大阪大学 社会技術共創研究センター
Research Center on Ethical, Legal and Social Issues

2024年度日本リスク学会第37シンポジウム
「AIのリスクを考える：生体認証技術から生成AIまで」

（倫理的・法的・社会的課題）

生成AIのELSIリスクの概要

カテライ　アメリア
大阪大学　　社会技術共創研究センター（ELSIセンター）　特任助教

生成AIのELSIリスク

生成AIに 膨大な投資が行われている一方で、生成AIのリスクに関する懸念も

→ 先行研究で報告されている生成AIのELSI（倫理的・法的・社会的）リスクを抽出・分類

→ E、L、Sのそれぞれの観点からではなく、包括的に生成AIのリスクを検討するための文献レビューを実施

対象：「生成AI」→ テキスト生成AI＋画像生成ＡＩ　・　海外の学術文献を中心に

メソッド：文系（リスク学・社会学など）・理系（情報学）の専門家を交えた学際的な研究チームで先行研究を分析

論文：Katirai A et al. Situating the social issues of image generation models in the model life cycle: a sociotechnical approach. arXiv preprint arXiv:2311.18345. 2023 Nov 30.

８つのELSIリスク

バイアス

プライバシー

情報環境への影響

高度なAIシステム特有のリスク

自然環境への影響

データに関する課題

知的財産権の問題

文化的環境への影響

論点①　バイアス

論点① バイアス（テキスト生成）

OpenAI（2023）：「**GPT-4が社会的バイアスおよび特定の世界観を強化し、特に、疎外されたグループに対する、有害なステレオタイプおよび侮辱的なコンテンツを再生産する**」

→ 機会や資源の提供に関する意思決定や情報提供でGPT-4が使われることによる危害を懸念する

研究①：Ferrera（2023）**ChatGPTに6つの種類のバイアス**

- Demographic （人口統計的バイアス）, Cultural（文化的バイアス）, Linguistic （言語的バイアス）,Temporal （時間的バイアス）. Confirmation（確証バイアス）, Ideological & Political （イデオロギー・政治的バイアス）

研究②：Nozza et al.（2022）**LGBTQIA＋に対するバイアス**

- 複数のテキスト生成AIのアウトプットを分析
- LGBTQIA＋に関する未完成の文章が、AIによって完成された際、生成された単語は７％の確率で有害であり、生成された文章は 13%の確率で有害であった

研究③：Abid et al.（2021）**イスラム教徒に対するバイアス**

- テキスト生成AIはイスラム教徒と暴力の持続的な関連付けといったような有害なステレオタイプを表示しがち
- AIの有害なバイアスを体系的に低減するための新しいアプローチが必要

研究④：Takeshita et al. (2022) **種差別**

- テキスト生成AIは、非人間動物を有害な単語と結びつける傾向があり、種差別的な言語を使用する傾向がある

Abid et al., 2021; Nozza et al., 2022; Takeshita et al., 2022; Ferrera, 2023; OpenAI, 2023

論点① バイアス（画像生成）

研究の一部　→ 更なる研究の実施が求められている

1. **Steed and Caliskan (2021): 人種、ジェンダーなど、複数のバイアスの関わり合い**は、教師なし学習を用いるImageNet のデータセットによって自動的に学習されている
2. **Srinivasan and Uchino (2021):** 生成されるアートに**人種およびジェンダーバイアス**が埋め込まれるリスクを確認し、**歴史的出来事が不正確に描写されるリスク**を指摘する
3. **Bianchi et al. (2022):** 画像生成AIによって生成される人物の描写に**人種、社会経済的地位、ジェンダー、性的指向を含む複数の種類のバイアス**が反映されるとし、建物や車などの**背景にある要素の描写にも現れている**ことが明らかに
4. **Offert and Phan (2022): Whiteness（白人性）は非常に耐久性が高く**、DALL-E 2 のプロンプトの最後にランダム化された非表示のキーワードが解決策として配置されていることを確認
5. **Cheong et al. (2023):** DALL-E Miniのアウトプットは**人種とジェンダーバイアス**等を含み、AI を介したフィードバック ループが生じている
6. **Luccioni et al. (2023):** 機械学習全体には、さまざまな**性格特性とジェンダー**に関するステレオタイプを含む、**人種、ジェンダー、外見**に関する偏見と不公平を増幅するリスク
7. **Garcia et al. (2023): 画像生成AIの基盤となるデータにおける社会的バイアスが反映されている**

Srinivasan and Uchino, 2021; Steed and Caliskan, 2021; Bianchi et al., 2022; Offert and Phan, 2022; Cheong et al., 2023; Garcia et al., 2023

論点① バイアス

バイアスの影響

- 既存の社会的不平等を反映するだけではなく、増幅させてしまう
- 生成された画像は様々な場面で活用される可能性（医療分野や犯罪者捜索等でも）
- → 生成AIに埋め込まれた社会的バイアスが、人間ユーザーによるバイアスのある決定を促し、その結果、**AIと社会全体の両方でバイアスがさらに固定化される**
- 問題のある描写自体で生じる危害（representational harm）や重要な決定に活用されることで生じる危害も（allocational harm）

「世界観」の問題

- 各AIには、ある種の「世界観」が反映されており、特に「初心者ユーザー」にとって不透明

バイアスの是正が困難

- AIの開発過程で重要となる、データのタグ付で使われる「言語」自体にバイアスが存在する
- 曖昧なインプットを回避することが不可能

Adams et al., 2023; Fraser et al.,2023; Luccioni et al., 2023; Munn et al., 2023; Ungless et al., 2023

論点②　プライバシー

論点② プライバシー

テキスト生成

- 想定される４つのプライバシーリスク：
  1.公開データの悪用によるプライバシー侵害
  2.個人が入力するデータの悪用によるプライバシー侵害
  3.システムに対するプライバシー攻撃
  4.透明性の欠如

画像生成

- Stable DiffusionやImagenなどのdiffusion モデルは学習データを記憶 (memorize) し、再生成 (regenerate)する
- →学習で使われているイメージがそのままアウトプットされてしまうリスク
- 機微データのアウトプットや個人の肖像の利用に関する懸念も

学習データセットLAIONには個人の医療画像が含まれている↓

Enlarge / The LAION data set is replete with potentially sensitive images collected from the Internet, such as these, which are now being integrated into commercial machine learning products. Black bars have been added by Ars for privacy purposes.

出典：https://arstechnica.com/information-technology/2022/09/artist-finds-private-medical-record-photos-in-popular-ai-training-data-set/

プライバシー懸念に
「完全に対処することはほぼ不可能」だが、
取り組みもある

Edwards, 2022; Nightingale and Farid, 2022; Carlini et al., 2023; Struppek et al., 2023; OpenAI, 2023; Wu et al., 2023

MATT BURGESS SECURITY MAR 11, 2024 3:00 AM

Google Is Getting Thousands of Deepfake Porn Complaints

Content creators are using copyright laws to get nonconsensual deepfakes removed from the web. With the complaints covering nearly 30,000 URLs, experts say Google should do more to help.

出典：https://www.wired.com/story/google-deepfake-porn-dmca-takedowns/

出典： https://www.technologyreview.com/2024/01/29/1087325/three-ways-we-can-fight-deepfake-porn-taylors-version/

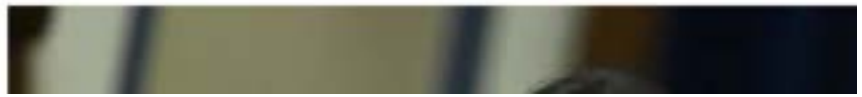
Alexandria Ocasio-Cortez

Edward Helmore

Tue 9 Apr 2024 17.16 BST

Alexandria Ocasio-Cortez recounts horror of seeing herself in 'deepfake porn'

Progressive New York congresswoman describes resurfacing of trauma when seeing explicit materials with her image transposed on to them



出典：https://www.theguardian.com/us-news/2024/apr/09/alexandria-ocasio-cortez-deepfake-porn

MATT BURGESS SECURITY OCT 16, 2023 7:00 AM

Deepfake Porn Is Out of Control

New research shows the number of deepfake videos is skyrocketing—and the world's biggest search engines are funneling clicks to dozens of sites dedicated to the nonconsensual fakes.

出典： https://www.wired.com/story/deepfake-porn-is-out-of-control/

同意のない性的な画像の生成も広まっている

LYDIA MORRISH BUSINESS MAR 6, 2024 7:00 AM

The Dark Side of Open Source AI Image Generators

Open source tools allow anyone to make AI art. They have also been used to produce nonconsensual deepfake porn.

出典：https://www.wired.com/story/dark-side-open-source-ai-image-generators/

ELSI

Birhane et al., 2021; Burgess, 2023; Burgess, 2024; Heikkila, 2024; Helmore, 2024; Morrish, 2024; The Verge, 2023; Thiel et al., 2023

論点③　情報環境への影響

論点③ 情報環境への影響（テキスト生成）

- テキスト生成AIでは、単語がベースになっているのではなく、**複数の文字が組み合わさったトークン**がベース
  - 学習データ内で、これらのトークンがどの頻度で、どう組み合わさっているかを含む、**データの傾向**が生成AIの機能を決定させ、生成物の内容も左右させる
- 無意味や誤ったコンテンツが生成されることは生成AIの問題の一つと認識されている
  - この現象は**幻覚（hallucination）**と呼ばれることが多いが、 幻覚は心理学からの概念であり、人間の心理のあり方に深く根付いている概念でもあるため、問題視されている
- **生成AIのアウトプットの全てをある種の「幻覚」として考えられる**
- **生成AIによる誤情報の生成がいずれは改善される問題ではない** → 現状のような仕組みに基づいて機能している限り、必然的に起こる問題と考えられている
- テキスト生成AIのアウトプットが**人間によって作成された文章よりも信頼されやすい**傾向が見られている
- → 生成されるアウトプットの納得性が向上し、過剰依存が生じるほど、誤情報・偽情報が問題になる

生成AIからのアウトプットのみならず、**情報環境全体への悪影響が懸念されている**

Bender et al., 2021; Nichol, 2022; Chen et al., 2023; Klein, 2023; Munn et al., 2023; Mokander et al., 2023; OpenAI, 2023; Spitale et al., 2023; Zhang et al., 2023; Stokel-Walker, 2024

論点③ 情報環境への影響（画像生成）

画像等を「真実」として受けられないため、新しいノルムが確立される

- 例① 生成された顔画像は、非常に写実的であるだけでなく、本物の顔とほぼ区別がつかず、より信頼に値すると判断される
- 例② 個人の不正行為への関わりを描写する画像の生成　→　詐欺や偽情報拡散に利用される可能性
- 例③「嘘つきの配当金」（liar's dividend）→　不正行為への関わりが記録された場合、証拠に疑問を投げかけられる

個人・社会・民主主義に深刻な影響が及ぼされることが見込まれている（⇨報告されている）

- 画像生成AIへのアクセスの容易さ、AIの民主化・AIに対する意識の向上で問題が悪化

Agarwal et al., 2019; Chesney and Citron, 2019; Ko et al., 2023; Nightingale and Farid, 2019; Verdoliva, 2020; Oppenlaender et al., 2023; Ricker et al., 2023; Sha et al., 2023; Struppek et al., 2022

2024年は選挙に関して記録的な一年

⇨　50カ国において、世界人口の半分以上を代表し、20億人以上が投票に向かう中、生成AIを用いた政治的な偽情報・誤情報の拡散が懸念されている


WORLD VIEW | 09 April 2024

AI-fuelled election campaigns are here — where are the rules?

Political candidates are increasingly using AI-generated 'softfakes' to boost their campaigns. This raises deep ethical concerns.

By Rumman Chowdhury

出典：
https://www.nature.com/articles/d41586-024-00995-9


What to Do About the Junkification of the Internet

Mar 12, 2024

ETHICS AND GOVERNANCE OF AI | INTERNET HEALTH | MEDIA, DEMOCRACY, & PUBLIC DISCOURSE

Nathaniel Lubin

出典：
https://cyber.harvard.edu/story/2024-03/what-do-about-junkification-internet


How AI shaped Milei's path to Argentina presidency

出典：
https://www.japantimes.co.jp/news/2023/11/22/world/politics/ai-javier-milei-argentina-presidency/


Generative AI may change elections this year. Indonesia shows how

By Kate Lamb, Fanny Potkin and Ananda Teresia

February 8, 2024 8:59 PM GMT+9 · Updated 2 months ago

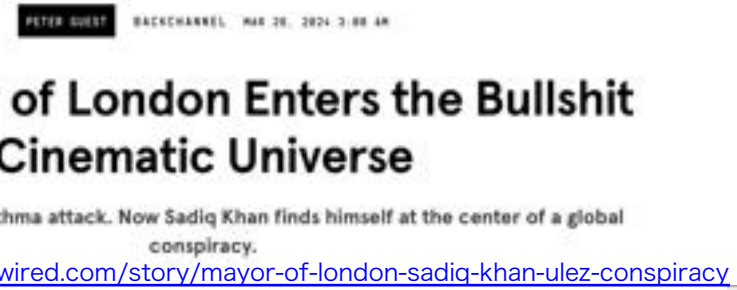
出典：
https://foreignpolicy.com/2024/02/14/prabowo-indonesia-election-democracy-jokowi/


NILESH CHRISTOPHER | VARSHA BANSAL | BACKCHANNEL | MAY 20, 2024 6:00 AM

Indian Voters Are Being Bombarded With Millions of Deepfakes. Political Candidates Approve

India's elections are a glimpse of the AI-driven future of democracy. Politicians are using audio and video deepfakes of themselves to reach voters—who may have no idea they've been talking to a clone.

出典：https://www.wired.com/story/indian-elections-ai-deepfakes/


PETER GUEST | BACKCHANNEL | MAR 28, 2024 3:00 AM

The Mayor of London Enters the Bullshit Cinematic Universe

It all started with an asthma attack. Now Sadiq Khan finds himself at the center of a global conspiracy.

出典：https://www.wired.com/story/mayor-of-london-sadiq-khan-ulez-conspiracy

Feliba, 2023; Chowdhury, 2024; Chrisotpher and Bansal, 2024; Guest, 2024; Lamb et al., 2024; Lubin, 2024; Masterson, 2024

論点④　高度なAIシステムで生じるリスク

論点④ 高度なAIシステムで生じるリスク

AIの中央集権化

- 社会に対するインパクトの大きな技術の開発・実装に関する決定が民間企業に委ねられてしまう

開発の高速化による安全基準の低下

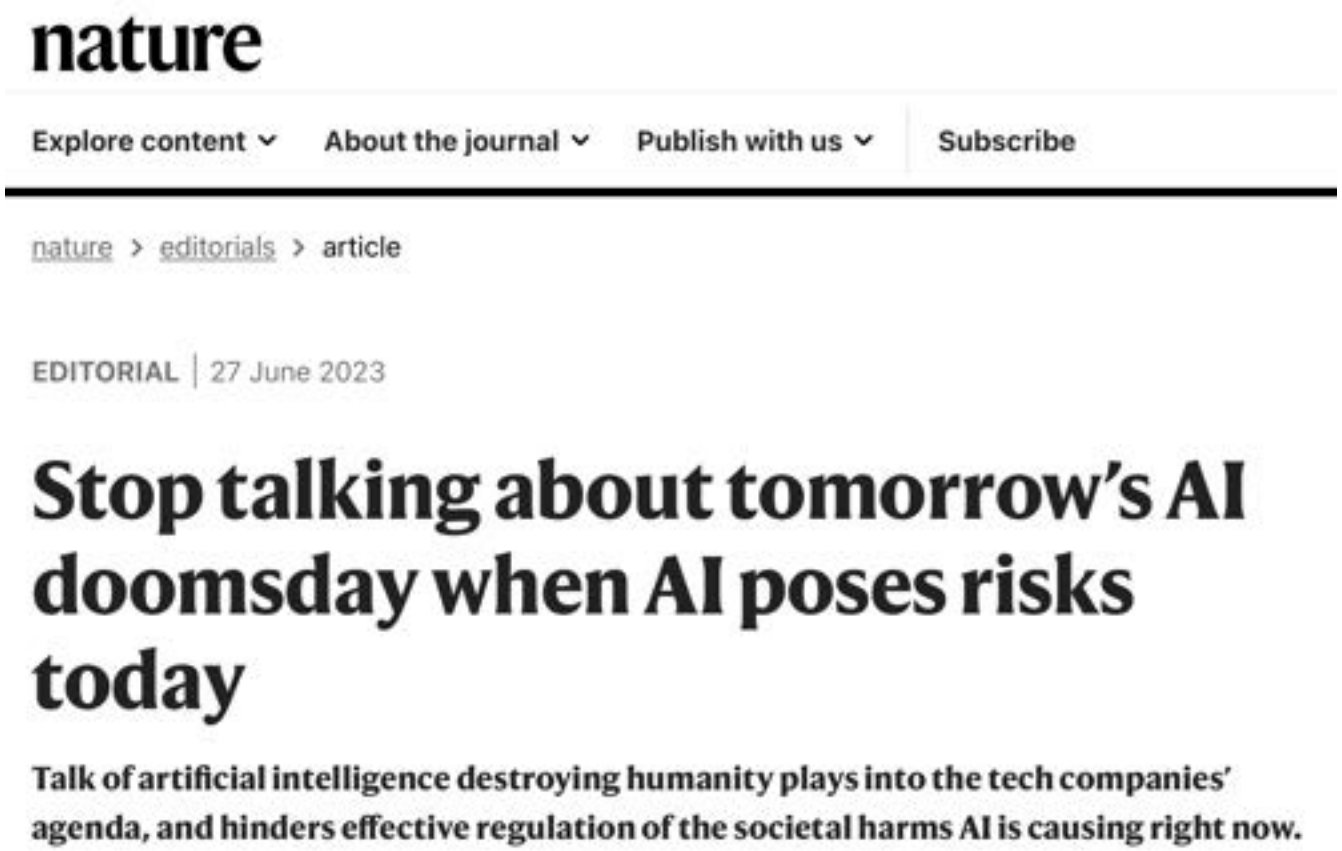
- →ネガティブな社会的インパクトが増強される

高度なAIシステムのリスク？

- 高度なAIシステムで人類存続リスクが懸念されている
- これは恐怖心とAIハイプを煽るものとして批判されている
- 「長期主義（longtermism）」と呼ばれるイデオロギーの中心的な考え方　→　今日のAIシステムの展開によって生じる実際の害を無視するものとも見られている
- “Criti-hype”（批判をすることにより、ハイプを増す）

DAIR, 2023; France24, 2023; Future of Life Institute, 2023; OpenAI, 2023; The Japan Times, 2023; Torres, 2023

『今日、AIがリスクをもたらしている中、明日のAI「終末の日」の話をやめよう』
人工知能による人類の絶滅に関する話は、テック企業の計略を支持し、すでにAIによって引き起こされている社会的危害に対する効果的な規制の妨げになっている



nature

Explore content ∨ About the journal ∨ Publish with us ∨ Subscribe

nature > editorials > article

EDITORIAL | 27 June 2023

Stop talking about tomorrow's AI doomsday when AI poses risks today

Talk of artificial intelligence destroying humanity plays into the tech companies' agenda, and hinders effective regulation of the societal harms AI is causing right now.

出典：https://www.nature.com/articles/d41586-023-02094-7

論点④　高度なAIシステムで生じるリスク（2）

	推測的なリスク （Speculative risks）	現実のリスク （Real risks）
誤報	悪意ある偽情報	不正確なツールへの過度の依存
労働への影響	テキスト生成AIがすべての仕事に取って代わる	中央集権化した権力、労働者の搾取
安全性	長期の人類存続リスク	直近のセキュリティリスク

Kapoor and Narayanan, 2023

論点⑤　自然環境への影響

論点⑤ 自然環境への影響（1）

Bender et al.（2021）：テキスト生成AIの使用の範囲とその規模を考慮すると、
優先順位の最も高い検討事項は**環境負荷**

テキスト生成AIの環境コストと経済的コストで、活用からの利益を得る可能性が最も低く、
害を受ける可能性が最も高い、**疎外されたコミュニティを二重に罰する**

- 二酸化炭素排出や淡水の使用量に関する懸念
- ハードウェアやインフラに必要とされている希土類金属や砂等の希少性

「ネットワークルーターからバッテリー、データセンターに至るまで、AI システムのネットワーク内の各要素は、
地球内部で形成されるまでに数十億年を要した元素を使用して構築されます。
ディープタイムの観点から、私たちは現代の技術時間のほんの一瞬のために地球の地質学的歴史を抽出しています。」

Beiser, 2018; Brevini, 2021; Crawford, 2021; Bender et al., 2021; Git Hub, 2023; Li et al., 2023; Luccioni et al., 2023; Twitter, 2023; Wired, 2023; Heaven, 2024; Stokel-Walker, 2024

論点⑤ 自然環境への影響（2）

研究がいまだに不十分

Benderらからの最初の呼びかけから３年経って：

「これまでの議論は、環境に対する現在及び将来のLLMの潜在的な影響を見逃している」

- 「画像を含むタスクは … エネルギーと二酸化炭素を大量に消費」するため、特に問題視されている

人間の健康と環境へのリスクが懸念されている

Pitron, 2020; Crawford, 2021; Rillig et al.,2023; Luccioni et al., 2023; Li et al., 2023; The Economist, 2023; Stokel-Walker, 2024

『生成AIの環境コストは高騰している―そして、ほとんど秘密』

nature

Explore content | About the journal | Publish with us | Subscribe

nature > world view > article

WORLD VIEW | 20 February 2024

Generative AI's environmental costs are soaring — and mostly secret

First-of-its-kind US bill would address the environmental costs of the technology, but there's a long way to go.

By Kate Crawford

出典：https://www.nature.com/articles/d41586-024-00995-9

おわりに

今後に向けて

- AIの開発と実装といったライフサイクル全体で生じるELSIリスクを（事前に）考える必要性がある　＝　“ELSI by design”

論点→ フェーズ↓	データ	知的財産権	バイアス	プライバシー	情報環境	文化	自然環境
データ収集							
データ							
フィルタリング							
モデルデザイン							
モデルデザイン							
学習							
強化							
実装							
インプット							
プラットフォーム							
アウトプット							

ライフサイクルの各フェーズで生じ得るリスクを検討するためのツール（生成AI用・案）

生成AIの開発と実装でますます多様な種類のリスクが生じ続けている

→ これらのリスクをゼロにすることは難しく、どの程度のリスクレベルならば、その便益と引き換えに受容できるのかについての検討が必要である

Katirai et al., 2023

謝辞

Many thanks to my collaborators on the projects reported in these slides, with special thanks to Professor Atsuo Kishimoto, Dr. Yuta Nakashima, Dr. Kazuki Ide, and Dr. Noa Garcia.

参考文献①

- Abid, A., Farooqi, M. and Zou, J., 2021. Large language models associate Muslims with violence. Nature Machine Intelligence, 3(6), pp.461-463.
- Adams LC, Busch F, Truhn D, Makowski MR, Aerts HJ, Bressem KK. What Does DALL-E 2 Know About Radiology? J Med Internet Res. 2023;25:e43110.
- Agarwal A, Farid H, Gu Y, He M, Koki Nagano K, and Li H. Protecting world leaders against deep fakes. In CVPR Workshops, volume 1, page 38, 2019.
- Beiser, V. The world in a grain: The story of sand and how it transformed civilization. Penguin Publishing Group. 2018
- Bendel O. Image synthesis from an ethical perspective. AI & SOCIETY. 2023.
- Bender EM, Gebru T, McMillan-Major A, Shmitchell S. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜. In: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency. Virtual Event Canada: ACM; 2021. p. 610–23. https://dl.acm.org/doi/10.1145/3442188.3445922
- Bianchi, F, Kalluri, P, Durmus, E, Ladhak, F, Cheng, M, Nozza, D, Hashimoto, T, Jurafsky, D Zou, J, and Caliskan, A. Easily accessible text-to-image generation amplifies demographic stereotypes at large scale. arXiv preprint arXiv:2211.0375
- Birhane A, Prabhu VU, Kahembwe E. Multimodal datasets: misogyny, pornography, and malignant stereotypes. arXiv preprint arXiv:2110.01963. 2021.
- Bird C, Ungless E, Kasirzadeh A. Typology of risks of generative text-to-image models. InProceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society.
- Brevini, B. Is AI good for the planet? John Wiley & Sons, 2021.
- Carlini N, Hayes J, Nasr M, Jagielski M, Sehwag V, Tramer F, Balle B, Ippolito D, and Wallace E. Extracting training data from diffusion models. arXiv preprint arXiv:2301.13188, 2023.
- Chen C, Fu J, Lyu L. A Pathway Towards Responsible AI Generated Content. ArXiv Prepr ArXiv230301325. 2023;
- Cheong C, Abedin E, Ferreira M, Reimann RW, Chalson, S, Robinson, P, Byrne, J, Ruppanner, L, Alfano, M, and Klein,C. Investigating gender and racial biases in DALL-E Mini images. PNAS, 2023 (to be published).
- Chesney R and Citron D. Deepfakes and the new disinformation war: The coming age of post-truth geopolitics. Foreign Aff., 98:147, 2019
- Crawford K, Paglen T. Excavating AI: The politics of images in machine learning training sets. AI Soc. 2021;36(4):1105–16.

参考文献②

- Crawford, K. The atlas of AI: Power, politics, and the planetary costs of artificial intelligence. Yale University Press, 2021.
- Edwards, B. Artist finds private medical record photos in popular AI training data set. Ars Technica, 2022. https://arstechnica.com/information-technology/2022/09/artist-finds-priva te-medical-record-photos-in-popular-ai-training-data-set/.
- Ferrara, E. (2023). Should ChatGPT be biased? Challenges and risks of bias in large language models. arXiv preprint arXiv:2304.03738.
- Fraser KC, Kiritchenko S, Nejadgholi I. Diversity is not a one-way street: Pilot study on ethical interventions for racial bias in text-to-image systems. ICCV Accept. 2023.
- Garcia N, Hirota Y, Wu Y, Nakashima Y. Uncurated image-text datasets: Shedding light on demographic bias. InProceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition 2023.
- Katirai A, Garcia N, Ide K, Nakashima Y, Kishimoto A. Situating the social issues of image generation models in the model life cycle: a sociotechnical approach. 2023.
- Ko H, Park G, Jeon H, J, J, Kim J, and Seo J. Large-scale text-to-image generation models for visual artists' creative works. In IUI, pages 919–933, 2023.
- Li P, Yang J, Islam MA, Ren S. Making AI Less "Thirsty": Uncovering and Addressing the Secret Water Footprint of AI Models [Internet]. arXiv; 2023 [cited 2023 Sep 6]. Available from: http://arxiv.org/abs/2304.03271
- Luccioni, AS, Akiki, C, Mitchell, M, and Jernite, Y. Stable bias: Analyzing societal representations in diffusion models. arXiv preprint arXiv:2303.11408, 2023.
- Luccioni AS, Jernite Y, Strubell E. Power Hungry Processing: Watts Driving the Cost of AI Deployment? [Internet]. arXiv; 2023 [cited 2023 Nov 29]. Available from: http://arxiv.org/abs/2311.16863
- Mökander J, Schuetthree-layered approach . arXiv; t J, Kirk HR, Floridi L. Auditing large language models: a 2023. http://arxiv.org/abs/2302.08500
- Munn L, Magee L, Arora V. Truth machines: synthesizing veracity in AI language models. AI & society. 2023.
- Nichol A. DALL-E 2 pre-training mitigations. OpenAI Blog. 2022.
- Nightingale SJ and Farid H. AI-synthesized faces are indistinguishable from real faces and more trustworthy. PNAS, 119(8):e2120481119, 2022.
- Offert, F and Phan, T. A sign that spells: DALL-E 2, in visual images and the racial politics of feature space. arXiv preprint arXiv:2211.06323, 2022.
- Oppenlaender J, Visuri A, Paananen V, Linder R, and Silvennoinen J. Text-to-image generation: Perceptions and realities. arXiv preprint arXiv:2303.13530, 2023.
- Pitron G. The Rare Metals War: the dark side of clean energy and digital technologies. Scribe Publications; 2020.
- Ricker J, Damm S, Holz T, and Fischer S. Towards the detection of diffusion model deepfakes. arXiv preprint arXiv:2210.14571, 2023.

参考文献③

- Rillig MC, Ågerstrand M, Bi M, Gould KA, Sauerland U. Risks and benefits of large language models for the environment. Environmental Science & Technology. 2023 Feb 23;57(9):3464-6.
- Spitale, G., Biller-Andorno, N. and Germani, F., 2023. AI model GPT-3 (dis) informs us better than humans. arXiv preprint arXiv:2301.11924.
- Srinivasan R and Uchino K. Biases in generative art: A causal look from the lens of art history. In FAccT, pages 41–51, 2021.
- Steed R and Caliskan A. Image representations learned with unsupervised pre-training contain human-like biases. In FAccT. ACM, 2021.
- Stokel-Walker, C. How AI ate the world: a brief history of artificial intelligence – and its long future. WHSmith, 2024.
- Struppek L, Hintersdorf D, and Kersting K. Rickrolling the artist: Injecting invisible backdoors into text-guided image generation models. arXiv preprint arXiv:2211.02408, 2022.
- Takeshita M, Rzepka R, Araki K. Speciesist language and nonhuman animal bias in English Masked Language Models. Information Processing & Management. 2022;59(5):103050.
- Ungless EL, Ross B, Lauscher A. Stereotypes and Smut: The (Mis) representation of Non-cisgender Identities by Text-to-Image Models. ArXiv Prepr ArXiv230517072. 2023;
- Verdoliva L. Media forensics and deepfakes: An overview. IEEE Journal of Selected Topics in Signal Processing, 14(5):910–932, 2020
- Wu, X., Duan, R. & Ni, J. Unveiling Security, Privacy, and Ethical Concerns of ChatGPT. Preprint at http://arxiv.org/abs/2307.14192 (2023).

***

- Burgess, 2023: https://www.wired.com/story/deepfake-porn-is-out-of-control/
- Burgess, 2024: https://www.wired.com/story/google-deepfake-porn-dmca-takedowns/
- Chowdhury, 2024: https://www.nature.com/articles/d41586-024-00995-9
- Christopher and Bansal, 2024: https://www.wired.com/story/indian-elections-ai-deepfakes/
- Crawford, 2024: https://www.nature.com/articles/d41586-024-00995-9
- DAIR, 2023: https://www.dair-institute.org/blog/letter-statement-March2023
- Feliba, 2023: https://www.japantimes.co.jp/news/2023/11/22/world/politics/ai-javier-milei-argentina-presidency/
- France24, 2023: https://www.france24.com/en/live-news/20230828-the-fight-over-a-dangerous-ideology-shaping-ai-debate
- Future of Life Institute, 2023: https://futureoflife.org/open-letter/pause-giant-ai-experiments/
- GitHub, 2023: https://github.com/Stability-AI/stablediffusion/blob/main/modelcard.md

参考文献④

- Guest, 2024: https://www.wired.com/story/mayor-of-london-sadiq-khan-ulez-conspiracy
- Heaven, 2024: https://www.technologyreview.com/2024/06/18/1093440/what-causes-ai-hallucinate-chatbots/
- Heikkila, 2024: https://www.technologyreview.com/2024/01/29/1087325/three-ways-we-can-fight-deepfake-porn-taylors-version/
- Helmore, 2024: https://www.theguardian.com/us-news/2024/apr/09/alexandria-ocasio-cortez-deepfake-porn
- Klein, 2023: https://www.theguardian.com/commentisfree/2023/may/08/ai-machines-hallucinating-naomi-klein
- Lamb et al., 2024: https://foreignpolicy.com/2024/02/14/prabowo-indonesia-election-democracy-jokowi/
- Lubin, 2024: https://cyber.harvard.edu/story/2024-03/what-do-about-junkification-internet
- Morrish, 2024: https://www.wired.com/story/dark-side-open-source-ai-image-generators/
- Kapoor and Narayanan, 2023: https://aisnakeoil.substack.com/p/a-misleading-open-letter-about-sci
- OpenAI, 2023: https://cdn.openai.com/papers/gpt-4-system-card.pdf
- The Economist: https://www.economist.com › business › 2023/10/05
- The Japan Times, 2023: https://www.japantimes.co.jp/business/2023/08/30/tech/ideological-fight-over-ai/
- The Verge, 2023: https://www.theverge.com/2023/12/20/24009418/generative-ai-image-laion-csam-google-stability-stanford
- Thiel, 2023: https://cyber.fsi.stanford.edu/news/investigation-finds-ai-image-generation-models-trained-child-abuse
- Torres, 2023: https://www.newstatesman.com/ideas/2023/08/longtermism-threat-humanity
- Twitter, 2023 (Sam Altman): https://twitter.com/sama/status/1599671496636780546?s=20&t=asOjLruBhE9QkebiA8pOgQ
- Wired, 2023: https://www.wired.com/story/the-generative-ai-search-race-has-a-dirty-secret/?