MUNI
LAW

Pojmy a dojmy: jak vnímáme dokumenty vytvořené AI?

Jakub HARAŠTA[α], Tereza NOVOTNÁ[α], Jaromír ŠAVELKA[β]

[α] Právnická fakulta, Masarykova Univerzita, ČR
[β] School of Computer Science, Carnegie Mellon University, USA

Jakub Harašta a Tereza Novotná byli podpořeni v rámci GAMU projektu *Forensic Support for Building Trust in Smart Software Ecosystems* (MUNI/G/1142/2022).

Prezentace vychází z preprintu:

HARAŠTA, Jakub, Tereza NOVOTNÁ a Jaromír ŠAVELKA. It Cannot Be Right If It Was Written by AI: On Lawyers' Preferences of Documents Perceived as Authored by an LLM vs a Human. *arXiv*, 2024, dostupné na https://arxiv.org/abs/2407.06798

tiny.cc/pojmy-dojmy

MUNI LAW

AI / (G)AI / LLM

- 30. listopad 2022: ChatGPT
- Označování syntetických výstupů (transparentnost)
- AI Act, …
- Viz Fernández-Llorca, D., Gómez, E., Sánchez, I. et al. An interdisciplinary account of the terminological choices by EU policymakers ahead of the final agreement on the AI Act: AI system, general purpose AI system, foundation model, and generative AI. *Artificial Intelligence and Law*, 2024

AI v právu

- Generování obsahu
  - Sumarizace
  - Překlady
  - Odpovídání na otázky
- Podpora k dílčím krokům
  - Odůvodňování (*legal reasoning*)
  - Podpora pro získávání právních informací
  - Přístup k právu/spravedlnosti
  - Předpovídání rozhodnutí
  - Anotace

AI v právu (II)

- Jak vnímáme obsah generovaný AI?
- Nutnost odlišovat úspěšnost / přesnost (*performance*) od vnímání (*perception*)
- „První dojem“ (sebevědomé vystupování; gramatické chyby; nesprávně používaná terminologie) → hraje roli u lidí; hraje roli u AI?

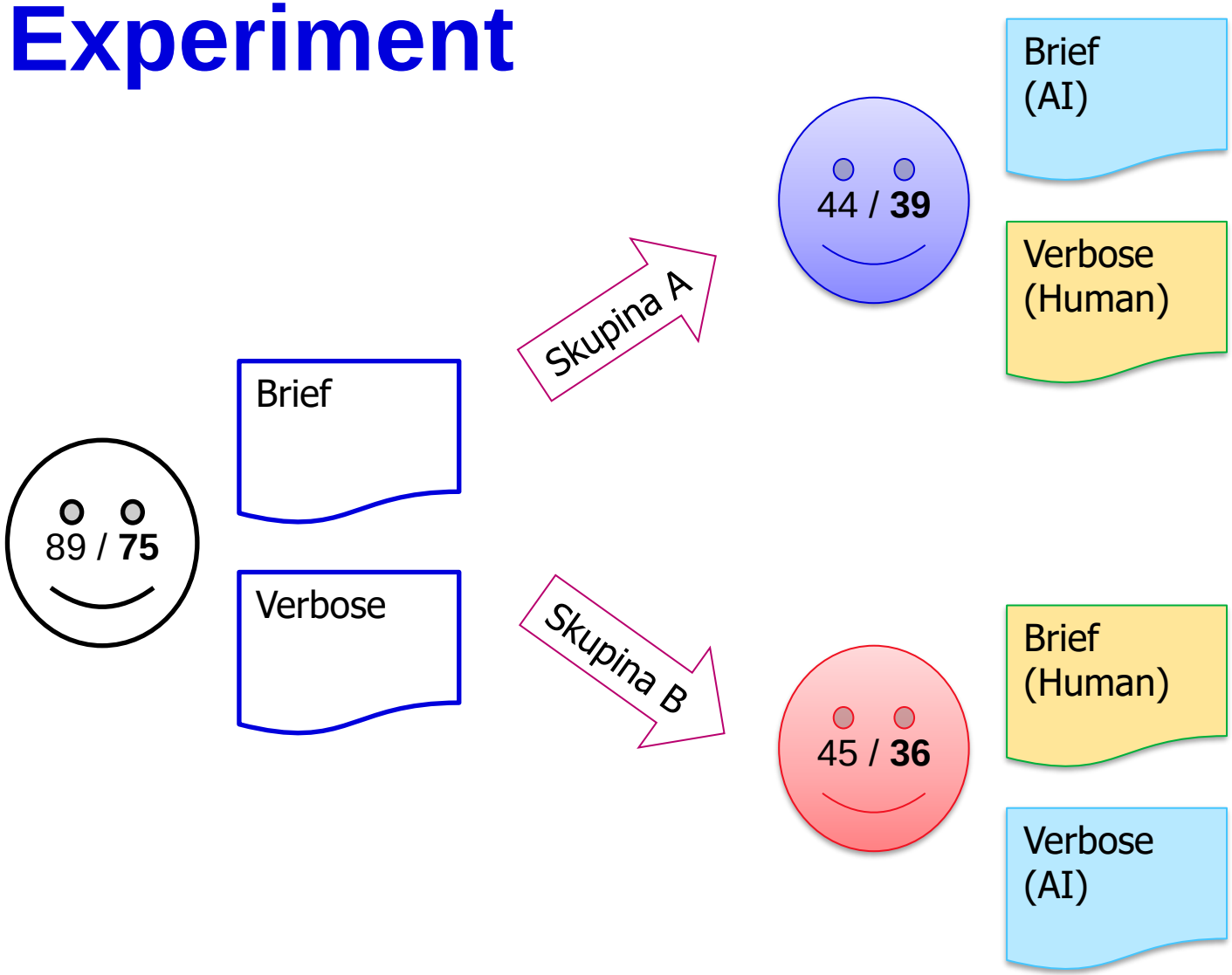
Vnímání obsahu

- Rizika, příležitosti v komunikaci zprostředkované AI (např. snižování důvěry v komunikaci vs. posílení komunikačních schopností) [Hancock et al., 2020]
- Nižší důvěra k AI [von Eschenbach, 2021; Castelo and Ward, 2021] (a algoritmům obecně [Jussupow et al., 2020])
- Rozhodování s morálním rozměrem snižuje důvěru [Bigman and Gray, 2018]
- Část viny za chyby v komunikaci přisuzujeme AI [Hohenstein and Jung, 2020]
- Nižší hodnocení obsahu generovaného AI v případě profilů na Airbnb [Jakesch et al., 2019], emailů [Liu et al. 2022], hudby [Shank et al., 2023], překladů [Asscher a Glikson, 2023], novinových zpráv [Waddell, 2018] nebo informací o zdravotních rizicích [Lim a Schmälzle, 2024]

Vnímání obsahu (II)

- Úroveň nedůvěry se liší když plněné úkoly vnímáme jako objektivní / subjektivní [Castelo et al., 2019]
- Halucinace [Cheong et al., 2024; Magesh et al., 2024]
- Nižší vzdělání zvyšuje riziko konzumování nespolehlivého obsahu [Oviedo-Trespalacios et al., 2023]
- Hodnocení správnosti odpovědi je náročné i pro VŠ studenty v rámci jejich oboru [Dahlkemper et al., 2023]

Experiment



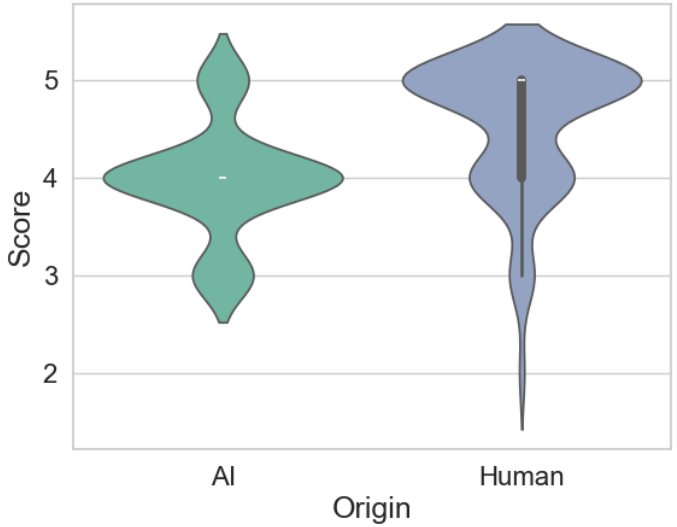
MUNI LAW

Hodnocení dokumentů

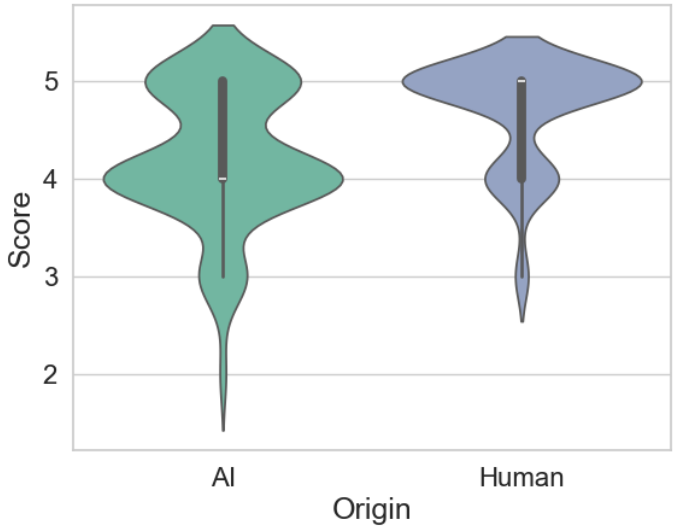
- Ohodnoťte jazykovou kvalitu dokumentů (uzavřená; 1 – nejhorší; 5 – nejlepší)
- Ohodnoťte správnost dokumentů (uzavřená; 1 – nejhorší; 5 – nejlepší)
- Okomentujte své hodnocení (otevřená; max. 100 slov)
- Zhodnoťte, zda je plná automatizace možná (otevřená; bez limitu)

AI vs. lidé (kvantitativně)

- Jazyková kvalita:

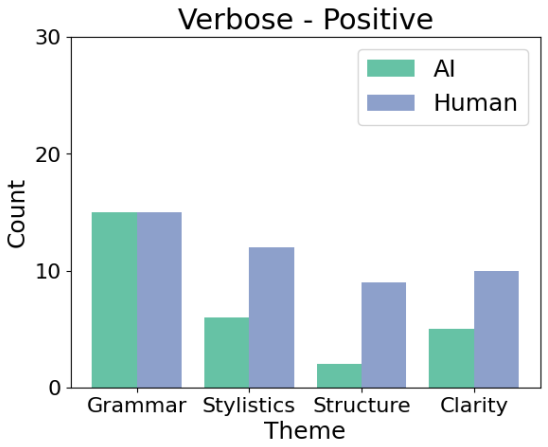


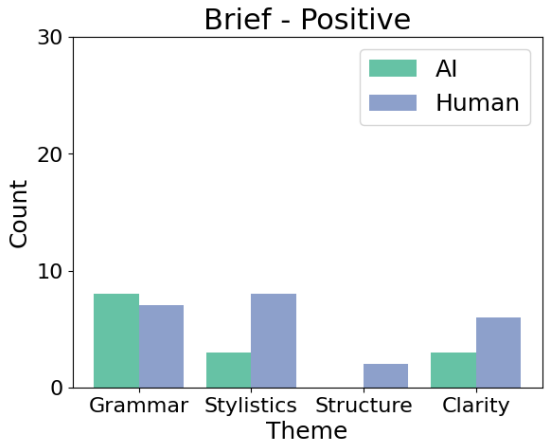
- Správnost:

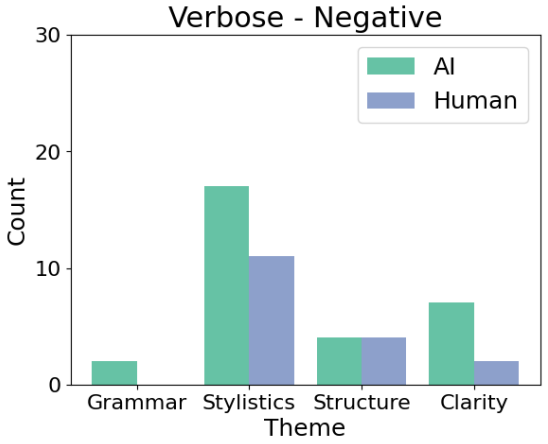


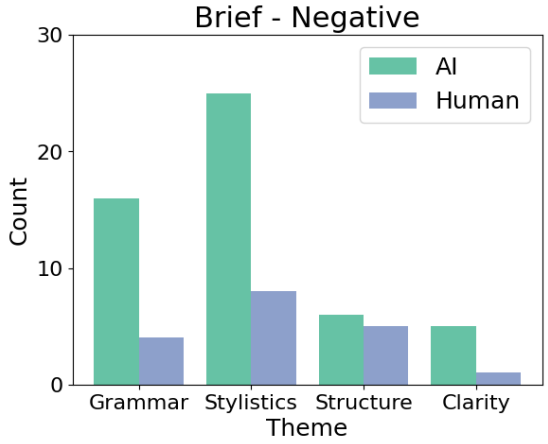
MUNI LAW

AI vs. lidé (kvalitativně; jazyková kvalita)



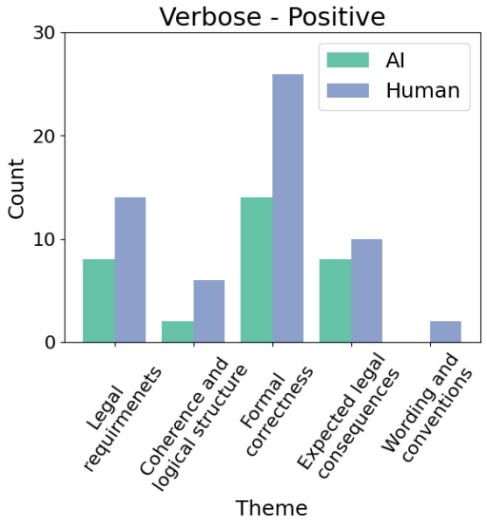


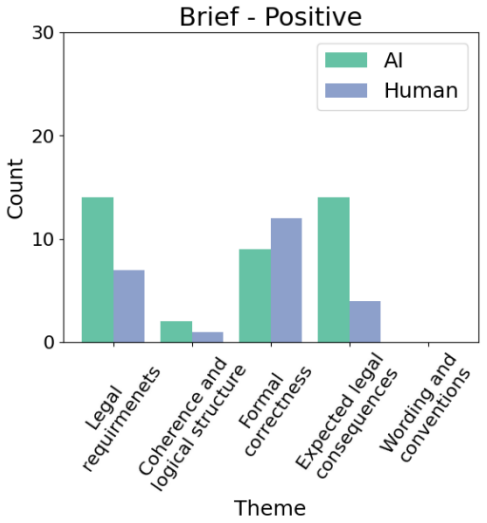


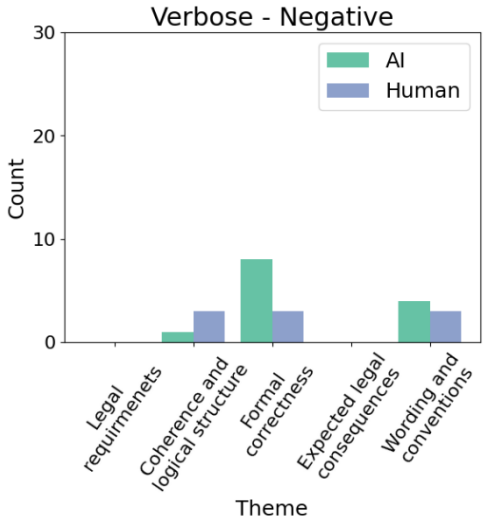


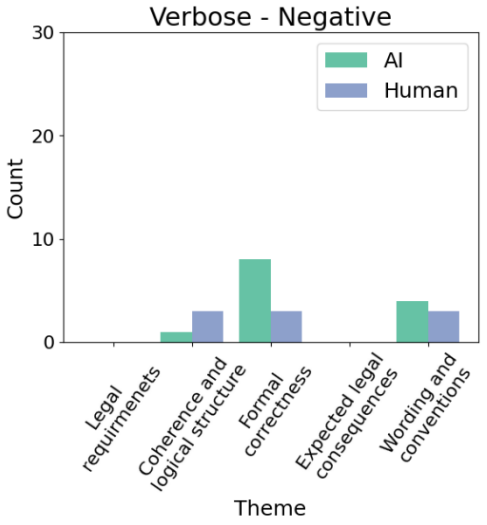
MUNI LAW

AI vs. lidé (kvalitativně; správnost)



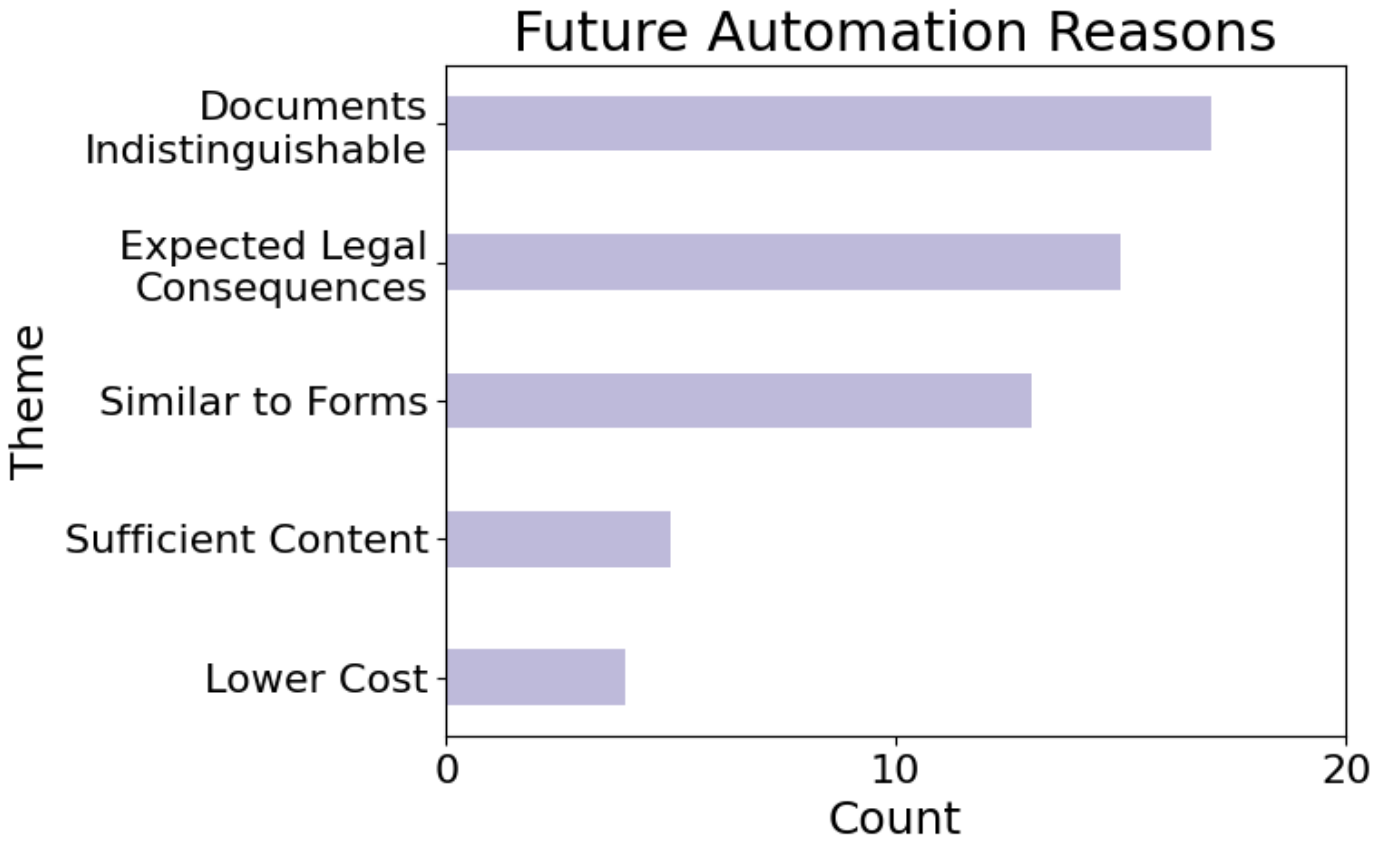






MUNI LAW

Budoucnost automatizace



MUNI LAW

Závěr

- Dokumenty prezentované jako generované AI jsou uživateli hodnocené hůře
- Část výsledků je statisticky signifikantní (vyšší počet respondentů?)
- Rozdíl mezi studenty a právníky (další skupiny?)
- Jednoduchý dokument (složitější dokumenty?)
- Relativně přímočaré (využití existujících modelů, např. TAM?)
- Nevyhnutelná automatizace (detailnější pohled?)
- Implikace?

Reference

- [Asscher a Glikson, 2023] Asscher O, Glikson E (2023) Human evaluations of machine translation in an ethically charged situation. *New Media & Society* 25(5):1087–1107.
- [Bigman and Gray, 2018] Bigman YE, Gray K (2018) People are averse to machines making moral decisions. *Cognition* 181:21–34.
- [Castelo and Ward, 2021] Castelo N, Ward AF (2021) Conservatism Predicts Aversion to Consequential Artificial Intelligence. *PLOS ONE* 16(12):1–19.
- [Castelo et al., 2019] Castelo N, Bos MW, Lehmann DR (2019) Task-Dependent Algorithm Aversion. *Journal of Marketing Research* 56(5):809–825.
- [Cheong et al., 2024] Cheong I, Xia K, Feng KJK, et al (2024) (A)I am not a lawyer, but...: Engaging legal experts towards responsible LLM policies for legal advice. *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24),* p 2454–2469.
- [Dahlkemper et al., 2023] Dahlkemper MN, Lahme SZ, Klein P (2023) How do physics students evaluate artificial intelligence responses on comprehension questions? A study on the perceived scientific accuracy and linguistic quality of ChatGPT. *Physical Review Physics Education Research* 19:010142/1–25.
- [Hancock et al., 2020] Hancock JT, Naaman M, Levy K (2020) AI-Mediated Communication: Definition, Research Agenda, and Ethical Considerations. *Journal of Computer-Mediated Communication* 25(1):89–100.
- [Hohenstein and Jung, 2020] Hohenstein J, Jung M (2020) AI as a moral crumple zone: The effects of AImediated communication on attribution and trust. *Computers in Human Behavior* 106:106190.
- [Jakesch et al., 2019] Jakesch M, French M, Ma X, et al (2019) AI-Mediated Communication: How the Perception that Profile Text was Written by AI Affects Trustworthiness. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI'19)*, p 1–13.

MUNI LAW

Reference (II)

- [Jussupow et al., 2020] Jussupow E, Benbasat I, Heinzl A (2020) Why Are We Averse Towards Algorithms? A Comprehensive Literature Review on Algorithm Aversion. *Proceedings of the 28th European Conference on Information Systems (ECIS)*.
- [Lim a Schmälzle, 2024] Lim S, Schmälzle R (2024) The Effect of Source Disclosure on Evaluation of AI-Generated Messages: A Two-part Study. *Computers in Human Behavior: Artificial Humans* 2(1):100058.
- [Liu et al. 2022] Liu Y, Mittal A, Yang D, et al (2022) Will AI Console Me when I Lose my Pet? Understanding Perceptions of AI-Mediated Email Writing. *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems (CHI'22).*
- [Magesh et al., 2024] Magesh V, Surani F, Dahl M, et al (2024) Hallucination-free? Assessing the reliability of leading AI legal research tools. arXiv:2405.20362.
- [Oviedo-Trespalacios et al., 2023] Oviedo-Trespalacios O, Peden AE, Cole-Hunter T, et al (2023) The risks of using ChatGPT to obtain common safety-related information and advice. *Safety Science* 167:106244.
- [Shank et al., 2023] Shank DB, Stefanik C, Stuhlsatz C, et al (2023) AI composer bias: Listeners like music less when they think it was composed by an AI. *Journal of Experimental Psychology: Applied* 29(3):676–692.
- [von Eschenbach, 2021] von Eschenbach WJ (2021) Transparency and the Black Box Problem: Why We Do Not Trust AI. *Philosophy & Technology* 34(4):1607–1622.
- [Waddell, 2018] Waddell TF (2018) A Robot Wrote This? How perceived machine authorship affects news credibility. *Digital Journalism* 6(2):236–255.

MUNI LAW

MUNI
LAW

Děkuji Vám za pozornost!

Otázky?

jakub.harasta@law.muni.cz

Jakub Harašta a Tereza Novotná byli podpořeni v rámci GAMU projektu *Forensic Support for Building Trust in Smart Software Ecosystems* (MUNI/G/1142/2022).