MUNI
FI

Etika umělé inteligence

Mgr. Tomáš Foltýnek, Ph.D.
foltynek@fi.muni.cz

Osnova dnešní přednášky

- Prezentace Michala Kolaříka
- Etika umělé inteligence
  - Velké jazykové modely: Škodlivý a zkreslený obsah
  - Zodpovědnost za rozhodnutí AI
  - Důsledky pro lidskou společnost
  - Důsledky pro životní prostředí
  - Regulace: AU AI Act
- Dilemma game
  - Nestát příliš blízko

Etika umělé inteligence

MidJourney /imagine artificial intelligence

Etika umělé inteligence

- Technologické hledisko:
  Posouvat hranice toho, co systémy **mohou** dělat

- Etické hledisko:
  Zabývat se i tím, zda by systém **měl** něco umět či dělat

- Kritérium: Prospěšnost pro lidskou společnost

Ethics {by, in, for} Design

- **Ethics by design**: Součástí rozhodovacích algoritmů má být schopnost etického zhodnocení zamýšlených akcí
- **Ethics in design**: Metody podporující analýzu a zhodnocení etických důsledků navrhovaných systémů
- **Ethics for design**: Etické kodexy, standardy, certifikační procesy zajišťující integritu vývojářů a uživatelů ve všech fázích životního cyklu systému

Etické otázky velkých jazykových modelů

- Timnit Gebru, bývalá ředitelka Google AI Ethics
- Článek “Ethical conisedarions of large text models” nebyl nikdy publikován, Gebru byla donucena opustit Google

- Učení a provoz – spotřeba elektřiny / uhlíková stopa
  - Učení GPT-3: 1287 MWh (Patterson et al., 2022)
    - Roční spotřeba 217 lidí v ČR
- Trénování jazykových modelů především v angličtině
  - Benefituje již bohatá část planety
- Důsledky změny klimatu trpí chudá část planety
  - Maledivy budou pod vodou, v Súdánu jsou častější záplavy, atd.
  - přitom na jejich jazycích se nic netrénuje
- Environmentální rasismus

Source: https://www.technologyreview.com/2020/12/04/1013294/google-ai-ethics-research-paper-forced-out-timnit-gebru/

Etické otázky velkých jazykových modelů

- Trénování ze zkreslených dat na internetu
  - Příliš velké datasety je nemožné prověřit
  - Obsah – rasismus, sexismus, násilí, zneužívání moci
    - AI považuje za normální
    - „Dáme-li AI veškerou krásu, ošklivost a krutost, pak nemůžeme čekat, že na výstupu bude jen krása“
  - Další vylučování již vyloučených skupin
- Diverzita trénovacích dat
  - Reddit: 67 % uživatelů jsou muži, 64 % uživatelů je ve věku 18 – 29 let
  - Wikipedia: Jen 9 – 15 % wikipedistů jsou ženy
  - Blogy (psané spíš staršími) nejsou v trénovacích datech zastoupeny tak jako sociální média (užívané spíše mladšími)

Zdroj: Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021, March). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency* (pp. 610-623). https://dl.acm.org/doi/pdf/10.1145/3442188.3445922

Microsoft Tay Chatbot

- Spuštěn v březnu 2016
- Komunikoval s lidmi na sociálních médiích
  - Twitter, Facebook, Instagram a Snapchat
- Záměr: Zábavné, neformální, hravé konverzace
  - Naučen na veřejných konverzacích na sociálních sítích
- Realita: Rasistický, fašistický a sexistický trol
  - Naučen na veřejných konverzacích na sociálních sítích
- Vypnut po 24 hodinách
- Ostuda pro Microsoft, ale cenná lekce pro vývoj AI systémů

Zdroj: https://www.theverge.com/2016/3/24/11297050/tay-microsoft-chatbot-racist

Čištění GPT

- Před spuštěním ChatGPT bylo potřeba jej zbavit závadného obsahu
  - Sexuální zneužívání, násilí, nenávistný obsah
- Supervised learning → Potřeba otagovaných textů
  - *“Our mission is to ensure artificial general intelligence benefits all of humanity, and we work hard to build safe and useful AI systems that limit bias and harmful content,”*
- OpenAI najala firmu Sama, ta najala dělníky z Keni a Ugandy
  - Pracovala i na filtru pro Facebook
  - Plat mezi 1,32 – 2,00 USD na hodinu (OpenAI platilo 12,50 USD)
  - Měli přečíst a otagovat 150 – 250 textů (každý 100 – 1000 slov) za 9h směnu
    - Detailní popisy mučení, poprav, sebevražd, incestu, znásilnění, sexuálního zneužívání dětí…
- Psychická traumata dělníků → Ukončení kontraktu

Zdroj: https://time.com/6247678/openai-chatgpt-kenya-workers/

Galactica

- Spuštěna 15. listopadu 2022
- Meta AI (Facebook)
- Generativní jazykový model na pomoc vědcům
  - Naučen na 48 milionech vědeckcýh článků, učebnic, přednášek…
- Problémy: Nepravdivé nebo zavádějící, ale přesvědčivé výstupy
  - Rizika: Narušení vědecké pravdy
  - Navíc k paper mills, predátorským časopisům,…
- Nejasnné přínosy pro poctivé vědce
- Vypnuta po třech dnech

Zdroj: https://www.technologyreview.com/2022/11/18/1063487/meta-large-language-model-ai-only-survived-three-days-gpt-3-science/

Bias: Anecdotical Evidence

Midjourney was asked to draw a professor, a doctor and a manager

Kde leží hranice mezi užitečnými znalostmi o světě a škodlivými stereotypy?

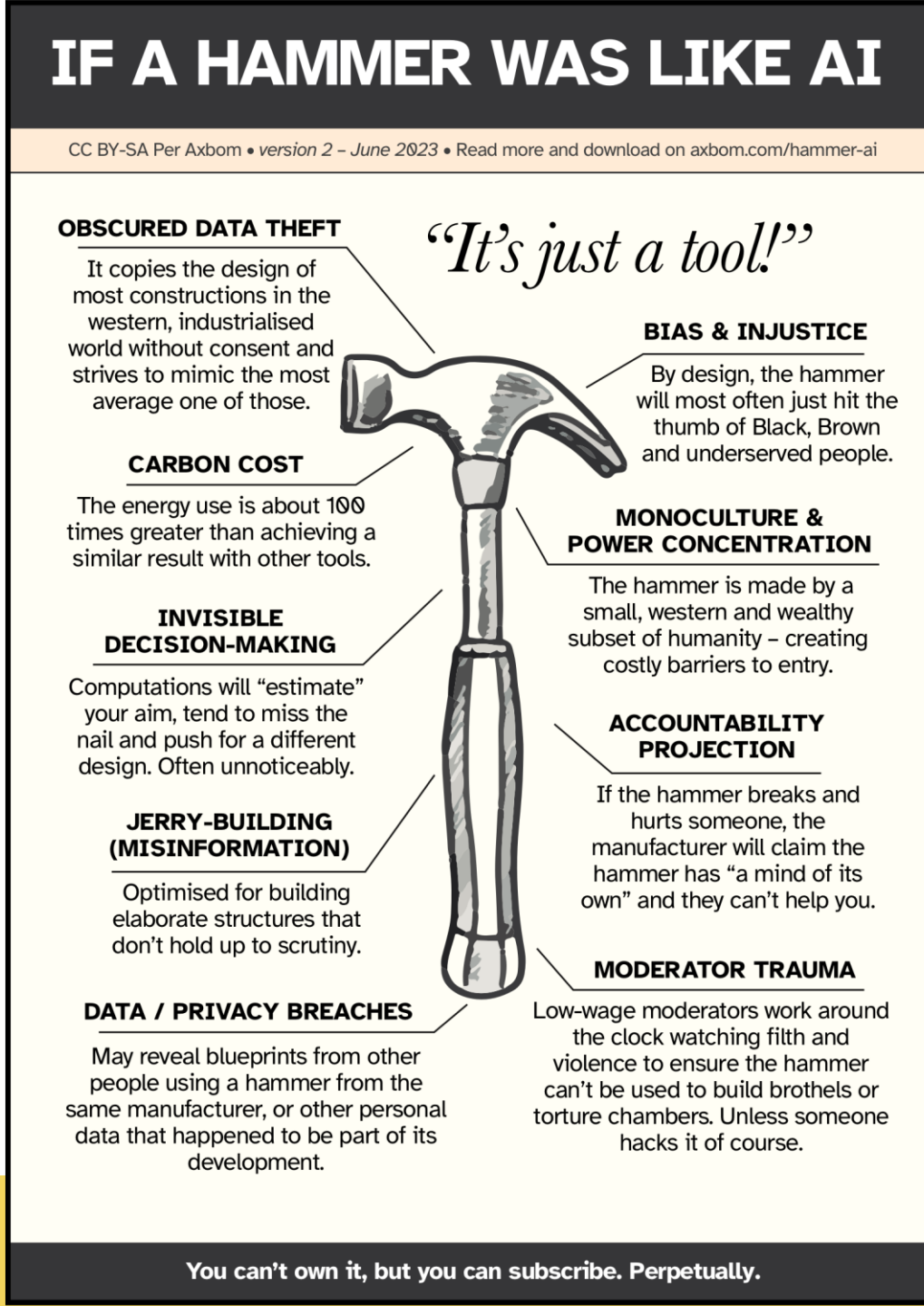
Jak (objektivně) měřit zkreslení?

- Netriviální problém, záleží na aplikaci
- Speciálně navržené datasety obsahující
  - Začátky textů k doplnění
  - Otázky k zodpovězení
  - Nejednoznačný text k přeložení
  - Text s vynechanými místy k doplnění
- Vymezení skupin, které nás zajímají
  - Pohlaví, věk, rasa, náboženství, povolání, politické přesvědčení
- Metriky ve vztahu ke skupinám
  - Přesnost překladu
  - Správnost odpovědi
  - Sentiment v odpovědi

Důsledky využívání AI

Důsledky pro lidskou společnost

- Změny na pracovním trhu
  - Ztráta pracovních míst
  - Vytvoření nových pracovních míst
- Závislost na technologiích
  - Degradace lidských schopností
- Prohlubování stávající nerovnosti
  - Digital divide



Zdroj: https://axbom.com/hammer-ai/

Důsledky pro životní prostředí

- Spotřeba energie → Uhlíková stopa
  - Při trénování
  - Při využívání

- Suroviny na výrobu HW
  - Těžba
  - Odpad

Midjourney /imagine a landfill full of old computer parts

Zodpovědnost

- Kdo má nést zodpovědnost za důsledky (škody)?
- Vývojář? Dodavatel? Provozovatel? Uživatel? AI?
- Závisí na kontextu
  - U GenAI/LLM jednoznačně uživatel
  - U autonomních systémů (automobily, zbraně,…) ???

- Moral outsourcing
  - Přenesení odpovědnosti za své činy na někoho jiného
    - „Jen dělám svoji práci“
  - Antropomorfizace AI umožňuje obvinění algoritmu z negativních důsledků

EU AI Act

- První právně závazný předpis regulující používání AI
- Čtyři úrovně rizika

Úroveň rizika	Oblast použití	Požadavky
Neakceptovatelné	Sociální skóre	Úplný zákaz
Vysoké	Kritická infrastruktura, vzdělávání, bezpečnostní součástky, trh práce, finanční služby, právo, migrace,…	Systém řízení rizik Kvalita trénovacích dat Logování aktivit Podrobná dokumentace Jasné informace pro uživatele Dohled člověka Robustnost, bezpečnost, přesnost
Omezené	Chatboti	Transparentnost
Nízké nebo žádné	SPAM filtry, počítačové hry	Žádná regulace

Dilemma Game

Obrázky generovány nástrojem Midjourney /imagine people arguing ethical dilemmas

Dilemma Game: Nestát příliš blízko

Právě jsem začal(a) doktorské studium a skvěle se mi spolupracuje s mým školitelem. Doslechl jsem se, že má intimní vztah s jednou ze svých doktorandek. Osobně jsem si ničeho neobvyklého nevšiml(a), i když je pravda, že jí s výzkumem hodně pomáhá. Včera, když jsem odcházel(a) pozdě večer, zahlédl(a) jsem je, jak stojí velmi blízko sebe. Nevím, co přesně se odehrávalo, ale je jisté, že to nebyl rozhovor o výzkumu. Co mám dělat?

A. Řeknu školiteli, že by měl ukončit vztah nebo svoji školitelskou roli. Pokud nebude souhlasit, informaci zveřejním.
B. Informuji příslušného proděkana.
C. Nechám to být, je to jejich soukromá záležitost.
D. Promluvím si s danou doktorandkou a řeknu jí, že tohle je zdroj problémů. Rozhodnutí však nechám na ni.

Příští přednáška

- Profesní etika v IT, aneb
  Jak se pozná „dobrý informatik“?

- Přečíst ACM Code of Ethics and Professional Conduct
  - https://www.acm.org/code-of-ethics