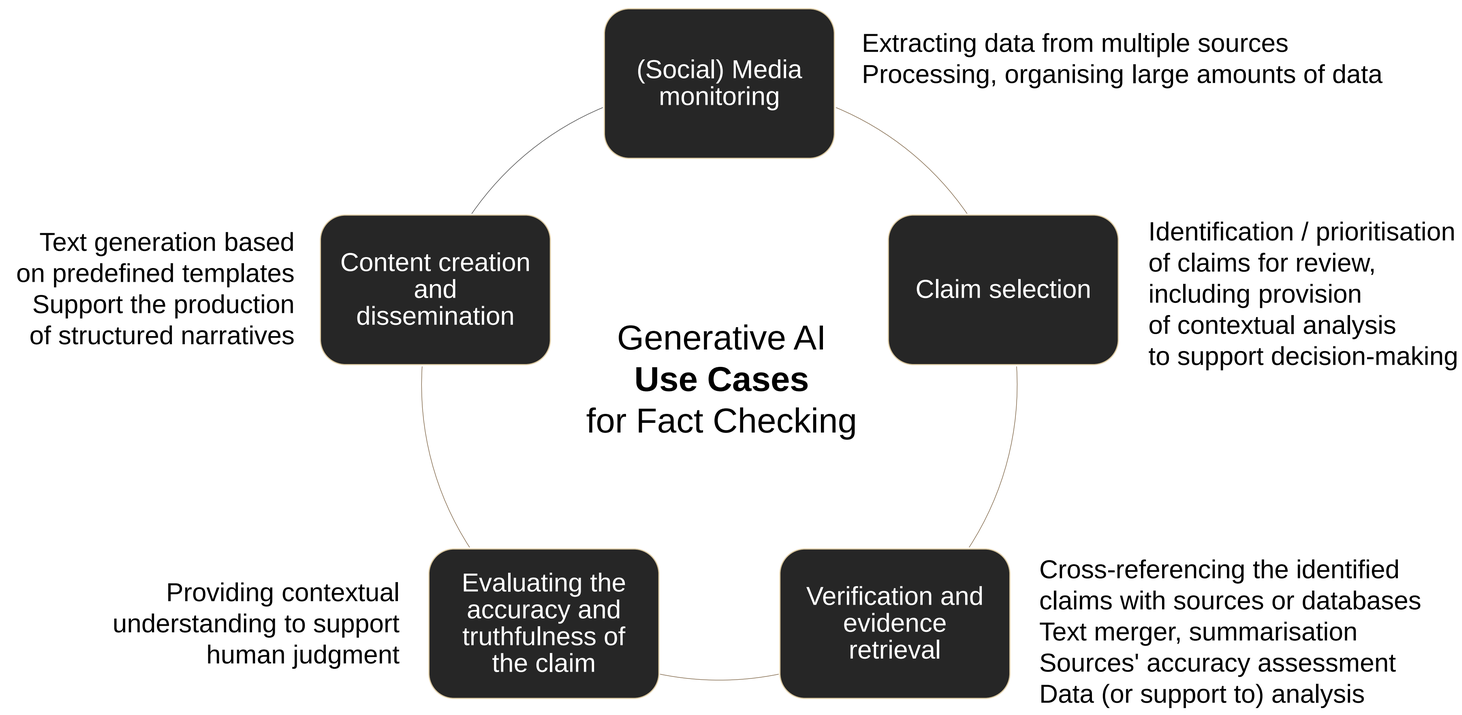
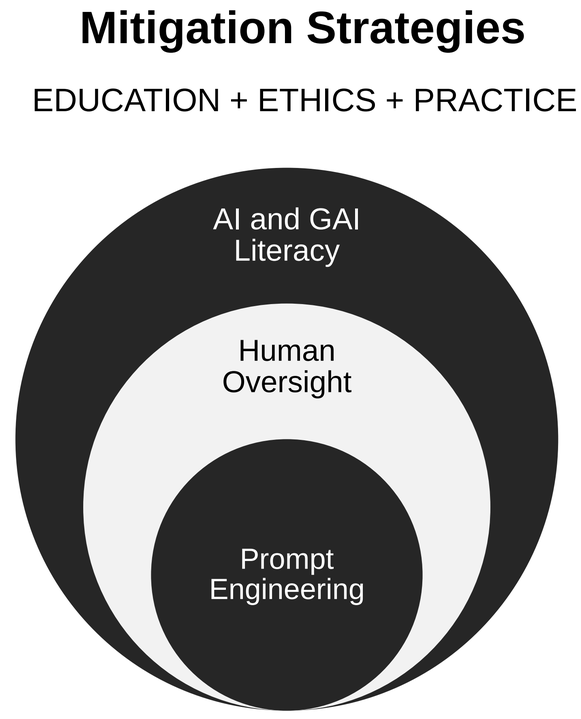
Striking the Balance in Using Generative AI for Fact-Checking

Presented by: Laurence Dierickx, Research Fellow DDCxTrygFonden
Co-authors: Arjen van Dalen (University of Southern Denmark), Carl-Gustav Lindén, and Andreas L. Opdahl (University of Bergen, Nordic Observatory for Digital Media and Information Disorder, NORDIS)

The launch of ChatGPT in November 2022 sparked a reflection on the usefulness of generative AI (GAI) technology in supporting fact-checking workflows and practices. At the same time, critics have questioned the fairness and reliability of the data collection and training on which GAI systems are based. The well-known phenomenon of artificial hallucinations adds an extra layer to these concerns, as does the fear of a proliferation of machine-generated content to create and spread mis- or disinformation. Given these ethical challenges and the inherent limitations of a GAI system, how can the risks be mitigated to foster responsible use among fact-checkers?





Identification of the potential of GAI in support of the fact-finding process.

Background

GAI systems operate by learning patterns and information from large datasets, which presents complexities in ensuring ethical use and reliable results. Opacity and bias in data sources, including copyrighted data, risks of spreading mis- or dis-information, and artificial hallucinations: the challenges are many.

At the same time, GAI systems have great potential to support fact-checking, highlighting the need to adopt risk mitigation strategies.

Methods

The narrative literature review (NLR) provides a comprehensive overview of existing knowledge in an emerging field from an interdisciplinary perspective. As such, it is considered a qualitative research tool that allows flexibility in exploring different methodologies.

It enabled the identification of three complementary mitigation strategies.

AI and GAI Literacy

Defined as the understanding that facilitates AI and GAI's recognition, management and ethical application, literacy encompasses both practical skills and ethical considerations.

Initial or ongoing training programmes can help professionals understand these challenges, equip them with the appropriate skills, develop a critical mindset to prevent further risks and avoid being taken in by the persuasive tone of LLMs such as ChatGPT.

Human Oversight

This principle, which is prioritised in almost all ethical guidelines for the use of AI and GAI in the news media sector, underlines the ongoing responsibility of humans to mitigate potential adverse effects in the absence of algorithmic accountability.

For fact-checkers, it can also be seen as a means of maintaining transparency to counterbalance the system's opacity.

Prompt Engineering

Research has shown that well-crafted prompts can increase explainability and reduce the generation of fabricated content (artificial hallucinations).

Prompting techniques facilitate user interaction and problem-solving, for example, by providing context in the prompt, and are a promising way to improve the accuracy and reliability of results. They can also be used to generate new prompts, exploiting the potential for self-adaptation.

Perspectives

Exploring the interplay between education, ethics and practice and how they complement each other provides valuable insights into optimising fact-checking workflows and routines.

This research paves the way for future work on integrated frameworks that aim to achieve acceptable levels of transparency and reliability.