MUNI
FI

Etika umělé inteligence
Profesní etika v IT

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

Osnova dnešní přednášky

- Prezentace Vojtěcha Jelínka: Ethical source
- Etika umělé inteligence
- Profesní etika v IT
- Dilemma game: Nestát příliš blízko
- Ukončení předmětu

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, nemůžeme očekávat, 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

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/

Profesní etika v IT

Aneb Jak se pozná dobrý informatik?

Co vlastně znamená „dobrý informatik”?

- Co znamená „dobrý/dobrá/dobré“?
- Co všechno může být dobré?
- Dobré srdce, dobré plíce, dobrý přítel, dobrý vědec, dobré auto, dobrá paměť, dobré odpoledne
- Slovo „dobré“ je velmi flexibilní

Typologie „dobrosti“

- Georg Henrik von Wright: Varieties of Goodness

- Instrumentální
  - Dobrý nůž / kladivo / kolo / počítač
- Technický (řemeslný)
  - Dobrý zedník / učitel / politik / programátor
- Funkční
  - Dobrý zrak / plíce / paměť
- Hedonický
  - Dobrá chuť / zábava / sex
- Užitečný
  - Dobrá rada / zprávy
- Morální
  - Dobrý člověk / vědec / kolega

- Které z uvedených kategorií „dobroty“ jsou zastoupené v IT? Jak?
- Které jsou nejdůležitější?
- Které jsou nejvíce opomíjené?

Integrity → Integrita, etika, sounáležitost

- Compliance with ethical and professional principles, standards, practices and consistent system of values, that serves as guidance for making decisions and taking actions.
- Dodržování etických a profesních zásad a postupů a konzistentní systém hodnot, které slouží jako vodítko pro rozhodování a pro jednání.

(definice dle European Network for Academic Integrity)

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

Hodnoty / ctnosti

- Základní hodnoty akademické etiky podle ICAI
  - Poctivost, důvěra, férovost, respekt, odpovědnost, odvaha
- Ctnosti podle Embassy of good science
  - Zodpovědnost, dostupnost, účelnost, kolegialita, odbornost, úcta k pravidlům, odvaha, kreativita, kritické myšlení, zvídavost, píle, svědomitost, empatie, férovost, poctivost, skromnost, loajalita, rozvaha, mravnost, objektivita, otevřená mysl, trpělivost, vytrvalost, pozitivita, přesnost, sebereflexe, spolehlivost, rozhodnost, zdvořilost, odpovědnost, obětavost, nezištnost, upřímnost, důkladnost, transparentnost, důvěryhodnost
- Jsou tyto hodnoty a ctnosti relevantní i pro IT?

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.

Ukončení předmětu

Autor: petesaloutos | Getty Images/iStockphoto

Požadavky na ukončení předmětu

- Předpokládá se zájem o téma
- Požadavek na kolokvium: 60 bodů ze 110 možných
- 30 bodů: Aktivní účast na přednáškách
  - 10 x 2 body: Odpovědníky/KvISy k zadaným článkům
  - 10 bodů: Zapojení do diskuse
- 10 bodů: Online aktivita
  - Tipy na zajímavé články do MS Teams
  - Příspěvky na sociálních sítích, blogu, atd.
- 20 bodů: ChatGPT ve zvoleném předmětu
- 40 bodů (nebo 2 x 20 bodů): Vlastní zpracování zvoleného tématu
  - Buď esej na dané téma
  - Nebo prezentace tématu na přednášce
- 10 bodů: Kritická zpětná vazba k eseji někoho jiného

Termíny

- ChatGPT ve zvoleném předmětu do 31. 3.
  - Za každý den zpoždění penalizace 2 body
- Aktivní účast na přednáškách do 15. 5.
- Online aktivita do 31. 5.
  - Přepočet průběžně udělovaných bodů do intervalu <0; 10>
- Odevzdání esejů do 4. 6.
  - Za každý den zpoždění penalizace 2 body
- Odevzdání zpětné vazby do 26. 6.
  - Za každý den zpoždění penalizace 2 body

- Uzavření hodnocení předmětu do 30. 6.

Požadavky na esej

- Smyslem je důkladné porozumění zvolenému tématu
  - Rozebíraného v rámci předmětu → Zpracování do větší hloubky
  - Jiné související téma
  - Případová studie
- Rozsah 3000 – 4000 slov (nebo 2x 1500 – 2000 slov)
- Dodržování zásad akademického psaní
  - Stavění vlastních myšlenek na tom, co bylo publikováno dříve
  - Oddělit vlastní myšlenky od cizích
  - Jasně vyznačit, co je produktem umělé inteligence
  - Pečlivé odkazování na použité zdroje
- Jazyk: čeština, slovenština, angličtina

Požadavky na zpětnou vazbu

- Přečíst esej někoho jiného
- Sepsat:
  - Co se mi líbilo a co se mi nelíbilo
  - S čím souhlasím, s čím nesouhlasím a proč
  - Co v eseji na dané téma chybělo či přebývalo
- Rozsah: 500 – 1000 slov
- Jazyk: čeština, slovenština, angličtina
  - Může se lišit od jazyka eseje

Konec ☺

MidJourney /imagine optimistic end