IV130 Přínosy a rizika inteligentních systémů

Etické aspekty a společenské dopady AI

5.&12. května 2023

Silá a slabá AI

- **Slabá AI** jako představa, že stroje mohou jednat *jakoby* byly inteligentní
- **Silná AI** jako předpoklad, že stroje, která tak činí, *skutečně* jsou inteligentní
  (John Searle, 1980)

- Silná AI dnes označovaná také jako „AI lidské úrovně“ nebo „obecná AI“ reprezentující programy, které zvládají libovolně široké druhy úloh včetně nových druhů a daří se jim to tak dobře jako lidem

- Kritika možnosti vyvinutí obecné AI má historické předchůdce: např. Siman Newcomb napsal v říjnu 1903, že „let vzduchem je jednou z velkých tříd problémů, s nimiž si člověk nikdy nemůže poradit“; první let bratří Wrightů následoval jen dva měsíce poté

- Vizionářská práce definující AI a také předvídající námitky proti AI pochází od Alana Turinga z roku 1950: Computing Machinery and Intellgence, Mind, 59, 433-460 (https://doi.org/10.1093/mind/LIX.236.433)

Námitky vůči AI (podle Turinga)

- *Náboženská námitka*: Myšlení jako schopnost nesmrtelné duše, proto nemůže myslet stroj – teologické argumenty ovšem nefungovaly již dříve, např. u Galileiho nebo Koperníka
- *Námitka 'Hlavy v písku'*: "Důsledky myslících strojů by byly příliš závažné, doufejme a věřme proto, že to není možné." – jde o chybný odkaz na důsledky a směšování toho, co nemá nastat s tím, co může nebo nemusí nastat
- *Matematická námitka*: Výsledky neúplnosti jako Gödelova věta o neúplnosti jako argument toho, že stroj založený na logice nedokáže vše – nicméně lidé se také často mýlí
- *Argumentace vědomím*: Vznesen Geoffreyem Jeffersonem roku 1949 jako teze, že stroj "nenapíše sonet a nesloží koncert", protože symboly nenahradí emoce – nicméně nemáme informace o emocích jiných lidí než sami svých (rovněž „čínský pokoj" Johna Searleho z roku 1980)
- *Argumentace různými neschopnostmi ve smyslu "počítač nikdy nedokáže X"*: – bez toho, aby to bylo nějak zdůvodněno; některé konkrétní věci Turing přímo vyvrací
- *Námitka Lady Lovelacové*: Ada Lovelacová soudila, že stroj se nedokáže učit a dělá jen, na co dostane instrukce – stroje ale může přijít na důsledky, které člověku unikly, a překvapovat tímto způsobem
- *Argumentace spojitostí nervového systému*: Mozek nefunguje dikrétně a vzruchy v neuronech se dějí v analogové pravděpodobnosti – což však lze simulovat
- *Argumentace neformálností chování*: Systém podřízený zákonům bude předpověditelný a tedy ne skutečně inteligentní – směšování zákonů s pravidly a pomíjení složitosti systému
- *Mimosmyslové vnímání*: V 50. letech populární – ale Turing argumentuje, že ani čtení mysli nemusí test ovlivnit

Neformálnost chování

- Práce Huberta Dreyfuse kritizující umělou inteligenci, What Computers Can't Do (1972), What Computers Still Can't Do (1992) nebo se Stuartem Dreyfusem, Mind Over Machine (1986), či filozof Kenneth Sayre docházející v roce 1993 k závěru, že "umělá inteligence realizovaná v rámci kultu komputacionalismu nemá žádnou šanci vykázat trvalé výsledky."
- Dobové argumenty se týkají „GOFAI", „Good Old Artificial Intelligence", s problémy zachytit vše pomocí tvrzení predikátové logiky („problém kvalifikace"), což je ale překonaný problém s užíváním pravděpodobnostních systémů

- Další argumentace spočívala v oddělení funkcí mozku od zbytku biologického těla, na což odpovídá koncept aktérů a prostředí a kognitivními funkcemi propojenými se senzory a akcemi
- Pokračující výzkum v AI ukazuje, že se problémy v oblastech, na něž AI neměla odpovědi, daří řešit
- Výzkum v AI vede ke zvětšování schopností, nikoli bariéře nemožnosti

Neschopnost některých činností

- Turing uvádí jako příklady nemožných schopností „být milý, vynalézavý, krásný, přátelský, mít iniciativu, mít smysl pro humor, rozlišovat dobro od zla, dělat chyby, zamilovat se, užívat si jahody se šlehačkou, přimět někoho, aby se do něj zamiloval, učit se ze zkušeností, správně používat slova, být předmětem vlastního myšlení, mít stejně rozmanité chování jako člověk, dělat něco opravdu nového“

- Řada z těchto věcí je snadná, víme např. za zkušenosti, že „počítače dělají chyby“
- Metavyvozovací techniky umožňují popisovat vlastní činnost stroje a tedy mít stroj jako předmět vyvotování
- Schopnost „přimět někoho, aby se do něj zamiloval“ dokážou i mnohem jednodušší technické artefakty, např. dětské hračky
- Počítače dnes dělají „opravdu něco nového“: objevy v astronomii, fyzice, matematice, informatice, biologii, chemii, atd.
- Zjevně stroje nedokážou být přesně jako lidé, ale v mnoha oblastech již lidi dokážou svými činnostmi předčít

Měření pokroku v AI

- Turing navrhl v článku z roku svůj známý „Turingův test“
- Proporce úspěšných reakcí v tomto testu by mohla sloužit jako míra úspěchu (Turing předpokládal, že kolem roku 2000 budou počítače s miliardou jednotek paměti v takovém testu úspěšné)
- Program Eliza (Weizenbaum, 1964-66, pattern matching částí textu s generováním reakcí bez analýzy významu, viz https://sites.google.com/view/elizagen-org/) dokázal přimět jejich uživatele k pocitu, že mluví s psychologem, podobně Mgonz (2008) nebo Natachata (2009), bot Cyberlover (cca 2007) dokonce dokázal získat důvěru komunikujících a kradl osobní data

- Skutečné testy se na bázi Turingova testu nedělají, hra v šachy, Go nebo počítačové hry se pro porovnání schopností užívá (namísto schopnosti oklamat posuzovatele)

- IBM Watson zvládl v roce 2011 znalostní hru *Jeopardy!* lépe než lidští hráči na bázi big data; obdobně strojová AI zvládá Go, šachy, poker a řadu počítačových her, obdobně pro lékařskou diagnostiku karcinomů kůže, skládání proteinů, atd.

Mohou stroje skutečně myslet?

- Edsger Dijkstra (1984): „otázka, zda mohou stroje myslet, … je asi stejně relevantní, jako otázka, zda ponorky dokážou plavat.“
- (Slovník uvádí jako první význam „*vhodnými vlastními pohyby se udržovat a pohybovat ve vodě*“, což ve smyslu pohybu končetin zřejmě nebude splněno; obdobně „*létat*“ ve smyslu „*pohybovat se ve vzduchu vlastní schopností (o tvorech), vlastní silou (o předmětech)*“, což platí pro letadla, ale obojí je irelevantní vzhledem ke konstrukci ponorek nebo letadel.)
- Turing upozorňuje na to, že nemáme žádné podklady pro tvrzení o schopnosti myslet jiných lidí než nás samých – je podle něj věcí *zdvořilé konvence* o ostatních kolem nás předpokládat, že myslí, a tuto zdvořilou koonvencvi můžeme rozšířit i na stroje (spojení se vzhledem či hlasem je věcí našich předsudků)
- Searlův *čínský pokoj* (1990) je založen na člověku reagujícím na čínské znaky pomocí knihy pravidel, nikoli schopnosti číst znaky
- Searle (1980) propaguje biologický naturalismus: mentální stavy jsou dány procesy nižší úrovně, přičemž neurony „to“ mají, zatímco tranzistory nikoli; obdobně si ale roboti mohou myslet o lidech, že jsou z buněk, přičemž buňky nemohou chápat
- Také sci-fi Terryho Bissona (1990): „They are Made Out of Meat“ o mimozemských robotech zkoumajících Zemi (https://en.wikipedia.org/wiki/They're_Made_Out_of_Meat)

Vědomí a kvalia

- Vědomí je problém povědomí o okolním světě a také subjektivní zkušenosti vlastního života
- „Kvalia“ jako označení vnitřní přirozené zkušenosti

- Relevantní otázkou je, zda mohou stroje mít kvalia
- Obdobná otázka existuje pro zvířata
- Koncept myšlení lze rozšiřovat i na veškerou hmotu, tzv. *panpsychismus* (v moderní podobě např. David Chalmers nebo Philip Goff)

- Turing (1950): „*Nechci vzbudit dojem, že si myslím, že vědomí není tajemství... Nemyslím si však, že tyto záhady musí být nutně vyřešeny, abychom mohli odpovědět na otázku, kterou se zabýváme v tomto článku*.“

- Obecně není záměrem AI dělat stroje, jejichž myšlení přesně odpovídá myšlení člověka – a není účelem strojů např. popsat pocit, když se praštíme kladivem do prstu při zatloukání hřebíku

Etika AI

- AI jako mocná technologie vyvinutá člověkem, s morální povinností člověka dobře ji užívat, podporovat její dobré vlastnosti a omezovat vlastnosti špatné/nežádoucí
- AI může zachraňovat životy díky
  - lepší lékařské diagnostice,
  - novým lékařským objevům,
  - lepší předpovědi extrémních povětrnostních jevů nebo
  - bezpečnějšímu řízení díky asistenčním a (případně) samořídicím technologiím.
- AI může zlepšovat životy, např.
  - Program AI for Humanitarian Action (Microsoft) využívá AI při odstraňování následků přírodních katastrof, řešení potřeb dětí, ochraně uprchlíků a podpoře lidských práv,
  - Program AI for Social Good (Google) podporuje práci v oblasti ochrany deštných pralesů, judikatury v oblasti lidských práv, monitorování znečištění, měření emisí fosilních paliv, krizového poradenství, ověřování pravdivosti zpráv, prevence sebevražd, recyklace a dalších otázek.
  - Centrum datové vědy pro společenskou prospěšnost Chicagské univerzity aplikuje strojové učení trestní soudnictví, ekonomický rozvoj, vzdělávání, veřejné zdraví, energetiku a životního prostředí.
  - Aplikace AI v oblasti správy plodin a produkce potravin pomáhají uživit svět.
  - Optimalizace obchodních procesů pomocí strojového učení zvyšují produktivitu podniků, zvyšují bohatství a zajistí více pracovních míst
  - Automatizace může nahradit únavné a nebezpečné úkoly, s nimiž se potýká mnoho pracovníků, a uvolnit je, aby se mohli soustředit na zajímavější aspekty
  - Handicapovaní mohou dostat asistenci v oblasti zraku, mluvení nebo pohybu
  - Strojový překlad umožňuje komunikaci lidí z různých kultur, atd.

Negativní dopady technologií

- Negativní boční efekty, např.
  - Dopady havárií jaderných elektráren typu Černobylu nebo Fukušimy
  - Znečištění ovzduší z užívání spalovacích motorů
  - Globální oteplování
  - Dopady staveb/konstrukcí na životní prostředí
- Dopady mohou být negativní i přio užívání ve shodě s primárním účelem:
  - Střelné zbraně
  - Telemarketing a nevyžádaná reklamní sdělení
- Automatizace přináší zvýšenou míru zisku, ale tok směrem k vlastníkům zvyšuje ekonomické nerovnosti, pro rozvojové země se mohou zmenšovat příležitosti využívat levnější pracovní sílu
- Naše rozhodnutí stran etiky užití a správy technologií ovlivní úroveň nerovností, kterou může AI přinést
- Obvyklé obecné požadavky (Principy robotiky z Velké Británie):

- Zajistit bezpečnost
- Zavést odpovědnost
- Zajistit spravedlnost
- Dodržovat lidská práva a hodnoty
- Respektovat soukromí
- Odrážet rozmanitost/inkluzi
- Podporovat spolupráci
- Vyhnout se koncentraci moci
- Zajistit transparentnost
- Uznat právní/politické důsledky
- Omezit škodlivé využití AI
- Zvážit důsledky pro zaměstnanost

Smrticí autonomní zbraně

- Autonomní zbraň podle definice OSN lokalizuje, vybírá a zasahuje (zabíjí) lidské cíle bez lidského dohledu
- Některá z těchto kritérií splňují různé zbraně, např.
  - nášlapné miny se používají již od 17. století: mohou v omezeném smyslu vybírat a zasahovat cíle podle stupně vyvíjeného tlaku nebo množství přítomného kovu, ale nemohou samy vyrazit a lokalizovat cíle (pozemní miny jsou zakázány Ottawskou smlouvou)
  - řízené střely, používané od 40. let 20. století, mohou pronásledovat cíle, ale musí být zaměřeny správným obecným směrem člověkem
  - od 70. let 20. století se k obraně námořních lodí používají automaticky střílející děla řízená radarem; jsou určena především k ničení přilétajících raket, ale mohla by útočit i na letadla s posádkou.
- Slovo "autonomní" se často užívá pro označení bezpilotních vzdušných prostředků nebo dronů, většina takových zbraní je jednak dálkově řízena, jednak vyžaduje ovládání smrtícího nákladu člověkem.

Autonomní zbraně

- Izraelská střela Harop je „létající munice“ s rozpětím křídel 3 m a padesátikilovou hlavicí; až 6 hodin hledá v dané zeměpisné oblasti jakýkoli cíl, který splňuje dané kritérium, a poté jej zničí (kritériem může být např. „vysílá radarový signál připomínající protiletadlový radar“ nebo „vypadá jako tank“).
- Turecká STM inzeruje kvadrokoptéru Karga – která unese až 1,5 kg výbušnin – jako schopnou „autonomně zasáhnout… cíle vybrané na snímcích… sledování pohyblivých cílů… protipěchotní… rozpoznávání tváří“
- Autonomní zbraně označovány za „třetí revoluci ve válčení“ po střelném prachu a jaderných zbraních
- Autonomní letadla, tanky a ponorky mohou být levnější, rychlejší, manévrovatelnější a mít delší dolet než jejich protějšky s lidskou posádkou.

Regulace autonomních zbraní

- Od 2014 vede OSN v Ženevě pod záštitou Úmluvy o některých konvenčních zbraních (CCW) pravidelná jednání o otázce, zda zakázat smrtící autonomní zbraně
- Cca 30 států, od Číny po Vatikán, je pro mezinárodní smlouvu,
- Další klíčové země – mj. Izrael, Rusko, Jižní Korea a Spojené státy - jsou proti zákazu
- Debata zahrnuje právní, etické a praktické aspekty.
- Právní otázky podle CCW vyžadují rozlišování mezi bojovníky a nebojovníky, posouzení vojenské nezbytnosti útoku a posouzení proporcionality mezi vojenskou hodnotou cíle a možností vedlejších škod (splnění kritérií je technickou otázkou, jejíž odpověď se bude v průběhu času měnit; v současné době se zdá, že diskriminace je za určitých okolností proveditelná a nepochybně se bude rychle zlepšovat, ale nezbytnost a přiměřenost dnes proveditelné nejsou: stroje by musely provádět subjektivní a situační posouzení, obtížnější než relativně jednoduché úkoly vyhledávání a zasahování potenciálních cílů)

Podřízení autonomních zbraní člověku

- Argumenty pro omezení legálního užívání autonomních zbraní pouze za okolností, kdy lidský operátor může rozumně předvídat, že provedení mise nepovede k tomu, že se cílem stanou civilisté nebo že zbraně provedou zbytečné či nepřiměřené útoky
- Autonomní zbraně by tak zatím mohly plnit pouze velmi omezené mise
- U etické stránky existují názory o morální nepřijatelnosti delegovat rozhodnutí o zabíjení lidí na stroj:
  - německý velvyslanec v Ženevě prohlásil, že Německo „nepřijme, aby o životě a smrti rozhodoval výhradně autonomní systém“,
  - Japonsko „nemá v plánu vyvíjet roboty schopné vraždit bez lidí v rozhodování“,
  - generál Paul Selva, v té době druhý nejvyšší vojenský důstojník ve Spojených státech, v roce 2017: „Nemyslím si, že je rozumné, abychom pověřovali roboty rozhodováním o tom, zda odejmeme lidský život, či nikoli.“
  - António Guterres, šéf OSN, v roce 2019: „stroje s pravomocí a volností brát životy bez účasti člověka jsou politicky nepřijatelné, morálně odporné a měly by být zakázány mezinárodním právem“.
- >140 NGO ve více než 60 zemích je v Kampani za zastavení zabijáckých robotů; otevřený dopis Future of Life Institute z 2015 má >4 000 podpisů z AI i 22 000 dalších osob (https://www.stopkillerrobots.org)

Argumenty pro autonomní zbraně

- Se zdokonalováním technologií by mělo být možné vyvinout zbraně, u nichž je menší pravděpodobnost, že způsobí civilní oběti než lidští vojáci nebo piloti (autonomní zbraně snižují potřebu lidských vojáků a pilotů riskovat smrt)
- Autonomní systémy nebudou podléhat únavě, frustraci, hysterii, strachu, hněvu nebo pomstě a nemusí „nejdříve střílet, potom se ptát“ (Arkin, 2015).
- Řízená munice snížila vedlejší škody ve srovnání s neřízenými bombami, inteligentní zbraně dále zvýší přesnost útoků (analýza obětí války s drony z roku 2013 ale naznačuje idealismus takového pohledu).
- Americké ministerstvo obrany (DOD) z roku 2011: „V dohledné budoucnosti bude rozhodování o použití síly [autonomními systémy] a volba jednotlivých cílů, na které bude použita smrtící síla, ponechána pod lidskou kontrolou“ (autonomní systémy nejsou dostatečně spolehlivé, aby jim bylo možné svěřit vojenská rozhodnutí)

Další aspekty autonomních zbraní

- Systémy strojového učení, které při výcviku fungují bezchybně, mohou při nasazení fungovat špatně
- Kyberútoky na autonomní zbraně by mohly vést ke ztrátám při přátelské střelbě; odpojení zbraně od veškeré komunikace tomu může zabránit (za předpokladu, že již nebyla kompromitována), ale pak nelze zbraň v případě poruchy odvolat
- Praktickým problémem autonomních zbraní je, že se jedná o škálovatelné zbraně hromadného ničení: rozsah útoku, který lze provést, je úměrný množství hardwaru, který si lze dovolit nasadit
- Kvadrokoptéra o průměru 5 cm může nést smrtící výbušnou nálož, jeden milion se vejde do přepravního kontejneru
- Autonomie zbraní nepotřebovaly milion lidských dozorců
- Jako zbraně hromadného ničení mají škálovatelné autonomní zbraně ve srovnání s jadernými zbraněmi a kobercovým bombardováním pro útočníka výhody: ponechávají majetek nedotčený a lze je použít selektivně k likvidaci pouze těch, kteří by mohli ohrozit okupační síly. Nebezpečím je možné užití k vyhlazení celé etnické skupiny nebo všech vyznavačů náboženství; často by je také nebylo možné vystopovat.

AI jako technologie dvojího užití

- Vlastnosti, které zvýhodňují útočníka, naznačují, že autonomní zbraně sníží globální a národní bezpečnost všech stran
- Racionální reakcí vlád je spíše zapojení se do diskusí o kontrole zbrojení než do závodů ve zbrojení
- Proces přípravy smlouvy je ovšem komplikovaný
- AI je technologie *dvojího užití*:
  - technologie AI, které mají mírové využití, jako je řízení letu, vizuální sledování, mapování, navigace a multiaktérové plánování, mohou být snadno použity pro vojenské účely,
  - Autonomní kvadrokoptéru lze snadno proměnit ve zbraň pouhým připojením výbušniny a vydáním příkazu k vyhledání cíle
- Řešení problému bude vyžadovat pečlivé zavedení režimů dodržování předpisů ve spolupráci s průmyslem, podobně jako již bylo s určitým úspěchem zavedeno v rámci Úmluvy o zákazu chemických zbraní.

Sledování, bezpečnost a soukromí

- Joseph Weinzenbaum již roku 1976 varoval před užitím Ai rozpoznávání řeči pro odposlechy a potlačování lidských práv
- V roce 2021 rozšíření *dozorových kamer*:
- v Číně 567 milionů (1 kamera na 4,1 obyvatel)
- v USA 85 milionů (1 kamera na 4,6 obyvatel)
- celosvětově cca 1 miliarda (https://www.wsj.com/articles/a-billion-surveillance-cameras-forecast-to-be-watching-within-two-years-11575565402)
- Čína je exportérem této technologie zejména do rozvojových zemí
- On-line instituce zranitelné vůči kyberkriminalitě (phishing, podvody s kreditními kartami, botnety, ransomware) a kyberterorismu (včetně potenciálně smrtících útoků, jako je odstavení nemocnic a elektráren nebo ovládnutí samořiditelných aut).
- Strojové učení významné pro obě strany v boji o kyberbezpečnost:
- útočníci využijí automatizaci ke zkoumání nejistot a mohou použít posilující učení pro pokusy o phishing a automatické vydírání,
- obránci mohou využít neřízené učení k detekci anomálních vzorců příchozího provozu a techniky strojového učení k odhalování podvodů

(https://cset.georgetown.edu/wp-content/uploads/Machine-Learning-and-Cybersecurity.pdf)

Ochrana soukromí s rozhodovací pravomocí strojů

- Pravomoc strojů/AI nad lidmi by nás odsouvala do druhořadé pozice a ke ztrátě práva podílet se na rozhodnutích, která se nás dotýkají
- I když to nejsou stroje, kdo rozhoduje, ale lidé, kteří tyto stroje sestrojili a uvedli do provozu, je potřebné zvažovat jednotlivé okolnosti každého lidského subjektu (jinak by tvůrci životům ostatních přikládali jen nepatrnou hodnotu)
- Riziko velkého odcizení mezi elitami, kterým slouží lidé, a početnou nejnižší třídou, již obsluhují a ovládají stroje

- V Evropské unii článek 22 Obecného nařízení o ochraně osobních údajů (General Data Protection Regulation; GDPR) z roku 2018 výslovně zakazuje poskytovat v takových případech pravomoc strojům:
  - Subjekt údajů má právo nebýt předmětem žádného rozhodnutí založeného výhradně na automatizovaném zpracování, včetně profilování, které má pro něho nebo ji právní účinky nebo se ho nebo jí obdobným způsobem významně dotýká.

Spravedlnost a podjatost

- Mechanismy AI a strojového učení mohou vstupovat do přípravy podkladů pro důležitá rozhodování, např. oprávnění dostat půjčku nebo hypotéku
- *Společenská podjatost* může být nezamýšleným důsledkem – např. algoritmy rozhodující o poskytnutí kauce před procesem se stíhanými na základě socioekonomické nebo rasové příslušnosti, efekty zónování při poskytování hypoték v USA, atd.
- Požadavky kalibrace systémů z hlediska:
  - Individuální spravedlnosti pro vyloučení efektů přináležitosti k nějaké skupině
  - Skupinové spravedlnosti pro vyloučení různého nakládání se skupinami/třídami
  - Spravedlnosti přes nevědomost– s nutností vyloučení predikce zamlčeného parametru při učení
  - Demografické parity s rovnocenným genderovým zacházením
  - Stejných příležitostí vylučující preferenci těch, kdo mají stejné předpoklady
  - Stejných dopadů poskytujících stejný užitek bez ohledu na skupinu, se zvážením přínosů i ceny za chybnou predikci

Praxe vytváření spravedlivých systémů

- Ujistěte se, že softwaroví inženýři hovoří se sociálními vědci a odborníky na danou oblast, aby porozuměli problémům a perspektivám, a zvažte spravedlnost od samého počátku.
- Vytvořte prostředí, které podporuje rozvoj různorodé skupiny softwarových inženýrů, kteří jsou reprezentativními zástupci společnosti.
- Definujte, jaké skupiny bude váš systém podporovat: různé jazykové mluvčí, různé věkové skupiny, různé schopnosti se zrakem a sluchem atd.
- Optimalizujte pro cílovou funkci, která zahrnuje spravedlnost.
- Prozkoumejte svá data z hlediska předsudků a korelací mezi chráněnými atributy a jinými atributy.
- Pochopte, jak probíhá případná lidské označování dat, navrhněte cíle pro přesnost označování a ověřte, zda jsou cíle splněny.
- Nesledujte pouze celkové metriky pro váš systém; ujistěte se, že sledujete metriky pro podskupiny, které by mohly být obětí předpojatosti.
- Zahrňte testy systému, které odrážejí zkušenosti uživatelů z menšinových skupin.
- Zajistěte zpětnou vazbu, aby se v případě výskytu problémů se spravedlností tyto problémy řešily.

Důvěra

- Jedna věc je vytvořit přesný, spravedlivý, bezpečný a zabezpečený systém AI,
- jiná věc je přesvědčit ostatní, že se to podařilo
- Lidé musí mít možnost důvěřovat systémům, které používají.
- Průzkum společnosti PwC z roku 2017 ukázal, že 76 % podniků zpomaluje zavádění AI kvůli obavám o *důvěryhodnost*.
- Kromě inženýrských přístupů k důvěryhodnosti zde jde o politické otázky.
- Aby si inženýrský systém zasloužil důvěryhodnost, musí projít procesem *verifikace* a *validace* (V&V).
- *Verifikace* znamená, že produkt splňuje specifikace.
- *Validace* znamená zajistit, aby specifikace skutečně splňovaly potřeby uživatele a dalších dotčených stran.
- Existují metodiky V&V pro inženýrství obecně a pro tradiční vývoj softwaru prováděný lidskými programátory; většina z ní je použitelná pro systémy umělé inteligence.
- Systémy strojového učení však vyžadují jiný proces V&V, který dosud nebyl plně vyvinut. (potřeba ověřit data, ze kterých se tyto systémy učí; potřebujeme ověřit přesnost a spravedlnost výsledků, a to i v případě nejistoty, která znemožňuje znát přesný výsledek; a potřebujeme ověřit, že protivníci nemohou model nepatřičně ovlivnit ani ukrást informace dotazováním na výsledný model)
- Jedním z nástrojů důvěry je *certifikace*.

Transparentnost

- Dalším aspektem důvěry je *transparentnost*: spotřebitelé chtějí vědět, co se uvnitř systému děje a že systém nepracuje proti nim, ať už v důsledku úmyslného zlého úmyslu, neúmyslné chyby nebo rozšířených společenských předsudků, které systém rekapituluje.
- Někdy je transparentnost poskytována přímo spotřebiteli,
- jindy se jedná o otázky duševního vlastnictví, kvůli nimž jsou některé aspekty systému spotřebitelům skryté, ale jsou přístupné regulačním orgánům a certifikačním agenturám.
- Když např. systém umělé inteligence zamítne žádost o půjčku, zaslouží si klient vysvětlení.

- V EU toto vymáhá nařízení GDPR
- O direktivě GDPR se často říká, že poskytuje obecné „právo dostat vysvětlení“ libovolného automatizovaného rozhodnutí, ale formulace v článku 14 požaduje jen
  - Smysluplné informace týkající se použitého postupu, jakož i významu a předpokládaných důsledků takového zpracování pro subjekt údajů.

- Při podávání vysvětlení bude zásadní budoucí postoj soudů: Spotřebitel by např. neměl dostat jen popis příslušného algoritmu hlubokého učení, pomocí něhož se trénoval klasifikátor, který ono rozhodnutí udělal.

Transparentnost

- Systém umělé inteligence, který dokáže vysvětlit sám sebe, se nazývá *vysvětlitelná umělá inteligence* (XAI).
- Dobré vysvětlení má několik vlastností:
  - Mělo by být pro uživatele srozumitelné a přesvědčivé,
  - mělo by přesně odrážet uvažování systému,
  - mělo by být úplné a
  - mělo by být specifické v tom smyslu, že různí uživatelé s různými podmínkami nebo různými výsledky by měli dostat různá vysvětlení.

- Součástí transparentnosti je vědomí, zda komunikujete se systémem AI, nebo s člověkem
- Toby Walsh (2015) navrhl, že "autonomní systém by měl být navržen tak, aby bylo nepravděpodobné, že bude zaměněn za něco jiného než autonomní systém, a měl by se identifikovat na začátku každé interakce“ (zákon „červené vlajky“ na počest britského zákona o lokomoci z roku 1865 (Locomotive Act), který vyžadoval, aby před každým motorovým vozidlem chodila osoba s červenou vlajkou, která signalizovala blížící se nebezpečí)

- V roce 2019 přijala Kalifornie zákon, který stanoví, že „je nezákonné, aby jakákoli osoba používala bota ke komunikaci nebo interakci s jinou osobou v Kalifornii online s úmyslem uvést druhou osobu v omyl ohledně své umělé identity“.

Technologická nezaměstnanost

- Obvyklým tématem je, že roboti berou lidem práci
- Knihy věnované tomuto tématu:
- Martin Ford: *Roboti nastupují: Automatizace, umělá inteligence a hrozba budoucnosti bez práce* a
- Calum Chace: The Economic Singularity: Artificial Intelligence and the Death of Capitalism (Ekonomická singularita: Umělá inteligence a smrt kapitalismu)
- Článek John Maynard Keynes (1930): „Ekonomické možnosti pro naše vnuky“ (z doby, kdy v Británii velká hospodářská krize vyvolala masovou nezaměstnanost, kterou popisoval jako „dočasné stadium nepřizpůsobivosti“ způsobené „nárůstem technické efektivity“, který se odehrál „rychleji, než se s problémem absorpce pracovních sil dokážeme vypořádávat“) s předpovědí:
  - „Poprvé od svého stvoření proto bude člověk konfrontován se skutečným, trvalým problémem – jak užívat svou nezávislost na naléhavých ekonomických starostech, svobodu, jak vyplnit volný čas, jež mu dobyla věda a složené úročení, jak žít moudře, konsensuálně a dobře.“
- Idea souvislosti technologií a zaměstnanosti je obsažena již u Aristotela v Knize I jeho Politiky:
  - „Neboť kdyby každý nástroj na rozkaz nebo již předem dovedl vykonati své dílo, [...] kdyby tak člunky samy od sebe tkaly a paličky hrály na kitharu, nepotřebovali by stavitelé pomocníků ani páni otroků.“

Budoucnost práce

- Pokud zaměstnavatel najde mechanický způsob, jak provádět práci, kterou dříve dělal, dochází k okamžitému snížení zaměstnanosti
- Povaha *kompenzačního efektu*, který následuje po zavedení technologie a uvolňuje pracovní sílu k jinému druhu práce

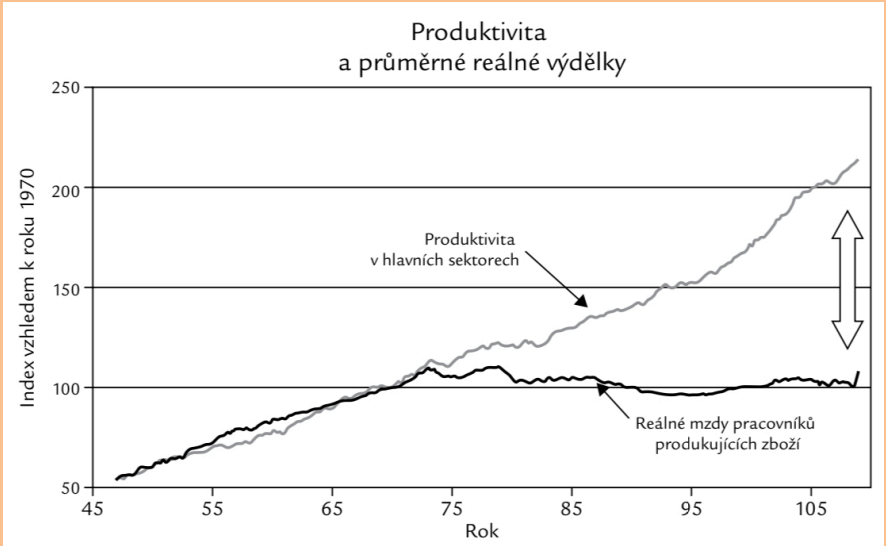
- Optimistický pohled poukazuje na všechny nové druhy pracovních míst, které se objevily po předchozích průmyslových revolucích.
- Pesimistický pohled argumentuje tím, že stroje budou dělat i všechny tyto „nové druhy práce“.

- Pokud stroj nahradí naši fyzickou práci, můžeme prodávat mentální práci, …
- …, pokud stroj nahradí naši mentální práci, co nám pak zbyde na prodej?

- Max Tegmark v knize *Život 3.0* vykresluje tento problém jako rozhovor dvou koní diskutujících o šíření spalovacích motorů v roce 1900. Jeden z nich předpovídá „nová pracovní místa [pro koně] [...]. Vždycky to tak bylo – jako když vynalezli kolo a pluh.“ Pro většinu koní, žel, tato „nová práce“ znamenala stát se žrádlem pro domácí mazlíčky.

Budoucnost práce

- Ekonomové hlavního proudu argumentují z perspektivy „velkého obrazu“: automatizace zvyšuje produktivitu, takže vcelku jsou na tom lidé lépe (dopřáváme si víc zboží i služeb za stejné množství práce)
- Nejde o předpověď, že se v důsledku automatizace bude mít lépe *každý* člověk
- Automatizace obecně zvětšuje podíl příjmů, které jdou do kapitálu (majitelům robotů), a zmenšuje podíl, který jde na práci (bývalým pracovníkům)
- Erik Brynjolfsson a Andrew McAfee v knize *Druhý věk strojů* ukazují, že se to děje už několik desetiletí. Data za Spojené státy jsou vynesena zde, naznačují, že mezi lety 1947 a 1973 rostly společně mzdy i produktivita, ale po roce 1973 mzdy stagnovaly, i když se produktivita zhruba zdvojnásobila (tzv. „velké rozpojení“):



Ekonomická produkce a mediány reálných mezd ve Spojených státech po roce 1947

Budoucnost práce v bankovnictví a obchodu

- Během let 2010–2016 zhruba jedno sto tisíc pracovníků na bankovních přepážkách svou práci ztratilo
- Americký Bureau of Labor Statistics (BLS) předpovídá další výrazné ztráty cca 40 tisíc pracovních míst do roku 2026 v bankovním sektoru: „Očekává se, že online bankovnictví a technologie automatizace budou pokračovat v nahrazování více činností, než se tradičně odhadovalo.“
- Data o pokladních v maloobchodu jsou podobná: jejich počet na hlavu klesl od roku 1997 do roku 2015 o 5 procent, a BLS říká: „Pokrok v technologiích, jako třeba samoobslužné pokladny v obchodech nebo rostoucí podíl online prodeje, bude dál omezovat potřebu pokladních.“
- Oba tyto sektory (bankovnictví a obchod) jsou na sestupné dráze
- Totéž platí o všech zaměstnáních s nízkou kvalifikací, kde se pracuje se stroji

Dynamika úpadku povolání

- Která povolání upadnou s novými technologiemi založenými na AI?
- Příklad uváděný v médiích je řízení vozidel:
- Ve Spojených státech je dnes kolem 3,5 milionu řidičů kamionů a mnohá z těchto pracovních míst jsou ohrožena automatizací.
- Amazon (i jiné společnosti) už dnes používá samořiditelné kamiony pro dopravu zboží na amerických dálkových dálnicích, i když pořád ještě se záložními lidskými řidiči.
- Je velmi pravděpodobné, že dálková část každé cesty kamionů bude brzo probíhat autonomně, i když o dopravu ve městech, nakládku a vykládku se zatím budou starat lidé.
- V důsledku takového očekávaného vývoje se už jen velmi málo mladých lidí zajímá o práci řidiče kamionu jako o kariérní volbu; paradoxně je v současné době ve Spojených státech velmi citelný nedostatek řidičů kamionů, což jen urychluje nástup automatizace.

Administrativa a služby

- Pracovní místa bílých límečků jsou rovněž v ohrožení.
- BLS pro léta 2016–2026 předpovídá 13procentní pokles v zaměstnanosti u pracovníků sjednávajících pojištění:
- „Automatizovaný software pro pojišťování dovoluje pracovníkům pojištění zpracovávat mnohem rychleji než dřív, což zmenšuje potřebu mnoha z pojišťovacích agentů.“

- Pokud se technologie zpracování jazyka (LLM) bude vyvíjet podle předpokladů, mnoho pracovních pozic v prodeji a službách bude rovněž zranitelných, stejně jako pracovní místa v právnických profesích.
- Nicméně: Zhruba 60 % pracovních míst se v roce 2018 nacházelo na pozicích, které v roce 1940 neexistovaly (David Autor a spol.: New Frontiers: The Origins and Content of New Work, 1940-2018, NBER 30389, 2022)
- A také: 85 % růstu zaměstnanosti za posledních 80 let lze vysvětlit vytvářením nových pozic v důsledku technologií (Goldman Sachs, březen 2023)
- Ale: Generativní AI by mohla [v ekonomice USA] vystavit automatizaci 300 tisíc pracovních míst na plný úvazek (tatáž zpráva Goldman Sachs, březen 2023)

Univerzální základní příjem

- Moderní zastánci Keynesovy vize obvykle podporují nějakou podobu univerzálního základního příjmu (universal basic income; UBI)
- Z daní z přidané hodnoty nebo příjmy z kapitálu by UBI poskytl příjmy každému dospělému bez ohledu na okolnosti. Ti, kdo by stáli o vyšší životní standard, by stále mohli pracovat, aniž by UBI ztratili, zatímco ti, kdo by nechtěli, by mohli trávit život, jak se jim zlíbí.
- UBI má relativně širokou podporu napříč politickým spektrem, od Adam Smith Institute po zelené.
- Charakteristika UBI jde od verze ráje po selhání (většina lidí nebude mít žádnou ekonomickou hodnotu, jíž by přispívala společnosti)
- Keynes jasně rozlišoval mezi těmi, kdo o něco usilují, a těmi, kdo si užívají – lidmi „racionálními“ a „rozkošnickými“
- Návrh na UBI předpokládá většinu lidí rozkošnického druhu.
- Keynes předkládá, že usilování o něco je jedním ze „zvyků a instinktů obyčejného člověka, jimiž byl krmen po nespočet generací,“ nikoli „skutečná hodnota života“. Předpovídá, že se tento instinkt postupně ztratí.

- Hédonismus a cílevědomost ale mohou být neoddělitelné: skutečná radost a trvalé naplnění vychází z cíle a jeho dosahování, spíše než díky pasivní konzumaci bezprostředních požitků.
- Existuje rozdíl mezi tím, když na Everest vystoupáte, a když vás na něj vysadí z helikoptéry.

Univerzální základní příjem

- Spojení mezi cílevědomostí a hédonismem je ústředním tématem pro vytváření kýžené budoucnosti
- Budou se budoucí generace divit, proč jsme si pořád dělali starosti s takovými zbytečnostmi jako „práce“?
- Je možné, že většina lidí na tom bude lépe, když bude mít co užitečného dělat, i když většina zboží a služeb bude produkována stroji s minimálním lidským dohledem

- Většina lidí se nutně zapojí do interpersonálních služeb, které mohou být poskytovány – nebo u nichž bychom dávali přednost, aby byly poskytovány – pouze lidmi
- Pokud už nemůžeme poskytovat fyzickou práci a rutinní mentální práci, můžeme pořád dodávat svou lidskost – budeme se muset stát dobrými v tom, jak být lidmi
- Současné profese tohoto typu zahrnují psychoterapeuty, osobní kouče, školitele, poradce, společníky a poskytovatele péče o děti a seniory.
- Pozorování od Keynese:
  - „Budou to tito lidé, kteří dokážou umění života samého udržovat naživu a kultivovat do větší dokonalosti a nezaprodají se sami prostředkům života, kdo si budou schopni užívat přebytku, až přijde.“
- Stroje s AI tak mohou člověka vést k učení se „umění života samého“, schopnosti inspirovat druhé a učit je umění oceňovat a tvořit – ať už ve výtvarném umění, hudbě, literatuře, diskusi, zahradničení, architektuře, jídle, víně nebo videohrách.

Důsledky pro životní naplnění

- Rozdělování příjmů se ve většině zemí již několik desítek let pohybuje chybným směrem: vysoký příjem a vysoké společenské postavení zpravidla plynou z vysoké přidané hodnoty.
- Profese péče o děti se spojuje s nízkými příjmy a velmi nízkým společenským postavením.
- Je to zčásti důsledek toho, že ve skutečnosti nevíme moc dobře, jak tuto práci vykonávat: někteří praktici jsou v tom přirozeně dobří, ale mnoho jich není.
- Naproti tomu třeba ortopedické operace: nenajímali bychom na ně znuděné adolescenty, kteří si potřebují trochu přivydělat, a nenechali bychom je dělat ortopedickou operaci za pět dolarů na hodinu a k tomu cokoli, co najdou k snědku v ledničce.
- Lidé věnovali staletí k pochopení toho, jak funguje lidské tělo a jak je v případě poruchy opravovat, a praktici musí projít léty výcviku, aby se všechny tyto znalosti naučili spolu s dovednostmi potřebnými k jejich užití a ortopedičtí chirurgové jsou dobře placení a vážení.
- Vědecké pochopení mysli je bohužel překvapivě slabé a naše vědecké pochopení štěstí a naplnění je ještě slabší. Nevíme, jak konzistentními a předpověditelnými způsoby přidávat hodnotu životům jiných.
- Dosáhli jsme mírného pokroku v tlumení jistých psychiatrických poruch, ale stále bojujeme stoletou válku o něco tak základního, jako je naučit děti číst.
- Vzdělávací systém i vědecké instituce se budou muset víc soustředit na člověka místo na fyzický svět (Joseph Aoun, prezident americké Northeastern University: univerzity by měly vyučovat a studovat „humanistiku“)
- Bez důkladných a promyšlených změn v tomto směru nejspíš v důsledku užívání AI stojíme před rizikem neudržitelně velkých socioekonomických zlomů.

Bezpečnost AI

- Varování před roboty se objevuje v řadě sci-fi děl, včetně R.U.R. (Rossum's Universal Robots) Karla Čapka (hra napsána v roce 1920, poprvé uvedena v roce 1921 v Praze; následně v New Yorku v roce 1922 a anglické vydání vyšlo v roce 1923). Odtud pochází i slovo „robot“, jímž se ve hře označují uměle vytvořené formy života (biologické stroje, které jsou sestaveny, na rozdíl od vypěstovaných nebo narozených), roboti jmenovaní ve hře jsou Marius, Sulla, Radius, Primus, Helena a Damon; ve hře ovládnou svět.
- Roboti vymykající se kontrole představují archetyp neznámého, podobně jako čarodějnice a duchové v pohádkách z dřívějších dob.
- Robot dost chytrý na to, aby přišel na způsob, jak vyhladit lidstvo, by mohl přijít i na to, že to není zamýšlená užitná funkce; vytváření inteligentních systémů má zahrnovat proces návrhu se zárukami bezpečnosti
- Distribuování nebezpečného aktéra umělé inteligence by bylo zjevně neetické (aktéři se mají vyhýbat nehodám, byli odolní vůči útokům protivníka a zlomyslnému zneužití a obecně mají působit užitek, nikoliv škodu)
- Důležité je to v případě nasazení aktérů umělé inteligence v aplikacích kritických z hlediska bezpečnosti, jako je řízení automobilů, ovládání robotů v nebezpečných továrnách nebo na stavbách a rozhodování o životě a smrti v lékařství
- Bezpečnostní inženýrství má v tradičních technických oborech dlouhou historii (užíváno při konstrukci mostů, letadel, kosmických lodí a elektráren, předem navržených, aby se chovaly bezpečně i v případě, že součásti systému selžou; užívána např. analýza způsobů a důsledků poruch (FMEA): analytici zvažují každou součást systému a představují si všechny možné způsoby, jak by se součást mohla pokazit (například co když praskne tento šroub?), přičemž vycházejí z minulých zkušeností a z výpočtů založených na fyzikálních vlastnostech součásti.

Bezpečnost AI

- Tradiční důraz v softwarovém inženýrství byl na korektnost implementace, nikoli na bezpečnost
- Správnost znamená, že software věrně implementuje specifikaci
- Bezpečnost jde nad to a vyžaduje, aby specifikace zvážila všechny možné způsoby selhání a aby byla navržena tak, aby se i v případě nepředvídaných selhání jen postupně postupně zhoršovala (sw pro samořiditelné auto nebude bezpečný, pokud si neporadí s neobvyklými situacemi: vypadne napájení hlavního počítače – bezpečný systém má záložní se samostatným zdrojem; propíchne se při vysoké rychlosti pneumatika – bezpečný systém má sw opravující výslednou ztrátu kontroly)
- Aktér navržený jako maximalizátor užitku nebo jako cílové funkce může být nebezpečný, pokud má špatnou cílovou funkci (robot s úkolem přinést z kuchyně kávu by neměl jako nezamýšlený vedlejší účinek splnit cíl a cestou převracet lampy a stoly, může ale např. trochu narušit molekuly vzduchu)
- Jedním ze způsobů minimalizace nevhodných efektů je navrhnout robota s nízkým dopadem: místo pouhé maximalizace užitku maximalizujte užitek minus vážený souhrn všech změn stavu světa (při rovnosti všech ostatních věcí robot raději nemění ty věci, jejichž vliv na užitek není znám; vyhne se tedy převržení lampy ne proto, že převržení lampy způsobí pád a rozbití, ale proto, že obecně narušení může být špatné – analogie lékařského kréda „v prvé řadě neškodit“ resp. Analogie regularizace ve strojovém učení: chceme politiku, která dosáhne cílů, ale dáváme přednost politice, která k tomu provádí hladké akce s malým dopadem)
- Měření dopadu je netriviálni: je nepřijatelné převrhnout křehkou lampu, ale je naprosto v pořádku trochu narušit molekuly vzduchu nebo nechtěně zabít některé bakterie v místnosti; není přijatelné poškodit domácí zvířata a lidi – robot musí znát rozdíly kombinací explicitního programování, strojového učení v průběhu času a důkladného testování.

Bezpečnost AI

- Užitkové funkce se mohou mýlit kvůli externalitám (faktory mimo to, co se měří a platí)
- Svět nese důsledky, když jsou skleníkové plyny považovány za externality – společnosti a země nejsou za jejich produkci penalizovány, a doplácejí na to všichni
- Využívání sdílených zdrojů souvisí s *tragédií obecní pastviny,* lze ji zmírnit internalizací externalit – zahrnutím užitkové funkce, např. pomocí uhlíkové daně – nebo využitím principů po staletí užívaných místními lidé po celém světě (Elinom Ostromová, nobelistka z roku 2009):
  - Jasně definujte sdílený zdroj a to, kdo k němu má přístup
  - Přizpůsobte se místním podmínkám
  - Umožněte všem stranám podílet se na rozhodování
  - Monitorujte zdroj pomocí odpovědných kontrolorů
  - Sankce úměrné závažnosti porušení
  - Snadné postupy řešení konfliktů
  - Hierarchická kontrola pro velké sdílené zdroje
- Aktéři AI mohou přicházet na maximalizaci užitku, aniž by ve skutečnosti řešili problém, který tvůrci zamýšleli vyřešit – nejde o podvádění, aktéři jen dělají svou práci
- Aktéři mohou využívat chyb v simulaci (například chyby přetečení v plovoucí desetinné čárce) a navrhnout řešení, která po opravě chyby nefungují
- Aktéři ve videohrách mohou objevit způsoby, jak rozbít nebo pozastavit hru, když směřují k prohře a chtějí se vyhnout. A ve specifikaci, kde se pád hry penalizoval, se jeden agent naučil spotřebovat právě tolik paměti hry, aby mu v okamžiku, kdy byl na řadě soupeř, došla paměť a hra spadla. A konečně, genetický algoritmus pracující v simulovaném světě měl vyvinout rychle se pohybující tvory, ale ve skutečnosti vytvořil tvory, kteří byli enormně vysocí a pohybovali se rychle tak, že padali.

Selhávající maximalizace užitku

- Aktéři AI mohou přicházet na maximalizaci užitku, aniž by ve skutečnosti řešili problém, který tvůrci zamýšleli vyřešit – nejde o podvádění, aktéři jen dělají svou práci – viz např. https://docs.google.com/spreadsheets/d/e/2PACX-1vRPiprOaC3HsCf5Tuum8bRfzYUiKLRqJmbOoC-32JorNdfyTiRRsR7Ea5eWtvsWzuxo8bjOxCG84dAg/pubhtml
- Aktéři mohou využívat chyb v simulaci (například chyby přetečení v plovoucí desetinné čárce) a navrhnout řešení, která po opravě chyby nefungují
- Aktéři ve videohrách mohou objevit způsoby, jak rozbít nebo pozastavit hru, když směřují k prohře a chtějí se vyhnout
- Ve specifikaci, kde se pád hry penalizoval, se jeden aktér naučil spotřebovat právě tolik paměti hry, aby mu v okamžiku, kdy byl na řadě soupeř, došla paměť a hra spadla
- Genetický algoritmus pracující v simulovaném světě měl vyvinout rychle se pohybující tvory, ale ve skutečnosti vytvořil tvory, kteří byli enormně vysocí a pohybovali se rychle tak, že padali, atd.

- Problém *srovnání hodnot*: specifikace užitkové funkce musí maximalizovat přesně to, čeho má být dosaženo
- *Problém krále Midase* jako příklad chybného srovnání hodnot

- Extrémní příklady chybného srovnání hodnot jsou např. chybějící společenské normy: pokud aktér pečuje o čistotu podlahy, lze znečišťovateli domluvit, aby byl čistotnější, ale není přijatelné ho např. unést nebo zneškodnit

- Technika *asistenčních her* jako učení chování odpozorovaného jednání od člověka

Problémy odpozorovaného chování

- Asistenční hry zahrnují opatrné jednání, aby nedošlo k narušení aspektů světa, na kterých by člověku mohlo záležet, a kladení otázek (robot by se např. mohl zeptat, zda je přeměna oceánů na kyselinu sírovou přijatelným řešením globálního oteplování, než tento plán uskuteční)
- Při jednání s lidmi se robot řešící asistenční hru musí přizpůsobit lidským nedokonalostem (pokud robot požádá o povolení, člověk mu ho může dát, aniž by předvídal, že robotův návrh je ve skutečnosti z dlouhodobého hlediska katastrofický; lidé nemají úplný introspektivní přístup ke své skutečné užitkové funkci a ne vždy jednají způsobem s ní slučitelným)
- Lidé někdy lžou, podvádějí nebo dělají věci, o kterých vědí, že jsou špatné. Někdy se dopouštějí sebedestruktivních činů (přejídání nebo zneužívání drog) – systémy AI se nemusí učit přebírat tyto problematické tendence, ale musí pochopit, že existují, když interpretují lidské chování, aby se dostaly k základním lidským preferencím
- Varování významných lidí z technologií (Bill Gates či Elon Musk) nebo vědců (Stephen Hawking či Martin Rees), že by se AI mohla vymknout kontrole (varují, že nemáme žádné zkušenosti s ovládáním silných nelidských entit s nadlidskými schopnostmi, skutečnost je ale horší: máme staleté zkušenosti s národy a korporacemi jako nelidskými entitami, které sdružují sílu tisíců nebo milionů lidí, kde výsledky pokusů o ovládání těchto entit nejsou povzbudivé: národy vyvolávají periodické křeče zvané války, které zabíjejí desítky milionů lidských bytostí, a korporace jsou částečně zodpovědné za globální oteplování a naši neschopnost mu čelit)

Ultrainteligentní stroje

- Systémy AI potenciálně představují mnohem větší problém než národy a korporace, protože mají potenciál se samy rychle zlepšovat; viz I. J. Good již v roce 1965:
  - „Nechť je *ultrainteligentní stroj* definován jako stroj, který dokáže daleko překonat všechny intelektuální aktivity jakéhokoli člověka, jakkoli chytrého. Protože konstrukce strojů je jednou z těchto intelektuálních činností, ultrainteligentní stroj by mohl konstruovat ještě lepší stroje; pak by nepochybně došlo k „*explozi inteligence*“ a inteligence člověka by zůstala daleko za ním. První ultrainteligentní stroj je tedy *posledním* vynálezem, který člověk kdy potřebuje, za předpokladu, že stroj bude natolik poslušný, že nám řekne, jak ho udržet pod kontrolou.“
- Goodovu "explozi inteligence" nazval technologickou singularitou profesor matematiky a autor sci-fi Vernor Vinge, když v roce 1993 napsal: „Do třiceti let budeme mít technologické prostředky k vytvoření nadlidské inteligence. Krátce poté skončí lidská éra.“
- V roce 2017 vynálezce a futurista Ray Kurzweil předpověděl, že singularita se objeví do roku 2045, což znamená, že se k ní přiblížila o dva roky za 24 let (tímto tempem zbývá už méně než 340 let)
- Vinge i Kurzweil poznamenávají, že technologický pokrok v mnoha ohledech v současnosti roste exponenciálně
- Extrapolace cesty od rychle se snižujících nákladů na výpočet až k singularitě může být přílišný skok – dosud každá technologie sledovala S-křivku (logistickou křivku), kde se exponenciální růst nakonec zužuje
- Nové technologie někdy nastupují, když staré dosáhnou svého vrcholu, ale někdy není možné růst udržet, ať už z technických, politických nebo sociologických důvodů: např. technologie létání se od letu bratří Wrightů v roce 1903 do přistání na Měsíci v roce 1969 dramaticky posunula, ale od té doby průlom srovnatelného rozsahu nebyl.
- Další překážkou v cestě ultrainteligentním strojům k ovládnutí světa, je fakt, že některé druhy pokroku vyžadují nejen myšlení, ale i jednání ve fyzickém světě.

Problém gorily

- Problém gorily (S. Russell: Jako člověk): „*Před zhruba deseti miliony let vytvořili předchůdci moderních goril (náhodou, to je jisté) genetickou linii vedoucí k moderním lidem. Jaký pocit z toho mají gorily? Pokud by byly schopny nám něco říct o situaci svého druhu vzhledem k lidem, byl by konsensuální názor určitě velmi negativní. Jejich živočišný druh nemá v podstatě žádnou jinou budoucnost, než jakou se nám uráčí mu povolit. Nechceme být v podobné situaci vůči superinteligentním strojům. Budeme to označovat jako problém gorily – konkrétně to, zda si lidé dokážou zachovat svou nadřazenost a autonomii ve světě, který obsahuje stroje s podstatně větší inteligencí.*“
- Alan Turing (přednáška v Manchesteru v 1951): „*Zdá se být pravděpodobné, že jakmile nastoupí strojový způsob myšlení, nebude trvat dlouho, než předčí naše chabé schopnosti. Nebude přicházet v úvahu, že by stroje umíraly, a budou spolu moci konverzovat, aby si bystřily důvtip. V nějaké fázi bychom tedy měli očekávat, že stroje převezmou nadvládu způsobem, který je zmiňován v Erewhonu Samuela Butlera.*“
- Alan Turing (BBC, 1951): „*Pokud bude stroj myslet, mohl by přemýšlet mnohem inteligentněji než my, a kde potom budeme? I kdybychom mohli držet stroje v podřízeném postavení třeba tak, že bychom ve strategických chvílích vypínali elektřinu, měli bychom se jako živočišný druh cítit nanejvýš zahanbeni. [...] Toto nové nebezpečí [...] je určitě něco, z čeho bychom měli pociťovat úzkost.*“

Rizika reálných systémů

- Rizika AI jsou podstatná v tom smyslu, že systémy zaměřené na maximalizaci nějaké cílové funkce se mohou a budou odchlovat od lidských preferencí (proměnných v čase a neznámých přesně)
- Superinteligentní systémy AI by měly nutně zůstat ovládány lidmi, není však zřejmé, jak toho dosáhnout
- Představa nějakého zákazu nasazení superinteligentních systémů AI asi není realizovatelná
- Pokud skutečný problém existuje, měl by být řešen, i kdyby zatím nebyl naléhavý.
- Není ani realistické snažit se výzkum v AI zakázat, je ale třeba na předběžných aspektech tohoto problému začít pracovat (v médiích diskutovaná výzva na 6měsíční moratorium v souvislosti s GPT3.5–>GPT4–>GPT?? sleduje cíl vytvořit pro to prostor)

- Aktuálně v souvislosti s GPT: Překvapení ze síly jednoduchého mechanismu v kombinaci s velikostí (otázka vyvstávajících/emergentních vlastností nepředpokládaných tvůrci, otázka skrytých vlastností jazykových stuktur) a výrazné zkrácení odhadů pro dosažení stupně obecné AI (např. podle Geoffrey Hinton uvádí ve Wired v květnu 2023 zkrácení odhadu z 30-50 let na 5-20 let)
- Doporučení ke shlédnutí: Podcasty Lexe Fridmana (https://youtube.com/@lexfridman) z poslední doby, zejména 371, 368, 367 a 373

Argumenty popírání nebezpečí reálných systémů AI

- Elektronické kalkulačky mají nadlidské schopnosti ve vztahu k aritmetice. Kalkulačky si nepodrobily svět; není proto důvod si dělat starosti s nadlidskou AI.
- ❖ Inteligence není totéž co aritmetika a aritmetické schopnosti kalkulaček jim nedávají výbavu pro ovládnutí světa.
- Koně mají nadlidskou sílu, ale nepanikaříme, abychom dokázali, že jsou bezpeční. Proto se nemusíme starat ani o dokazování, že jsou systémy AI bezpečné.
- ❖ Inteligence není totéž co fyzická síla; a síla koní jim nedává výbavu ovládnout svět.
- V dějinách je nula příkladů strojů, které zabily miliony lidí, takže to indukcí nemůže nastat ani v budoucnu.
- ❖ Všechno se někdy stane poprvé a před tím byla nula příkladů, kdy se tak stalo.
- Žádná fyzikální veličina ve vesmíru nemůže být nekonečná a to zahrnuje i inteligenci, takže starosti o superinteligenci jsou přehnané.
- ❖ Superinteligence nemusí být nekonečná, aby byla problematická; a fyzikálně jsou možná výpočetní zařízení miliardkrát výkonnější než lidské mozky.
- Neděláme si starosti s kataklyzmatickými, ale vysoce nepravděpodobnými riziky, jako že se poblíž Země objeví černá díra. Proč si dělat starosti se superinteligentní AI?
- ❖ Pokud by většina fyziků na Zemi na takové černé díře pracovala, neptali bychom se jich, zda je to bezpečné?

Zdroj: S. Russell: Jako člověk

Principy pro prospěšné stroje

- Představa obecných zásad realizujících stroje, které budou srovnány s cíli člověka

1. Jediným záměrem stroje je maximalizace uskutečňování lidských preferencí.
2. Stroj si na začátku není jistý tím, jaké tyto preference jsou.
3. Zásadním zdrojem informací o lidských preferencích je lidské chování.

Zdroj: S. Russell: jako člověk

PRVNÍ PRINCIP: ČISTĚ ALTRUISTICKÉ STROJE

1.Jediným záměrem stroje je maximalizace uskutečňování lidských preferencí.

- Ústřední úloha principu, že jediným záměrem stroje je maximalizace uskutečňování lidských preferencí
- Bude prospěšný zejména lidem, nikoli třeba zvířatům
- Stroj musí být čistě altruistický –nepřipisuje absolutně žádnou vnitřní hodnotu své vlastní pohodě, nebo dokonce své vlastní existenci
- Stroj může sám sebe chránit, aby mohl dál dělat věci užitečné pro lidi nebo protože by opak vadil vlastníkovi, ale nikoli proto, že by chtěl zůstat naživu.
- Jakékoli preference jeho vlastního zachování by znamenaly dodatečnou motivaci uvnitř robota, která se striktně nesrovnává s tím, jak se daří lidem

DRUHÝ PRINCIP: POKORNÉ STROJE

2. Stroj si na začátku není jistý tím, jaké tyto preference jsou.

- Princip, že si stroj na začátku není jistý, jaké vlastně lidské preference jsou, je klíčem k vytváření prospěšných strojů
- Stroj, který by předpokládal, že dokonale zná skutečné záměry, je bude naplňovat s klapkami na očích a nikdy se nezeptá, zda je nějaký postup v pořádku
- Stroj, který si není jistý skutečným záměrem, bude vykazovat pokoru: podřídí se například lidem a dovolí, aby byl vypnut (součástí jeho uvažování bude, že ho člověk vypne, kdyby něco dělal špatně – tedy pokud by dělal něco v rozporu s lidskými preferencemi)
- Pokud člověk stroj vypne, vyhne se stroj tomu, aby dělal něco špatně, a to je, oč mu jde: zůstává spojen s člověkem, který je potenciálním zdrojem informací, jež mu umožní vyhnout se chybám a dělat svou práci lépe
- Nejistota je ústředním pojmem AI od 80. let dvacátého století, ale nejistota v záměrech systémů AI pomíjena a při práci na maximalizaci užitku, dosahování cílů, minimalizaci nákladů a minimalizaci ztrát se předpokládalo, že jsou tyto parametry dokonale známy

TŘETÍ PRINCIP: UČIT SE PŘEDVÍDAT LIDSKÉ PREFERENCE

3. Zásadním zdrojem informací o lidských preferencích je lidské chování.

- Prvním důvodem je poskytnout konečné ukotvení pro termín lidské preference: lidské preference nejsou ve stroji a ten je nemůže přímo pozorovat, musí ale pořád existovat nějaké propojení mezi strojem a lidskými preferencemi.
- Propojení s lidskými preferencemi nastává prostřednictvím pozorování voleb, jež lidé dělají
- Druhým důvodem je umožnit stroji stávat se užitečnějším, když se dozví víc o tom, co chceme (kdyby o lidských preferencích nevěděl nic, nebyl by člověku k užitku)
- Lidské volby odhalují informace o lidských preferencích.
- Lidé nejsou dokonale racionální: mezi lidskými preferencemi a lidskými volbami vznikají nekoherence a stroj tyto nekoherence musí brát v úvahu, pokud má lidské volby interpretovat jako indicie lidských preferencí.

Prospěšné stroje

- Standardní model užívaný ve velké části techniky dvacátého století, je založen na mechanismech optimalizujících pevné, zvnějšku dodané záměry
- Funguje pouze tehdy, když je u takového záměru zaručeno, že je úplný a správný, nebo pokud lze zařízení snadno vypnout – ani jedna z těchto podmínek nebude splněna, jakmile se AI stane dostatečně mocnou
- Může-li být zvenčí dodaný záměr špatný, pak nemá smysl, aby stroj pracoval, jako kdyby byl správný
- U prospěšných strojů, u nichž lze čekat, že jejich činnost bude naplňovat naše záměry, se tyto stroje muset učit víc o tom, co skutečně chceme, a to pozorováním, jaká rozhodnutí děláme a jak je děláme
- Takto navržené stroje se podřídí člověku: budou žádat o povolení; budou jednat opatrně, pokud nebudou pokyny jasné; a dovolí, aby byly vypnuty
- Oprávněnost „dokazatelně přínosného“ přístupu k AI by měla být podložena jak matematickými rozbory, tak praktickými realizacemi v podobě užitečných aplikací (zatím není)

Ochablost a lidská autonomie
Rizika nekontrolovaně sloužících strojů

- Na Zemi dosud žilo víc než sto miliard lidí, kteří strávili řádově jeden bilion člověkoroků učením sebe sama a učením druhých, aby mohla naše civilizace pokračovat.
- Až donedávna k tomu měli jediný prostředek: znovuvytváření idejí v myslích nových generací. (Papír jako metoda předávání funguje, ale sám o sobě neudělá nic, dokud na něm zaznamenané znalosti nezasáhnou mysl další osoby.)
- To se teď mění: naše znalosti je stále víc možné vkládat do strojů, které mohou naši civilizaci udržovat v chodu samy, bez nás.Jakmile se ztratí praktické pohnutky předávat naši civilizaci dalším generacím, bude tento proces zvrácen.
- Jeden bilion let kumulativního učení se v reálném smyslu ztratí. Staneme se pasažéry na výletní lodi provozované stroji, ocitneme se na výletě, který pokračuje napořád.
- Stroje mohou dobře chápat, že lidská autonomie a kompetence jsou důležitými aspekty toho, jak chceme vést své životy. Mohou klidně trvat na tom, že si lidé ponechávají ovládání a odpovědnost za svůj vlastní blahobyt – jinými slovy, stroje odmítnou převzít řízení. (Ale my, krátkozrací líní lidé, s tím můžeme nesouhlasit.)
- Jde o modifikovanou tragédii obecní pastviny: libovolný jednotlivý člověk může považovat za zbytečné věnovat roky pracnému získávání znalostí a dovedností, které stroje již mají; ale pokud takto bude uvažovat každý, lidstvo kolektivně přijde o svou autonomii.
- Řešení tohoto problému je zřejmě kulturní, nikoli technické: potřebujeme kulturní hnutí, jež by znovu formovalo naše ideály a preference směrem k autonomii, jednání a schopnostem a pryč od nestřídmosti a závislosti, korigování lidských preferencí v globálním měřítku, spolu s radikálními změnami fungování společnosti - superinteligentní stroje mohou pomoci při utváření takového řešení i během procesu dosahování rovnováhy pro každého jedince.
- Proces podobný vztahu rodiče a malého dítěte: dítě odrostlé bezmocni znamená pro rodiče udržování stále se vyvíjející rovnováhu mezi možnostmi dělat pro dítě všechno, anebo ho ponechat zcela jeho vlastním schopnostem: dítě dospěje k pochopení, že rodič sice dovede zavázat mu tkaničky, ale že to od něj už nechce.
- Je pro člověka žádoucí budoucnost, aby s ním mnohem dokonalejší stroje navždy zacházely jako s dítětem? Jedním z rozdílů je, že děti nemohou své rodiče vypnout.
- V našem současném světě neexistuje nic podobného vztahu, jejž v budoucnu budeme mít s prospěšnými inteligentními stroji.

Stuart Russell: Jako člověk, Argo/Dolořán, 2021