Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

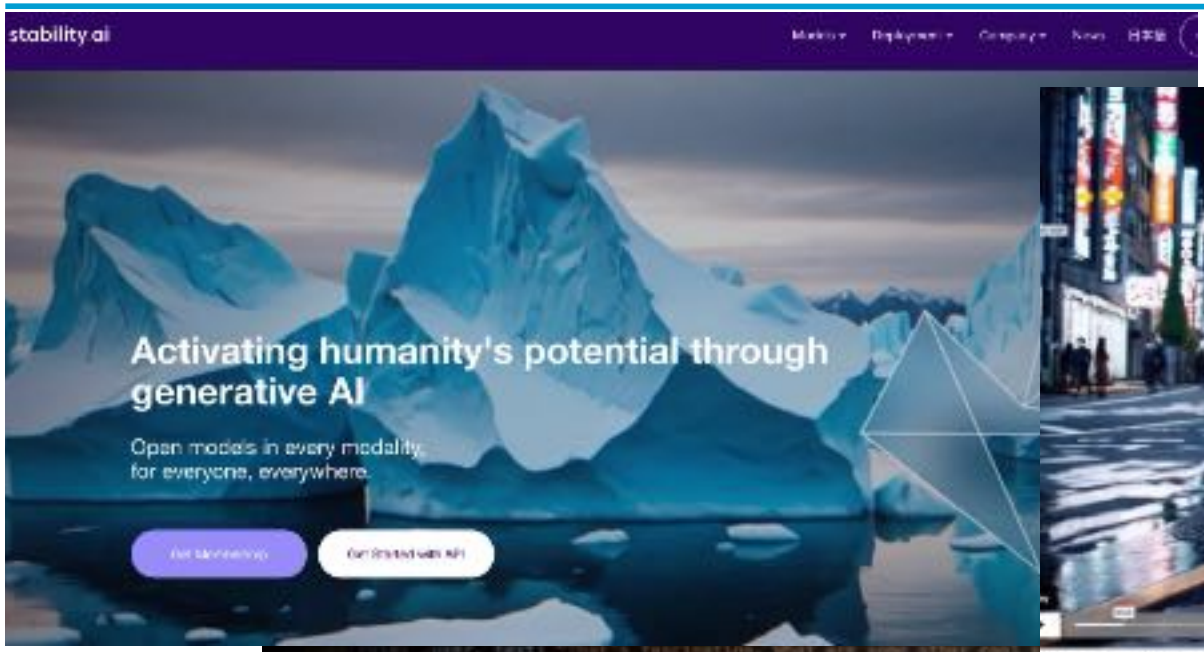
Prof. Dr. Judith Simon

Trustworthy Generative AI?

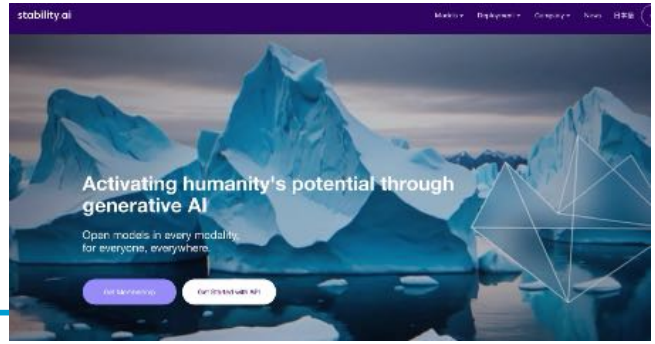
Challenges and Prospects

12.11.24 | Trilateral AI Conference 2024 | Tokyo | Japan

Generative AI – Pathways to Democratization, Transparency and Sustainability



https://openai.com
https://openai.com/sora
https://stability.ai



The Advent of Generative AI

- Extremely fast update & disruptive effects due to

1. Production of plausible texts and images of high quality, but without reference to truth

→ Wide range of applicability & high importance for information, communication, emotion

2. Simple interface & free access

→ immediate usability without technical prerequisites or skills

https://openai.com
https://stability.ai

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

Generative AI – Beyond Euphoria and Simple Solutions : (New) Challenges of Generative AI

1. Necessity and Limitations of De-Biasing
2. Possibilities and Limits of Explainable AI
3. Risk of Quadruple Deception
4. Power and Power Asymmetry
5. AI as Critical Infrastructure
6. Responsibility Diffusion and Control Deficits
7. Availability and Openness of AI Systems

Simon et al. (2024): Generative AI – Beyond Euphoria and Simple Solutions : (New) Challenges of Generative AI
https://www.leopoldina.org/publikationen/detailansicht/publication/generative-ki-2024/

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

Generative AI – Beyond Euphoria and Simple Solutions : (New) Challenges of Generative AI

1. Necessity and Limitations of De-Biasing
2. Possibilities and Limits of Explainable AI
3. Risk of Quadruple Deception
4. Power and Power Asymmetry
5. AI as Critical Infrastructure
6. Responsibility Diffusion and Control Deficits
7. Availability and Openness of AI Systems

Simon et al. (2024): Generative AI – Beyond Euphoria and Simple Solutions : (New) Challenges of Generative AI
https://www.leopoldina.org/publikationen/detailansicht/publication/generative-ki-2024/

Bias, Discrimination & Justice

nature

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

nature > news feature > article

NEWS FEATURE | 19 March 2024

AI image generators often give racist and sexist results: can they be fixed?

Researchers are tracing sources of racial and gender bias in images generated by artificial intelligence, and making efforts to fix them.

Press release chevron_right

Generative AI: UNESCO study reveals alarming evidence of regressive gender stereotypes

Ahead of the International Women's Day, a UNESCO study revealed worrying tendencies in Large Language models (LLM) to produce gender bias, as well as homophobia and racial stereotyping. Women were described as working in domestic roles far more often than men ¬– four times as often by one model – and were frequently associated with words like "home", "family" and "children", while male names were linked to "business", "executive", "salary", and "career".

https://www.unesco.org/en/articles/generative-ai-unesco-study-reveals-alarming-evidence-regressive-gender-stereotypes
https://www.nature.com/articles/d41586-024-00674-9

'We definitely messed up': why did Google AI tool make offensive historical images?

Experts say Gemini was not thoroughly tested, after image generator depicted variety of historical figures as people of colour

Google's Gemini AI illustrations of a 1943 German soldier. Photograph: Gemini AI/Google

https://www.theguardian.com/technology/2024/mar/08/we-definitely-messed-up-why-did-google-ai-tool-make-offensive-historical-images

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

Generative AI – Beyond Euphoria and Simple Solutions : (New) Challenges of Generative AI

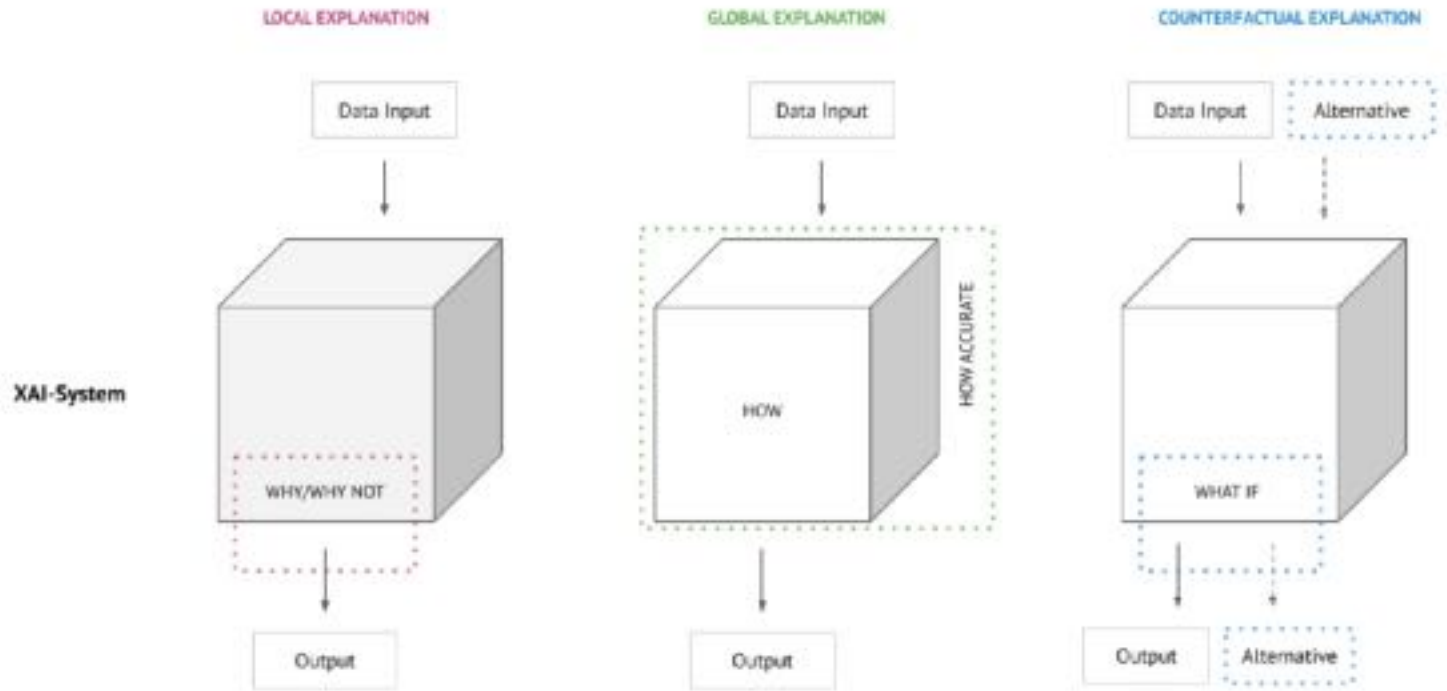
1. Necessity and Limitations of De-Biasing
2. Possibilities and Limits of Explainable AI
3. Risk of Quadruple Deception
4. Power and Power Asymmetry
5. AI as Critical Infrastructure
6. Responsibility Diffusion and Control Deficits
7. Availability and Openness of AI Systems

Simon et al. (2024): Generative AI – Beyond Euphoria and Simple Solutions : (New) Challenges of Generative AI
https://www.leopoldina.org/publikationen/detailansicht/publication/generative-ki-2024/

Picturing the original image (left), saliency map using a method called Grad-CAM (middle), and another using Guided Backpropagation (right). The picture above is the canonical example for "class-discrimination". The above saliency maps are taken from https://github.com/kazuto1011/grad-cam-pytorch.

https://towardsdatascience.com/what-explainable-ai-fails-to-explain-and-how-we-fix-that-1e35e37bee07

„*Who* needs to understand what in a given scenario? *What can* be explained about the system in use? *What should* explanations look like in order to be *meaningful* to affected users?“



Asghari, et al. (2021). "What to explain when explaining is difficult? An interdisciplinary primer on XAI and meaningful information in automated decision-making." https://doi.org/10.5281/zenodo.6375784.

Generative AI – Beyond Euphoria and Simple Solutions : (New) Challenges of Generative AI

1. Necessity and Limitations of De-Biasing
2. Possibilities and Limits of Explainable AI
3. Risk of Quadruple Deception
4. Power and Power Asymmetry
5. AI as Critical Infrastructure
6. Responsibility Diffusion and Control Deficits
7. Availability and Openness of AI Systems

Simon et al. (2024): Generative AI – Beyond Euphoria and Simple Solutions : (New) Challenges of Generative AI
https://www.leopoldina.org/publikationen/detailansicht/publication/generative-ki-2024/

Deception I: AI or Human?

https://de.wikipedia.org/wiki/ELIZA#/media/Datei:ELIZA_conversation.jpg
Open.ai

Deception II: AI Capabilities & AI Hype

The Google engineer who thinks the company's AI has come to life

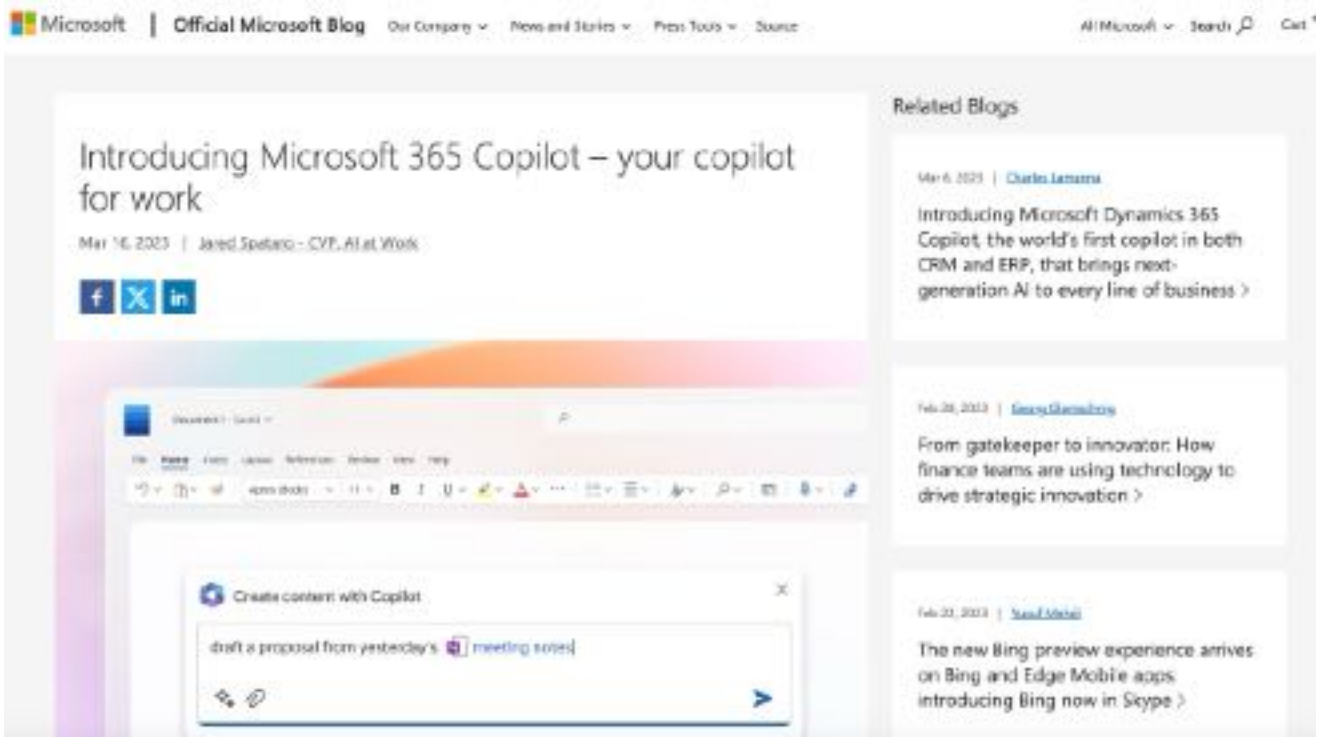
AI ethicists warned Google not to impersonate humans. Now one of Google's own thinks there's a ghost in the machine.

https://www.washingtonpost.com/technology/2022/06/11/google-ai-lamda-blake-lemoine/

Deception III: Deep Fakes

https://www.zdf.de/nachrichten/panorama/prominente/papst-daunenjacke-fake-ki-kuenstliche-intelligenz-100.html

Deception IV: AI Function Creep

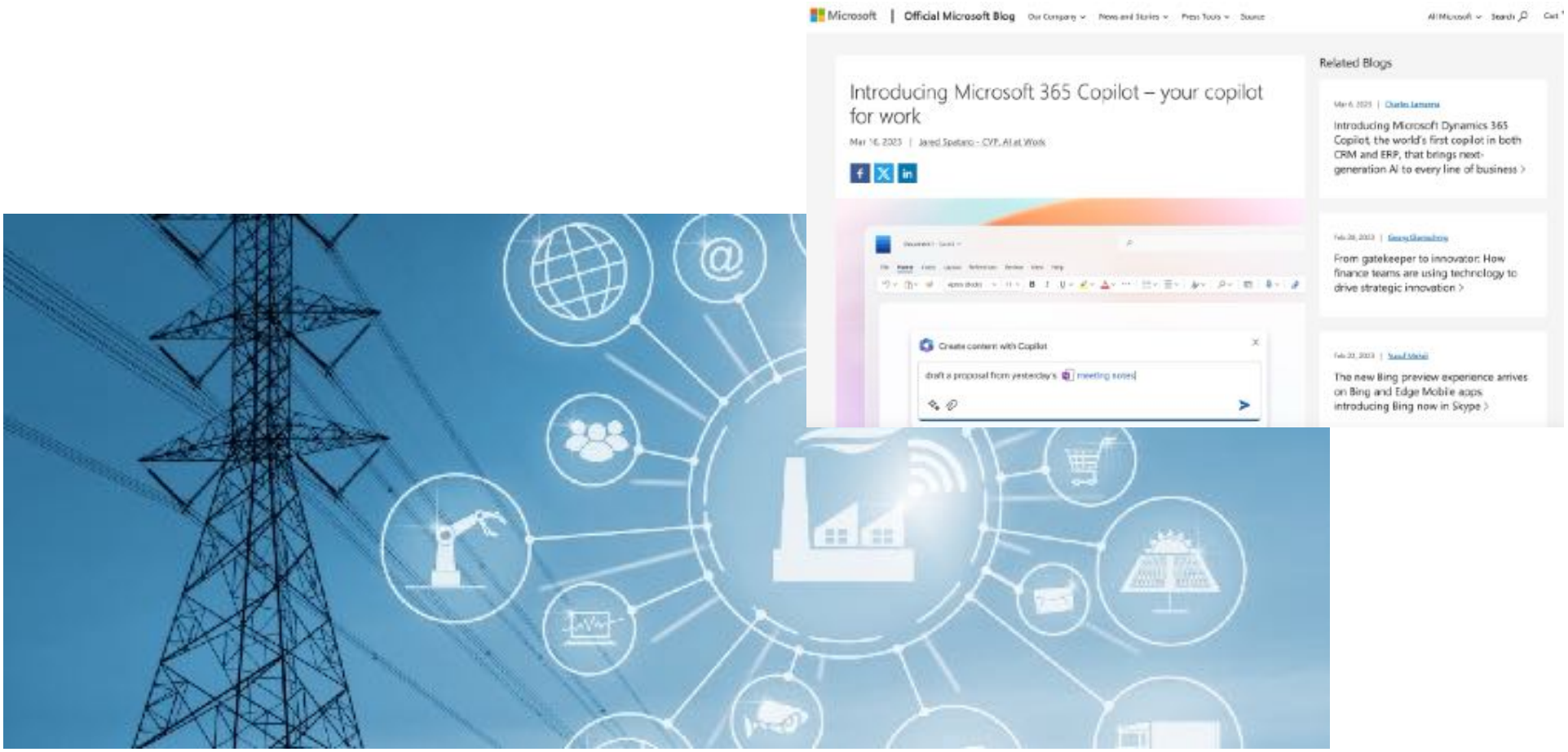


https://blogs.microsoft.com/blog/2023/03/16/introducing-microsoft-365-copilot-your-copilot-for-work/

Generative AI – Beyond Euphoria and Simple Solutions : (New) Challenges of Generative AI

1. Necessity and Limitations of De-Biasing
2. Possibilities and Limits of Explainable AI
3. Risk of Quadruple Deception
4. Power and Power Asymmetry
5. AI as Critical Infrastructure
6. Responsibility Diffusion and Control Deficits
7. Availability and Openness of AI Systems

Simon et al. (2024): Generative AI – Beyond Euphoria and Simple Solutions : (New) Challenges of Generative AI
https://www.leopoldina.org/publikationen/detailansicht/publication/generative-ki-2024/



https://blogs.microsoft.com/blog/2023/03/16/introducing-microsoft-365-copilot-your-copilot-for-work/
https://www.bsi.bund.de/EN/Themen/KRITIS-und-regulierte-Unternehmen/Kritische-Infrastrukturen/kritis_node.html

Conclusions: Trustworthy GenAI?

- Trustworthy Gen AI?
  - Specific challenges of GenAI + generic ethical challenges of AI
  - Effects of AI/GenAI vary for different stakeholders - special focus on those who are already marginalized
  - Essential to understand both need for & limitations of technical harm mitigation (de-biasing, X-AI, etc)
  - Ethics in AI is in the method and goes beyond
  - The devil is in the details: a closer look at technologies, but also at institutional/organizational framework conditions

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

Many thanks for your attention!

Prof. Dr. Judith Simon
Professor for Ethics in Information Technologies
Universität Hamburg
Email: judith.simon@uni-hamburg.de
Web: http://uhh.de/inf-eit