MUNI
FI

Etická hlediska generativní AI v univerzitním vzdělávání

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

Generativní AI a univerzity

- Pokroky v generativní AI bez většího zájmu
- Listopad 2022: ChatGPT
- Zděšení, obavy, výzvy, naděje, příležitost…

- Reakce a postoje univerzit různé
  - Žádná
  - Zákaz
  - Deklaratorní dokumenty
  - Změny ve vnitřních předpisech
  - Metodická podpora

Midjourney /imagine large language model with transformer architecture using artificial neural network

Co je AI?

Artificial Intelligence refers to systems that appear to have “intelligent behaviour by analysing their environment and taking actions – with some degree of autonomy – to achieve **specific goals**” (EU, 2018)

Generativní AI: specific goals = generování obsahu

Midjourney /imagine artificial intelligence

Současný stav

- Nástroje založené na umělé inteligenci mohou transformovat nebo generovat **jakýkoliv druh obsahu**: text, obrázky, umění, hudbu, programový kód
- **Je obtížné (až nemožné) rozlišit** obsah vygenerovaný umělou inteligencí od obsahu vytvořeného člověkem
- Široká dostupnost nástrojů **prohlubuje existující hrozby** pro akademickou etiku
  - práce psané na zakázku, fabrikace a falšování dat, atd.

ENAI: Testing of AI Detection Tools

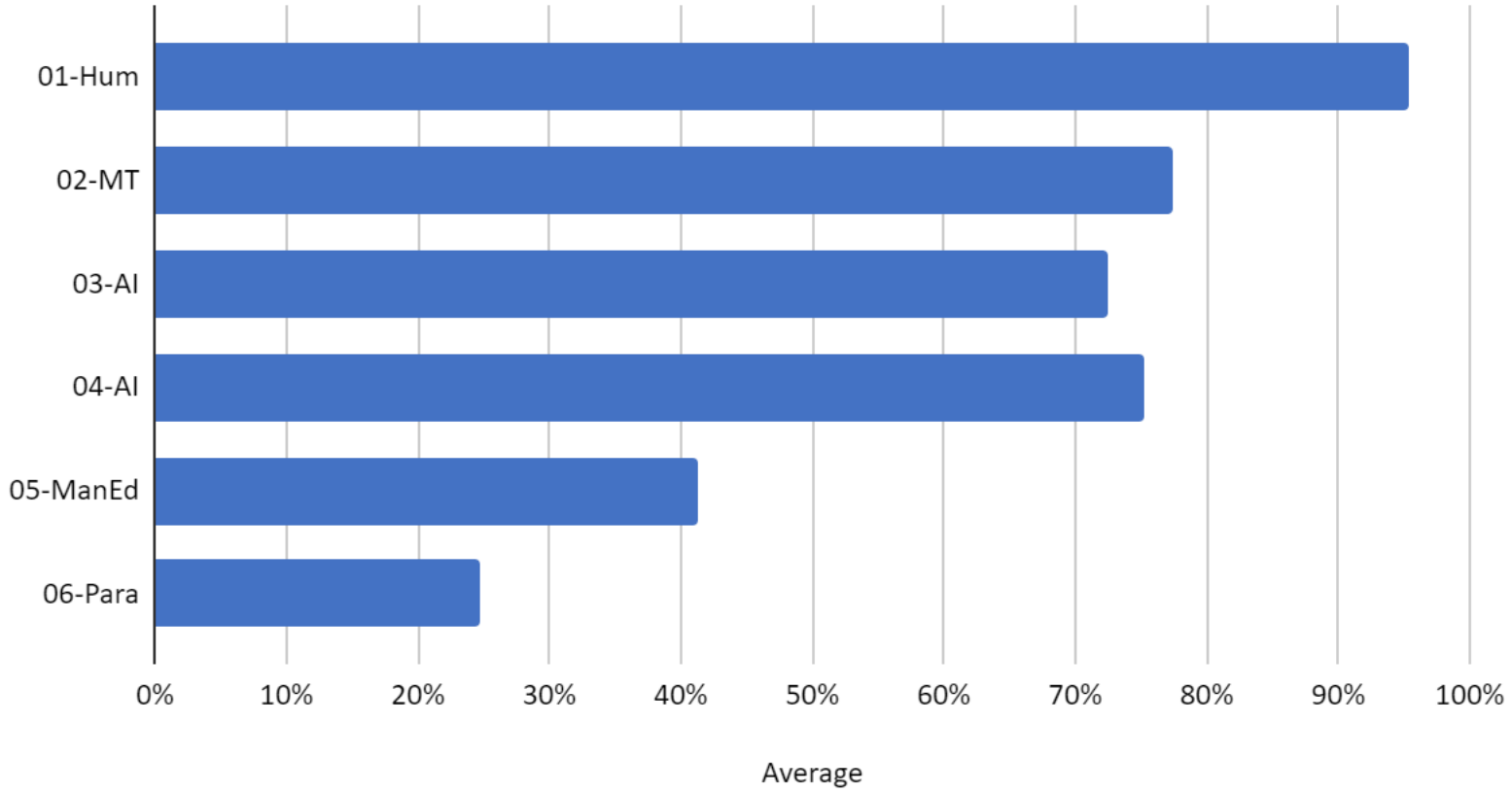
- 12 volně dostupných a 2 komerční nástroje
- 54 dokumentů v 6 kategoriích

  01-Hum: human-written
  02-MT: human-written + machine translation to English
  03-AI: AI-generated text
  04-AI: AI-generated text
  05-ManEd: AI-generated text + manual edits
  06-Para: AI-generated text + machine paraphrase

- Weber-Wulff et al. Testing of Detection Tools for AI-Generated Text
  - https://arxiv.org/abs/2306.15666

Fungují nástroje na detekci textu vytvořeného umělou inteligenci?

- Nefungují ☺
- Zkreslení směrem k “napsáno člověkem”
- I tak produkují falešně pozitivní výsledky
- Neposkytují důkaz
  - Nelze prokázat discipliární přestupek
  - Nemožnost obrany
- Text vytvořený AI a parafrázovaný AI je většinou klasifikován jako napsaný člověkem



Současný stav

- Nástroje založené na umělé inteligenci mohou transformovat nebo generovat **jakýkoliv druh obsahu**: text, obrázky, umění, hudbu, programový kód
- **Je obtížné (až nemožné) rozlišit** obsah vygenerovaný umělou inteligencí od obsahu vytvořeného člověkem
- Široká dostupnost nástrojů **prohlubuje existující hrozby** pro akademickou etiku
  - práce psané na zakázku, fabrikace a falšování dat, atd.

Rizika GAI ve vědeckých publikacích

- Nepřesný, nerelevantní a nepravdivý obsah
  - “any section of a manuscript written by an NLP system should be checked by a domain expert for accuracy, bias, relevance, and reasoning”
  - “If a section of a manuscript written by an NLP system contains errors or biases, coauthors need to be held accountable for its accuracy, cogency, and integrity”
- Vědecká integrita
  - “researchers should not use NLP systems to fabricate empirical data or falsify existing data”

Hosseini, M., Rasmussen, L. M., & Resnik, D. B. (Jan 2023). **Using AI to write scholarly publications**. *Accountability in Research*. https://doi.org/10.1080/08989621.2023.2168535

Kdy je využití AI v pořádku?

- **Dovolené a přiznané** využití nástrojů umělé inteligence je obecně v pořádku
- Kdy se jedná o pochybení?
- Záleží na **smyslu úkolu**
- Pochybení nastává, pokud využití AI
  - není dovoleno, nebo
  - není přiznáno, nebo
  - přináší neférové zvýhodnění

edintegrity.biomedcentral.com/articles/10.1007/s40979-023-00133-4

BMC Part of Springer Nature

International Journal for Educational Integrity

Home About Articles Submission Guidelines Submit manuscript

Editorial | Open Access | Published: 01 May 2023

ENAI Recommendations on the ethical use of Artificial Intelligence in Education

Tomas Foltynek, Sonja Bjelobaba, Irene Glendinning, Zeenath Reza Khan, Rita Santos, Pegi Pavletic & Július Kravjar

International Journal for Educational Integrity **19**, Article number: 12 (2023) | Cite this article

9106 Accesses | **107** Altmetric | Metrics

Unauthorised Content Generation

Production of academic work, in whole or part, for academic credit, progression or award, whether or not a payment or other favour is involved, using unapproved or undeclared human or technological assistance.

Midjourney /imagine artificial intelligence writing student essay in blue and yellow, white background

Co by mělo být přiznáno?

- Jakékoliv využití jiných osob, zdrojů a nástrojů, které ovlivnilo myšlenky nebo vytvářelo obsah, by mělo být řádně přiznáno
  - Tedy i využití AI, pokud ovlivňuje myšlenky nebo obsah, musí být řádně přiznáno
  - Konkrétní podoba se může lišit
  - Pokud je to možné, měl by být uveden i prompt
- Využití služeb, zdrojů a nástrojů, které ovlivňují pouze formu, je v pořádku
  - jazyková a stylistická korektura, slovník synonym,…

Jak přiznat využití AI?

- MU: Doporučení k využití nástrojů umělé inteligence při plnění studijních povinností
  - Obecná deklarace
  - Přímý odkaz dle citační normy
  - Popis využití s podrobným komentářem (metodika)
  - Dataset promptů jako příloha práce
- Hlavní zásada: Být **transparentní**
- Co je **smyslem** odkazování?
  - U “klasických” zdrojů: Dohledat původ myšlenky
  - U generativní AI: Posoudit intelektuální přínos autora

Zodpovědnost

- Výstupy generativní AI mohou být zkreslené, nepřesné nebo nesprávné.
- Ani nástroj, ani jeho poskytovatel nenese zodpovědnost za vygenerovaný obsah
- Zodpovědnost je vždy na uživateli!

- Nástroj **AI nemůže být uveden jako (spolu)autor** publikace
  - Viz kritéria autorství ICMJE/COPE

Co přinese budoucnost?

- AI může přijít s novými objevy
  - kritéria autorství?
  - duševní vlastnictví?
  - recenzní řízení?
- Výstupy budou stále více výsledkem **spolupráce** člověka a AI
  - rozlišování člověk vs. AI pozbyde smysl
- Otázky ohledně **smyslu** práce
  - Co je smyslem semestrální / bakalářské / diplomové práce?
  - Co je smyslem vědeckého článku?
  - Co má být předmětem hodnocení? Text?

Midjourney /imagine bright optimistic future of education and science with artificial intelligence

Důsledky pro výuku

- Jasně studentům vysvětlit **smysl** úkolu
- Jasně **deklarovat**, zda je využití AI vhodné / dovolené
- **Přiznat si**, že nedovolené využití AI nejsme schopní poznat a prokázat
- **Nastavit hodnocení předmětu** tak, aby jeho férovost nebyla ohrožena přispěním někoho / něčeho jiného
- V rámci výuky **ukazovat** vhodné a nevhodné způsoby využití AI

Děkuji za pozornost

foltynek@fi.muni.cz

twitter.com/TFoltynek

MidJourney /imagine optimistic end