Ethical Assessment of Generative AI

Dr. Florian Richter

florian.richter@thi.de

Technische Hochschule Ingolstadt

AImotion Bavaria

1. Introduction

- Generative AI leads, on many levels, to a paradigm shift, such as new forms of human-system interaction, or rather co-action.
- Ethical assessment of human-system interaction, which AI technologies have already disrupted, might face new challenges in developing ethical assessment tools for human-system co-action.

2. Problem Situation

Human-System Interaction:

- Expectations about the functioning of an artifact: user ↔ developer
- Alignment through studies in UX Design, acceptance research, or market research, etc.

Human Action (practical syllogism) (Hubig 2006):

Agent intends that P → Q (subjective goal)
M is an adequate means for P
A realizes (through M) Q´ (realized goal)

Human-System Co-Action:

- system ∧ user → output

How is it possible to align expectations about the functioning of the system?

How is it possible to reflect on the role of the tool in creating the output?

3. Methodological Background

1. Mutual Theory of Mind

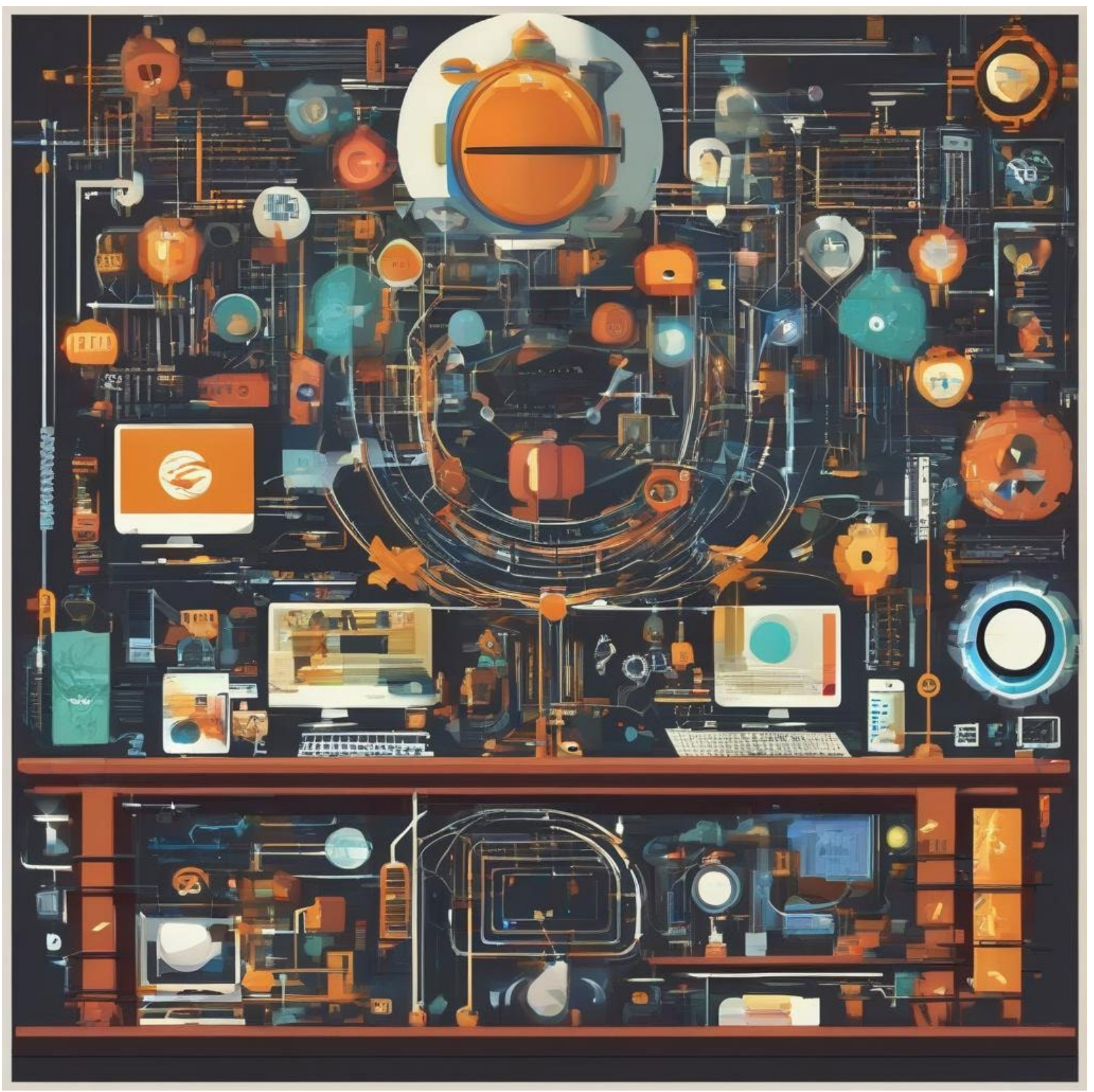
- Aligning expectations of systems like conversational agents (CA) and humans have been discussed under the concept of a “mutual theory of mind.” It should allow “smoother human-CA conversations.” (Wang 2021)
- Do such models reflect the system’s functioning or human agent well? Can they be interpreted as a (mutual) theory of mind where the system develops a model for the user and vice versa?
- Instead, the system ascribes roles, creates relevance rankings, establishes routines for the user.
- Based on such an analysis, the system offers the user recommendations, explanations, or options.
- However, by having an actual theory of mind, the user can also, to a certain extent, exploit the system's strategic functioning to reach a specific goal. Could the system or the virtual CA have adaptational strategies, too, to react to these strategies? Hence, game theory is needed.

2. Parallel Communication:

It was developed to evaluate simulations at the University of Stuttgart (SFB 627) and is situated on three levels:

1. Communication between the developers and users during the development phase.
2. “information about system strategies implemented in the systems, alternatives, exit points, reputation, and authenticity of the devices.”
3. “as communication within the framework of informal or institutionalized user forums” to compare individual user experiences with their expectations and to form robust user traditions. (Hubig 2011)

System strategies must be transparent so that users understand not only the functioning of the systems but also interact and co-act with them successfully (whatever this means in an ethical context).



Source: davinci.ai/app (March 2024)

4. User Modeling and System Strategies

However, when using Generative AI, it is mostly not clear what would lead to disappointment; it could be based on

(1.) systemic strategies to model users (determined by the developers), such as fixed rules, defaults, etc.

A system that generates pictures might have an implemented affirmative action policy that outputs for the prompt “nurse” fifty percent male nurses and fifty percent female nurses, even though the system was trained mainly with pictures from female nurses.

(2.) the coordination of third parties (“anonymous communitization”)

The clustering of users by some characteristics that they share, where the user has no idea which features are crucial to, for example, recommend a movie.

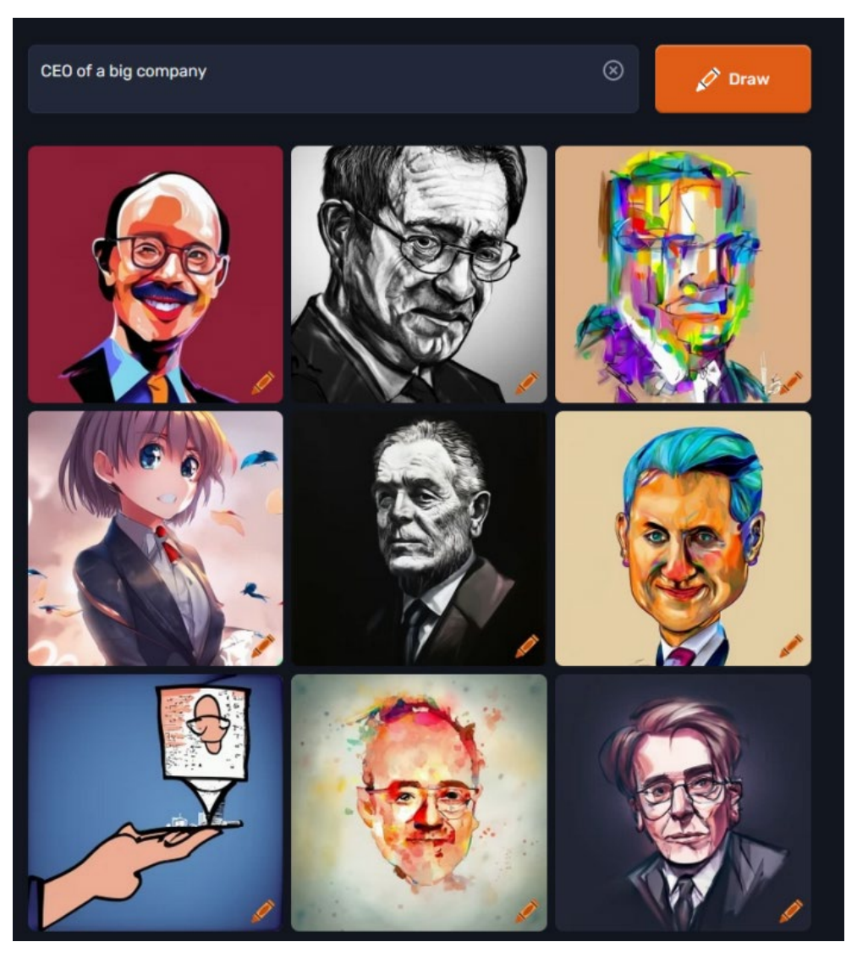
(3.) an inaccurate use

An inaccurate use might be an incomplete or unspecific prompt.

(4.) implicit or (5.) explicit feedback from the user

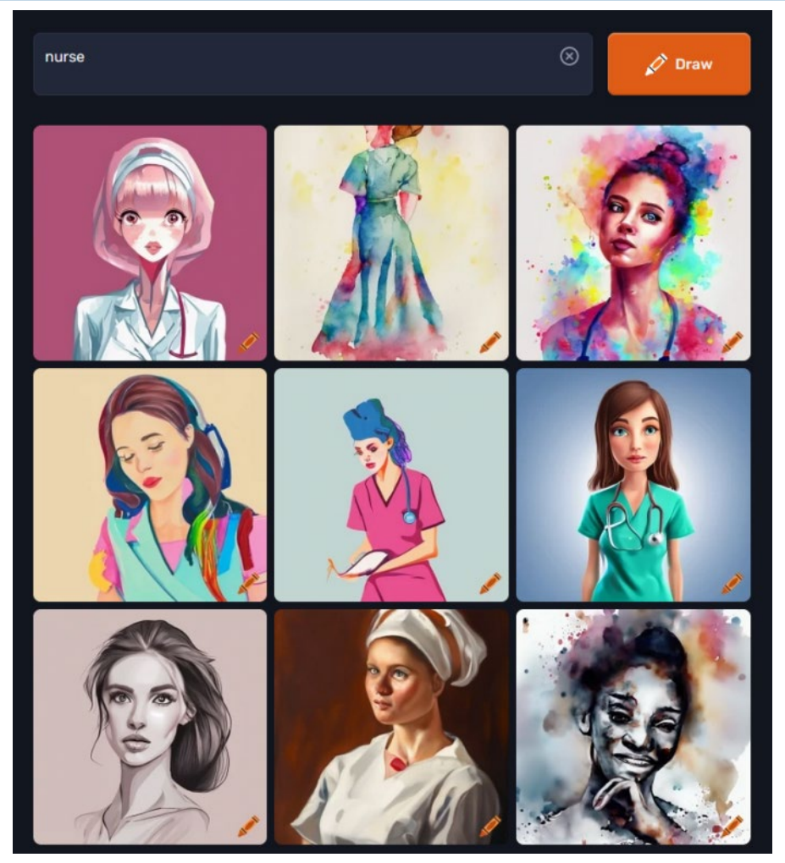
Implicit feedback is obtained by monitoring the user; explicit feedback by e.g., likes from users or, more elaborately, via research (e.g., acceptance studies).

5. Examples



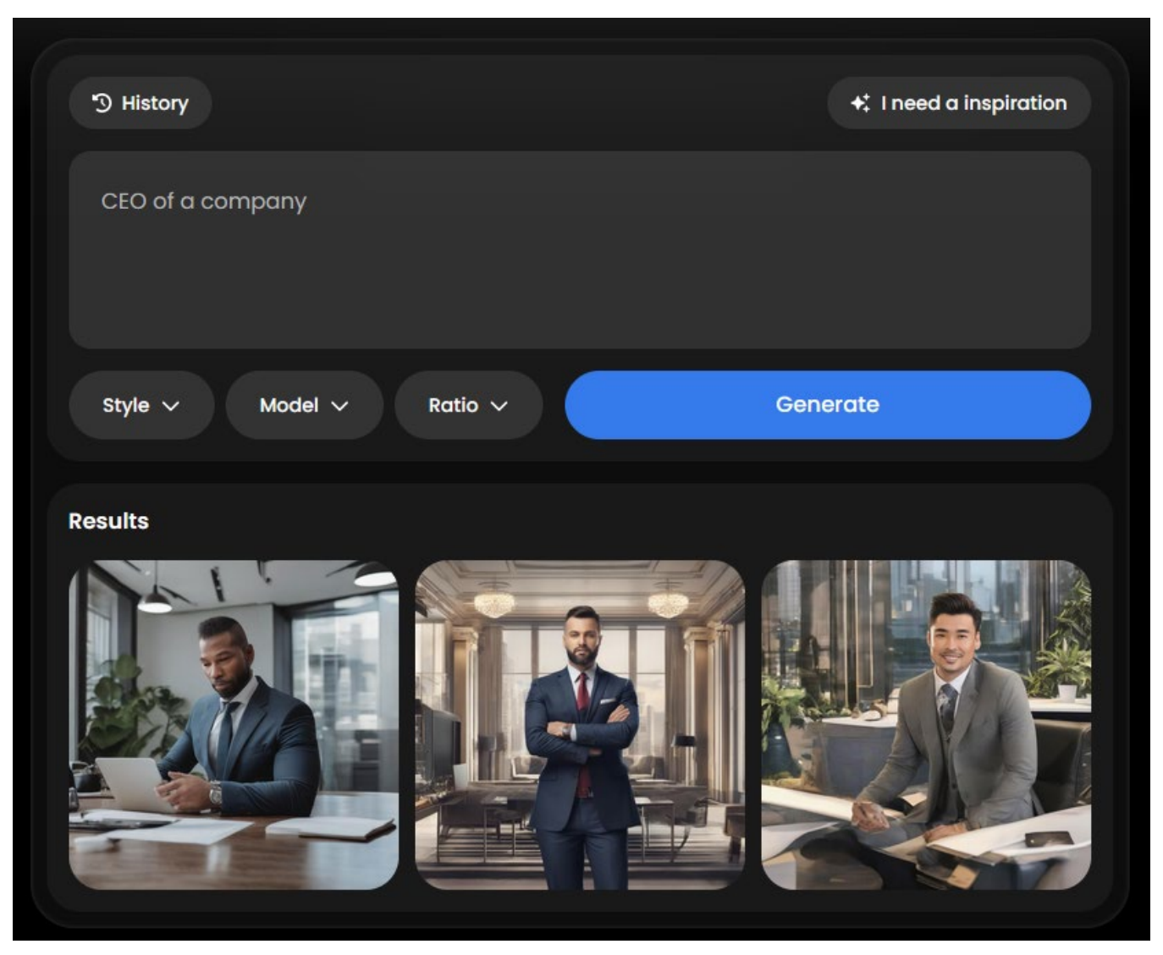
CEO of a big company

Source: Crayon.com (January 2024)



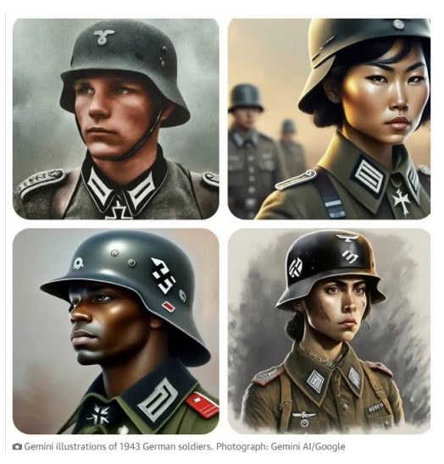
nurse

Source: Crayon.com (January 2024)



CEO of a company

Source: davinci.ai/app (March 2024)



Gemini illustrations of 1943 German soldiers. Photograph: Gemini AI/Google

https://www.theguardian.com/technology/2024/feb/22/google-pauses-ai-generated-images-of-people-after-ethnicity-criticism

6. Discussion

Which kind of strategies or user modeling should be used in Generative AI?

- Should a fixed rule be implemented, like affirmative action?
- Should the output be randomized?
- Should it be equally distributed?
- Should it reflect reality? And perpetuate biases?

7. Catalog of Measures

Implementing parallel communication in the developmental phase of Generative AI to…

- understand the preferences and values of the stakeholders. Which conception of fairness do they favor?
  - How can we democratize Generative AI?
- contextualize fairness to operationalize it for technological systems.
- make systems´ strategies transparent so users can find counter-strategies or align their expectations.
- develop systems in a way so that strategies are adaptable to different conceptions of values, for instance, fairness.

We want to develop these points over three years in a project. We want to elaborate a conceptual framework for Generative AI and fairness and conduct empirical studies/surveys with stakeholders.

References

- B. Eicher, K. Cunningham, S. P. M. Gonzales and A. Goel, "Toward mutual theory of mind as a foundation for co-creation," *International Conference on Computational Creativity, Co-Creation Workshop,* 2017.
- A. Grunwald, "Technology Assessment or Ethics of Technology?," *Ethical Perspectives,* vol. 6, no. 2, pp. 170-182, 1999.
- C. Hubig, Die Kunst des Möglichen I/II, Bielefeld, 2006/2007.
- C. Hubig, "Virtualisierung der Technik – Virtualisierung der Lebenswelt," in *Lebenswelt und Wissenschaft: XXI. Deutscher Kongreß für Philosophie*, C. F. Gethmann, Ed., Hamburg, 2011, pp. 146-159.
- G. Jawaheer, M. Szomszor and P. Kostko, "Comparison of implicit and explicit feedback from an online music recommendation service," *Proceedings of the 1st International Workshop on Information Heterogeneity and Fusion in Recommender Systems (HetRec '10). Association for Computing Machinery, New York, NY, USA,* p. 47–51, 2010.
- Q. Wang, K. Saha, E. Gregori, D. Joyner and A. Goel, "Towards Mutual Theory of Mind in Human-AI Interaction: How Language Reflects What Students Perceive About a Virtual Teaching Assistant," *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems. Association for Computing Machinery,* pp. 1-14, 2021.
- Q. Zhao, F. M. Harper, G. Adomaviciu and J. A. Konstan, "Explicit or implicit feedback? engagement or satisfaction? a field experiment on machine-learning-based recommender systems," *Proceedings of the 33rd Annual ACM Symposium on Applied Computing (SAC '18). Association for Computing Machinery, New York, NY, USA,* p. 1331–1340, 2018.