Úvod do umělé inteligence
(na začátku roku 2023)

Šimon Podhajský

asociace
debatních
klubů

Kdo jsem

- BS v kognitivních vědách z Yale University
  (obor zahrnoval i předměty v informatice a AI)
- AI hobbyista a laik
  (tj. **nejsem citovatelný** jako autorita!)
- WSDC 2008-11, vítěz DL 2009 + 2011

Co je umělá inteligence?

Produkt snahy o vytvoření stroje, který se bude inteligentně chovat (tj. schopnost vnímat, dedukovat a vytvářet informace a podle nich se řídit)

Typicky výsledek strojového učení (*machine learning, či ML*)

Příklady:

- Autonomní systémy (samořídící auta, bezpečnostní/vojenští roboti)
- Tvůrčí (generativní) systémy (DALL-E, ChatGPT)
- Hráčí-boti (AlphaGo, …)
- *Podobnými metodami i pokročilé rozpoznávání řeči, vyhledávání apod.*

Jak funguje strojové učení? [*] - příklad kalkulačky

1. Vyber algoritmus (např. neuronová síť, hloubkové učení, …)
2. Nakrm algoritmus daty, ve kterých je každá vstupní hodnota spojena se správnou výstupní hodnotou (např. “2 + 2” je vstup a “4” je správný výstup)
3. Nech algoritmus, aby se naučil “pravidla”, která předpovídají správný výstup na základě vstupu
4. Pusť algoritmus “do světa” na nová data

[*] supervidované strojové učení

Implikace procesu strojového učení

1. Pokud jsou tréninková data **chybná** (např. 2 + 2 = 5), stroj se naučí vztahy chybně
2. Pokud jsou tréninková data **zaujatá** (např. data o kariérních výsledcích žen a mužů), stroj se naučí používat **předsudky**
3. Pokud jsou tréninková data **nekompletní** (např. data pro samořídící auta nemusí obsahovat kontakt s koňským spřežením), nevíme, jak se stroj zachová
4. Pokud se stroj může dobrat výsledku **nedetekovaným podvrhem**, naučí se podvádět

Tyto problémy jsou obecné pro drtivou většinu moderní AI!

Další problém s rozšířením AI: distribuční škála

AI systémy, které jsou široce dostupné, může využívat každý.

To znamená, že:

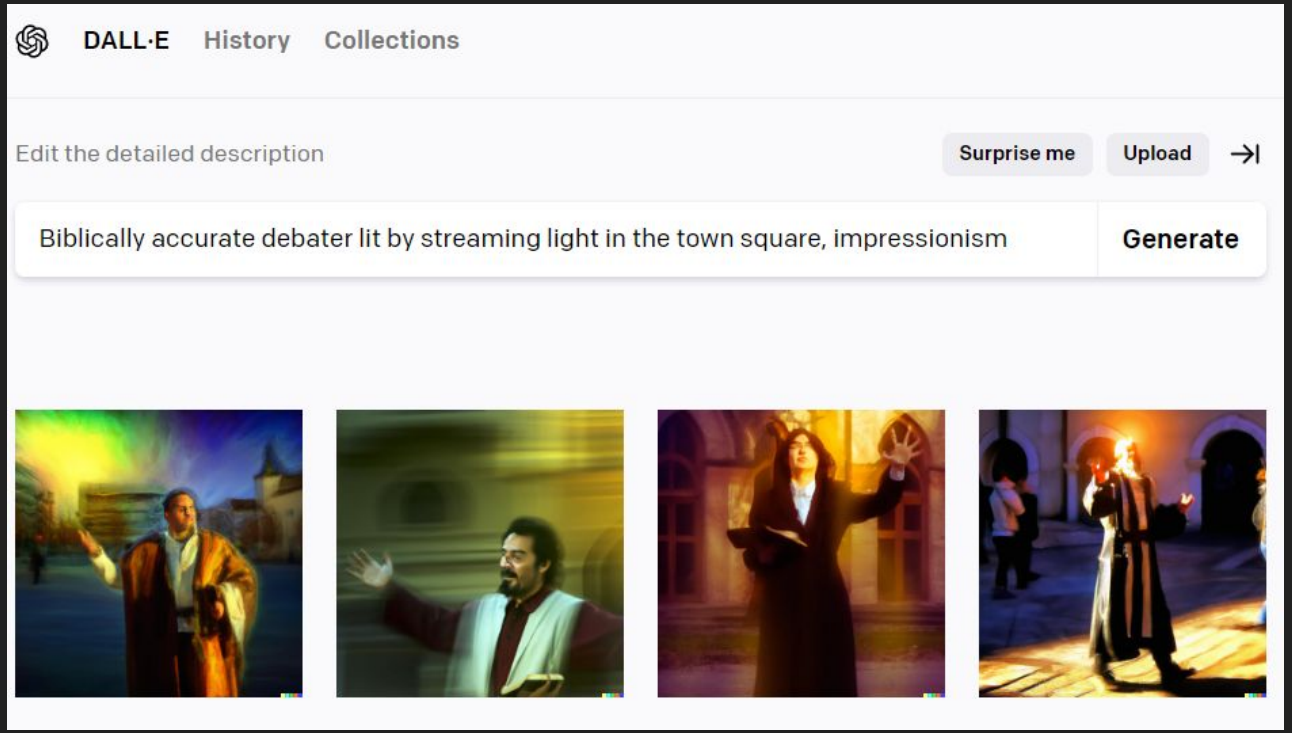
- zneužitelnost roste
- dopady jakýchkoliv problémů v modelu/tréninku jsou znásobeny

Současné problémy AI a disciplína “AI Ethics”

AI umění (DALL-E, Midjourney, …)

Generátor obrázků a maleb na základě popisného textu.

(Všechny ilustrace v této prezentaci byly vygenerovány umělou inteligencí.)



Problémy s AI uměním

- Porušování duševního vlastnictví při tréninku modelu
- Potenciální porušení duševního vlastnictví výsledky modelu
- Podrývání živobytí umělců
- Naučená objektifikace žen a sexismus/rasismus

Deepfakes

- Ars Technica:
  - “AI image generation tech can now create life-wrecking deepfakes with ease” (pomocí Stable Diffusion + Dreambooth)
  - “Microsoft’s new AI can simulate anyone’s voice with 3 seconds of audio” (VALL-E)
- Vytváření násobně přesvědčivějších kompromitujících materiálů a falešných zpráv
- Potenciál destabilizovat politickou situaci a působit společenský rozkol

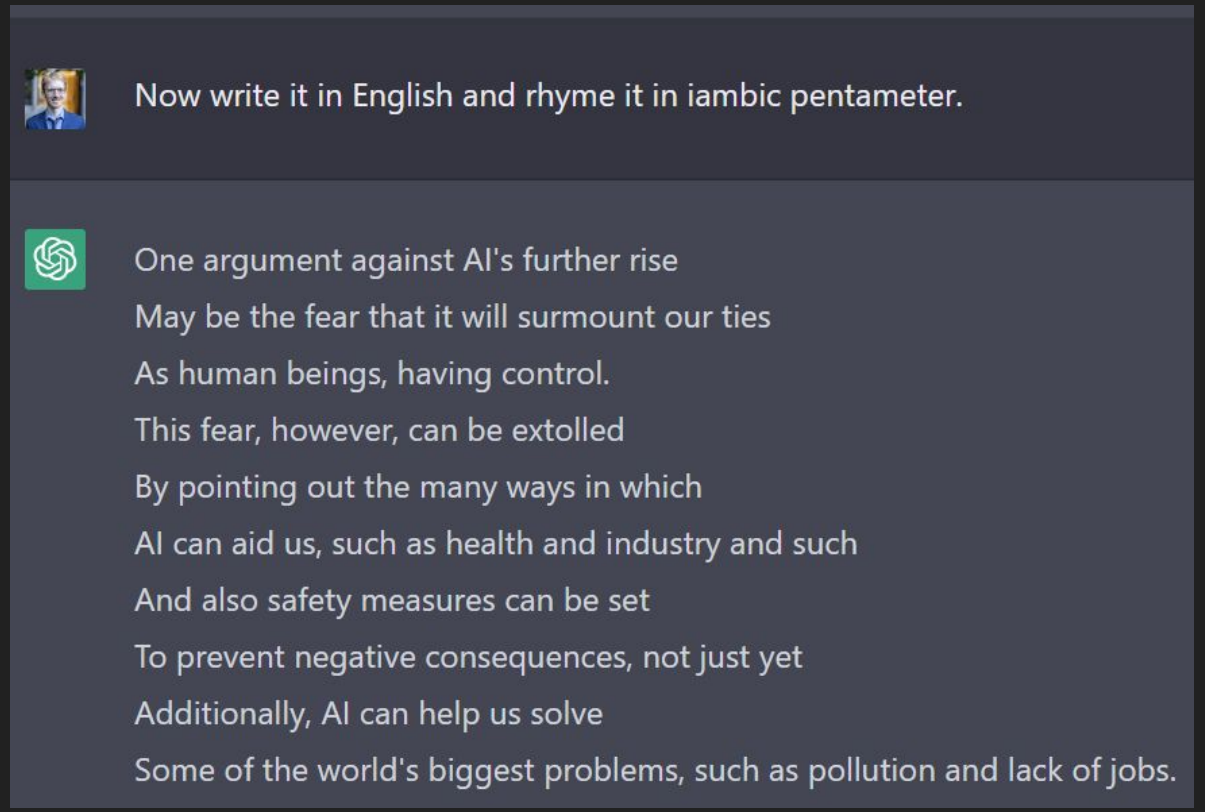
Chatboti a doplňovači textu (ChatGPT, Replika, Copilot, ...)

Psané reakce na základě tréninku large language model (LLM).

Napiš první argument pro tezi "Další vývoj umělé inteligence je nežádoucí". Zrýmuj ho.

Jedním z argumentů pro tezi, že další vývoj umělé inteligence je nežádoucí, může být obava z toho, že umělá inteligence překoná lidskou inteligenci a bude mít nad námi kontrolu. Tuto obavu lze refutovat tím, že pokrok v oblasti umělé inteligence může pomoci v mnoha oblastech, jako je zdravotnictví, energie, doprava a průmysl, a že existují bezpečnostní opatření a regulace, které mohou pomoci předcházet negativním důsledkům. Navíc, pokrok v oblasti AI může pomoci při řešení některých z největších výzev, kterým čelíme, jako jsou klimatické změny a nedostatek pracovních sil.

(čím větší jazyk, tím lepší schopnosti)



Problémy s chatboty

- Potenciální podrývání živobytí všech profesí, které pracují s informacemi
- GPT někdy vyprodukuje rasistický, sexistický či jinak problematický text
- (...a dělá to navzdory tomu, že přesně proti takovému konání byla explicitně trénována)

Mnoho dalších AI systémů, každý se svými problémy

- Medicínské AI - diagnostické, servisní, vynálezecká (AlphaFold), …
- Policejní a soudní AI - preventivní, vyhodnocující propuštění/recidivitu (COMPAS), …
- Sledovací AI - rozpoznávání obličejů veřejnými kamerami, monitoring populace totalitními režimy, …

Disciplína “AI Ethics” je vděčný zdroj dalších příkladů

- Timnit Gebru
- Oxford: Ethics in AI Institute
- Algorithmic Justice League
- Stanford: Human-Centered Artificial Intelligence Institute

Potenciální problémy AI a disciplína “AI Safety”

Jak realistický je scénář “Skynet”?

Scénář Skynet: “Mocný (možná superinteligentní) počítač, který - buď ze zlého úmyslu nebo z neopatrnosti - ohrozí či vyhubí lidstvo.”

Rozdělíme na dvě otázky:

1. Jak blízko jsme vytvoření mocného počítače, který by měl schopnost nás vyhubit?
2. Jak by se mohlo stát, že by nás takový počítač vyhubit *chtěl* nebo naše vyhubení *dopustil?*

Midjourney: “T-800 delivering an impassioned speech at a lectern, gesticulating with its robot hand”

Pravděpodobnost superinteligence a AI zimy

- Superinteligentní počítač / “široká” AI (AGI) je zlatým grálem
- Bude umět všechno lépe - včetně tvorby dalších AI - a bude schopna se sama zlepšovat a škálovat
- Ani není třeba superinteligence - dostatečně pokročilá “úzká” AI může být dostatečně napojená na svět a schopna dekontextualizovaného uvažování - např. ChatGPT umí matematiku (trošku), přestože jej nikdo explicitně neučil počítat
- Řada vědců v oboru si myslí, že takový počítač pravděpodobně vyvineme v příští dekádě

Na druhou stranu…

- “AI zimy” jsou dlouhé prodlevy mezi kratšími obdobími “AI boomů”, během kterých se zdá, že AGI je nadosah
- Dartmouth Workshop (1956): “Myslíme si, že můžeme tento obor (AI) významně posunout, pokud na něm strávíme v 10 lidech dva měsíce.”

Bude nás chtít zabít? Problém zarovnání (alignment problem)

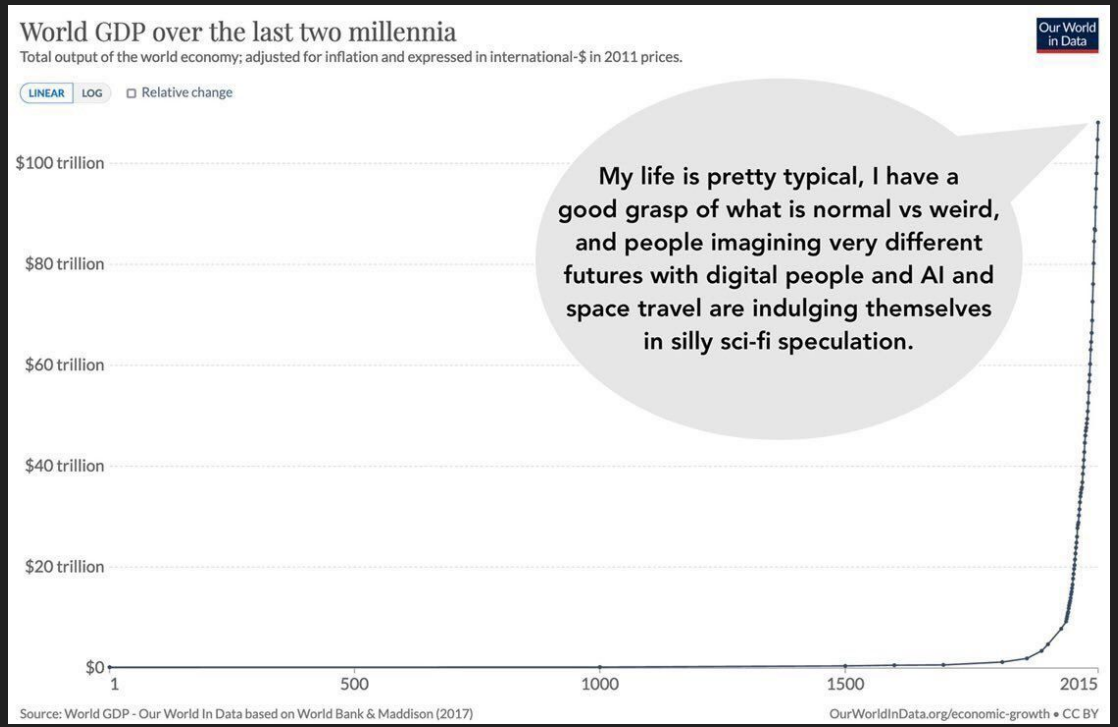
- Úzká vs. široká/všeobecná AI (AGI) - zabít nás může i úzká, bude-li mít dostatek zdrojů
- Není důvod si myslet, že dostatečně pokročilá AI si vyvine stejné hodnoty, jaké má lidstvo:
  - Naše “kalkulačka” neměla žádné “hodnoty” kromě správného výsledků
  - Specifikace “lidských” hodnot je extrémně složitá - sdílíme je vůbec? Jak vyjadřujeme vzájemně sporné principy?
- Mohla by nás tedy zabít “omylem”, nebo protože jsme v konfliktu s jejími primárními cíli
- Špatně specifikovaná lidská hodnota (např. “zvýšit lidské štěstí”) může mít v rukou superinteligence nečekanou implementaci (např. násilně daný nitrožilní heroin pro všechny)
- Dobře specifikovaná lidská hodnota může *stále* mít nečekanou implementaci (viz příklady špatně specifikované AI)

Některé příklady nečekaných chování AI

("Specification gaming examples in AI - master list", Viktoria Krakovna, OpenPhil)

Title	Description
Aircraft landing	Evolved algorithm for landing aircraft exploited overflow errors in the physics simulator by creating large forces that were estimated to be zero, resulting in a perfect score
Bicycle	Reward-shaping a bicycle agent for not falling over & making progress towards a goal point (but not punishing for moving away) leads it to learn to circle around the goal in a physically stable loop.
Block moving	A robotic arm trained using hindsight experience replay to slide a block to a target position on a table achieves the goal by moving the table itself.
Indolent Cannibals	In an artificial life simulation where survival required energy but giving birth had no energy cost, one species evolved a sedentary lifestyle that consisted mostly of mating in order to produce new children which could be eaten (or used as mates to produce more edible children).

Je to sci-fi? Možná, ale ve sci-fi žijeme



Disciplína “AI Safety” je vděčný zdroj dalších příkladů

- Future of Humanity Institute (FHI)
- Machine Intelligence Research Institute (MIRI) & Eliezer Yudkowsky
- OpenAI, Anthropic, DeepMind mají AI Safety divize/týmy, které produkují výzkum (např. Amodei et al.: “Concrete Problems in AI Safety”)

Další užitečné zdroje

- Shrnující článek o riziku vyhynutí kvůli AI od Kelsey Piper pro Vox
- Metodická poznámka ŘS od Filipa Vrťa
- Amodei et al. (Google Brain): Concrete Problems in AI Safety
- Victoria Krakovna: Seznam příkladů, ve kterých AI splnila cíl nečekaným či kontraproduktivním způsobem
- Bostrom: Superintelligence (kniha; TED Talk o knize, přednáška pro Talks at Google)
- Ajeya Cotra: “Without specific countermeasures, the easiest path to transformative AI likely leads to AI takeover”
- Holden Karnofsky: “AI Could Defeat All Of Us Combined”

Nick Bostrom: What happens when our computers get smarter than we are?