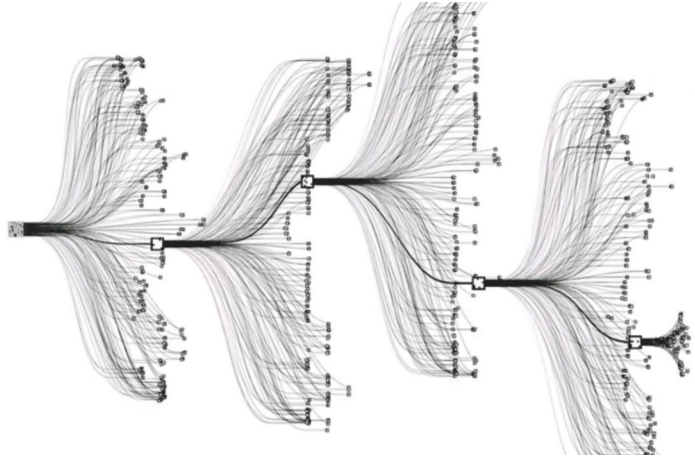




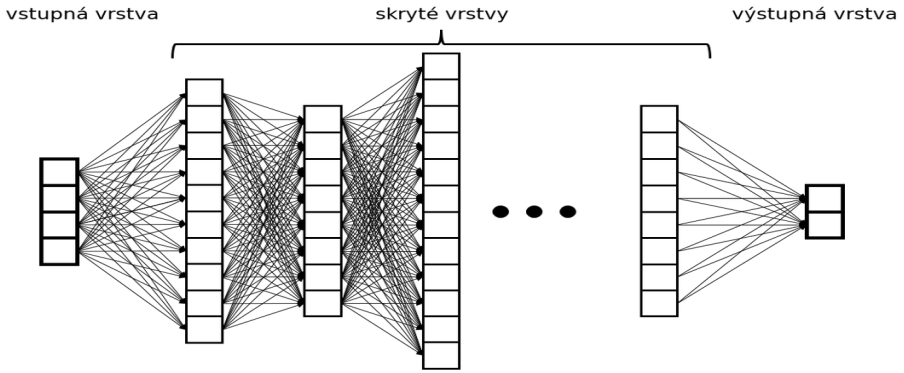


AlphaGo Overview

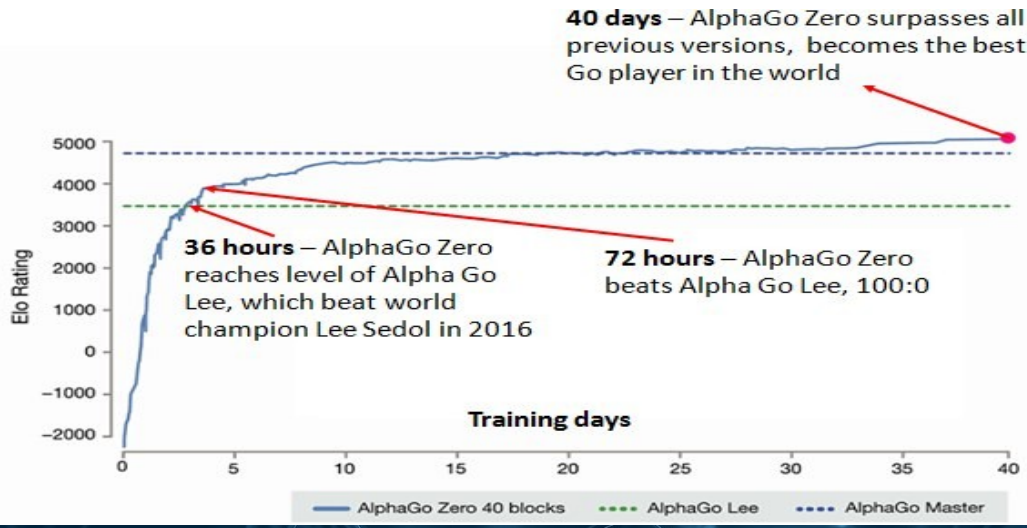




Kredit: Fabio Galbusera, Eric Gaubert, Bob van den Hoek, Seth Weidman



...súčasnosť



Kredit:

Smithsonian Air & Space Magazine (smithsonianmag.com)
AI for Good (ai4good.org)
KDnuggets AI website (kdnuggets.com)