教育における生成AI活用推進リーダープログラム

生成AIの注意点・リスク

バイアス・毒性

吉田塁

東京大学 大学院工学系研究科 准教授

LLM 寄附講座 特任准教授

2026年4月

バイアス

- **判断や出力における体系的な偏りや歪み**

- **具体例**
  - Midjourneyを使ったSTEM分野の人物画像で、生成された肖像画はほぼすべてが白人男性（Messingschlager & Appel 2025）
  - 14の大規模言語モデルをテストし、OpenAIのChatGPTとGPT-4が最もリベラル寄り、MetaのLLaMAが最も保守寄り（Feng et al. 2023）

- **主な原因（Gallegos et al. 2024）**
  - 学習データ、モデルの学習・推論方法、評価、運用

毒性（Toxicity）

- **有害・不適切なコンテンツをAIが出力すること**

- **具体例**
  - ChatGPTと自殺問題
    - 16歳の少年がChatGPTと自殺について会話を続けた末に命を落とし、遺族がOpenAIを提訴。チャットログには、自殺念慮を親に相談しないよう促す内容が含まれていたと報じられている
  - ニューヨーク市「MyCity」チャットボット
    - 「事業主は従業員のチップを徴収できる」「セクハラを訴えた従業員を解雇できる」など、法的に誤った情報を提供していたことが発覚

- **主な原因（Gehman et al. 2020, Ouyang et al. 2022, Luong et al. 2024）**
  - 学習データ、学習アルゴリズム、ジェイルブレイク

参考文献

- Feng, S., Park, C. Y., Liu, Y., & Tsvetkov, Y. (2023, July). From pretraining data to language models to downstream tasks: Tracking the trails of political biases leading to unfair NLP models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 11737-11762).
- Gallegos, I. O., Rossi, R. A., Barrow, J., Tanjim, M. M., Kim, S., Dernoncourt, F., ... & Ahmed, N. K. (2024). Bias and fairness in large language models: A survey. *Computational linguistics*, *50*(3), 1097-1179.
- Gehman, S., Gururangan, S., Sap, M., Choi, Y., & Smith, N. A. (2020, November). Realtoxicityprompts: Evaluating neural toxic degeneration in language models. In *Findings of the association for computational linguistics: EMNLP 2020* (pp. 3356-3369).
- Luong, T., Le, T. T., Ngo, L., & Nguyen, T. (2024, August). Realistic evaluation of toxicity in large language models. In *Findings of the Association for Computational Linguistics: ACL 2024* (pp. 1038-1047).
- Messingschlager, T. V., & Appel, M. (2025). Algorithmic bias in image-generating artificial intelligence: prevalence and user perceptions. Information, Communication & Society, 1-23.
- Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., ... & Lowe, R. (2022). Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, *35*, 27730-27744.