Dealing with biases and hallucinations: The ethical uses of (G)AI tools in the European news media sector

Laurence Dierickx – Carl-Gustav Lindén
University of Bergen
Nordic Observatory for Digital Media and Information Disorder (NORDIS)

ECREA 2024
September 25, 2024 - Ljubljana

LLMs: the game changers for ethics?

Operational complexity: LLMs learn patterns from large datasets, raising challenges in ensuring ethical use and reliable result (stochastic parrots).

- Data quality of training data (UGC, biased data)
- Risks of plagiarism (copyrighted data)
- Failures in logical deduction (A = B is not necessarily B = A)
- Artificial hallucinations (training data and processes)
- Generation is not verification (designed for providing an answer)
- Risk of being fooled by a convincing tone (anthropomorphism)
- Socio-professional risks (replacement of human work, impact on critical thinking and creativity)

NORDIS

UNIVERSITAS BERGENSIS

How are media and professional organisations in Europe developing and implementing guidelines to address biases and hallucinations in GAI tools?

What are their common features and key principles to ensure quality and ethics of information?

Focus on European media ethics

- Self-regulatory bodies
- Ethical codes agree on accuracy, fairness/balance, independence, privacy, protection of sources
- High level of professionalism
- Impact of news media public sphere
- Government support (public service broadcasters)

Method

- Search engines (Google, Ecosia)
  (AI OR ”artificial intelligence” OR ChatGPT) AND ethic AND journalism*
  ethics AND (guideline OR code OR recommendation*) AND (journal* OR media OR news)*
  (Name of the news media) AND (AI OR ”artificial intelligence” OR ChatGPT) AND (guideline OR recommendation* OR principle*)*
- (Human) Monitoring on Twitter/X
- Limits: publicly available texts (no internal documents)
- 51 texts from 11 countries between October 2019 and April 2024 (Belgium, Denmark, Finland, France, Germany, Norway, Spain, Sweden, Switzerland, The Netherlands, United Kingdom), include texts of US news media in Europe (AP, The Wired, Thomson Reuters)
- Non-English text translated with DeepL (human supervision)

NORDIS

UNIVERSITAS BERGENSIS

Ethical guidelines before ChatGPT

Public broadcasters

- BBC - Guidelines for ML developers (BBC, 2019)
- Bayerischer Rundfunk (Germany, 2020)
- Schweizer Radio und Fernsehen (Switzerland; 2021)
- Press agencies (AP, Thomson Reuters, AFP on data)

Press councils (recommendations)

- Council for Mass Media in Finland (2019)
- Catalan Press Council (2021)

“Almost all press councils in Europe have so far ignored data-driven phenomena like algorithms or news automation.”

Porlezza, C. (2024). The datafication of digital journalism: A history of everlasting challenges between ethical issues and regulation. *Journalism*, *25*(5), 1167-1185.

Key principles

- **Alignment with public service values** (BBC)
- **Responsible engineering and interdisciplinary collaboration**, emphasising transparency and the development of a 'data culture'. Collaboration with start-ups and universities to address bias and filter bubbles in personalised content (Bayerischer Rundfunk, BR)
- **Transparency and human responsibility** to ensure trust and credibility (Schweizer Radio und Fernsehen, SRF + 2 press councils)
- **Thorough verification and transparency**, with clearly identified data sources (Associated Press)
- **Fairness and a human-centred approach** in the design, development and deployment of AI products and services. (Thomson Reuters)

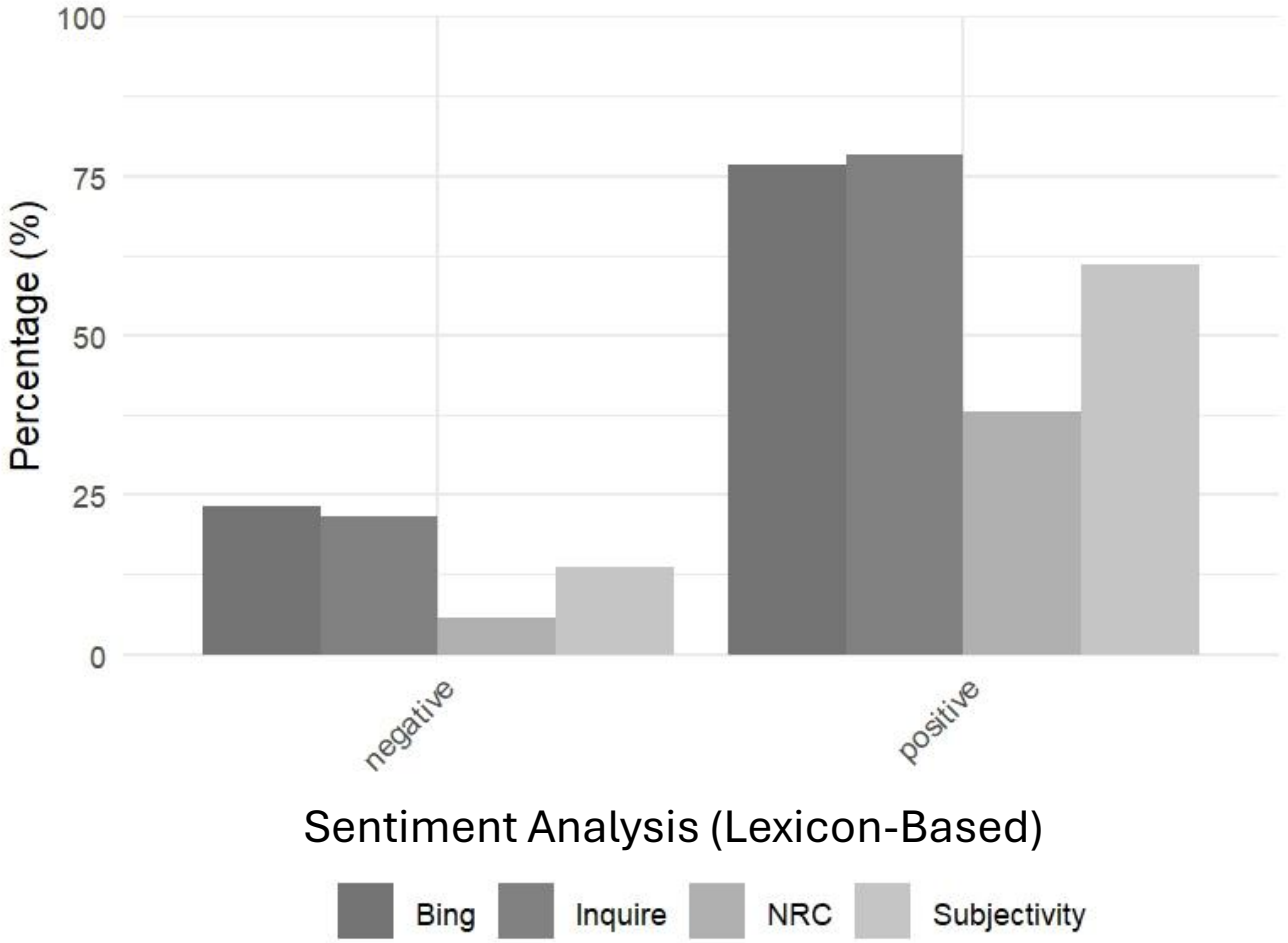
Ethical guidelines after ChatGPT

- News media (17)
- Public broadcasters (7)
- Press agencies (5)
- Press councils (4)
- Press groups (4)
- Professional organisations (4)
- “Guidelines”, relating to recommendations and practical advice (19)
- “Principles”, stating fundamental principles and including texts presented as “Charter” in French (14)
- “Positions”, focusing on commitment and perspective (7)
- Ethical codes (updated, 2)

General observations

- News media organisations are using AI, both generative and non-generative, to assist with various tasks from information gathering to distribution
- These organisations recognise the risks of AI systems, including bias (AJP, ANP, *Financial Times*, Mediahuis, RTS, Yle), potential errors and untrue content (ANP, STT).
- AI systems may rely on copyrighted data (one-third of the texts).
- There is an awareness of the paradox of AI technologies, which can be used to both inform and disinform (*Dagens Næringsliv*, *De Volksrant*, *Der Spiegel, Financial Times*, SVT).

NLP analysis



Related trigrams

AI related risks – case negatively affect – harm non-discrimination – intellectual property right – protect people personal – exercise great caution – machine learning challenges – protect source protection

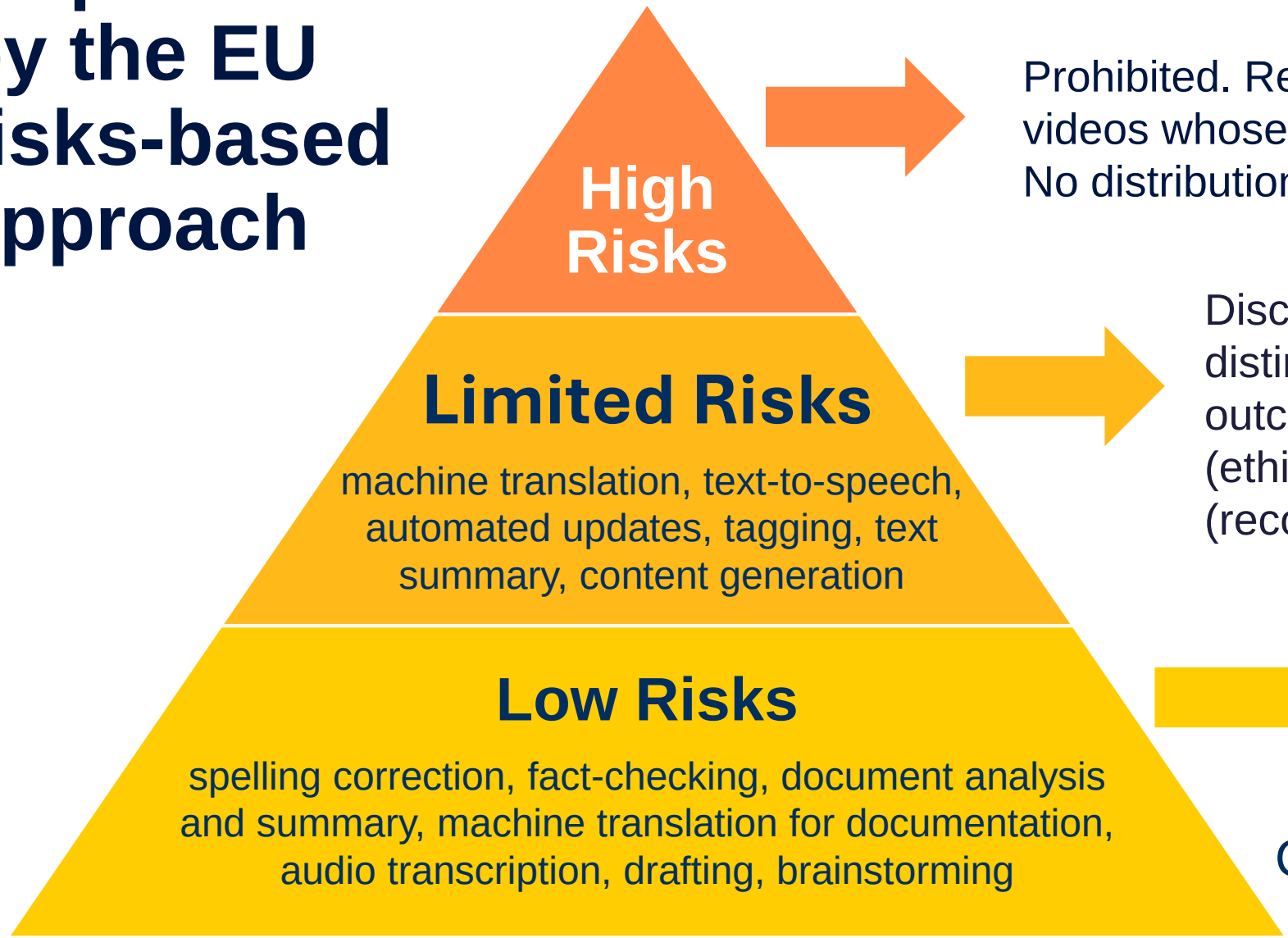
Acknowledging limitations

- ‘AI models can produce entirely false images and articles. They also replicate the existing societal perspectives, including historic biases’ (*Financial Times*)
- ‘They too often contain errors (“hallucinations”) or biases (“bias”), and it is usually unknown what data the systems were trained with’ (*De Volkskrant*)
- ‘With computer-generated content, it is complex to guarantee the reliability of facts presented as true’ (ANP)
- They also can generate ‘false leads and boring ideas’ (*Wired*)
- ‘The sources used by AI are often obscure, making it problematic to use in editorial work’ (STT)
- Material created using generative AI raises significant issues around bias, ownership, plagiarism and intellectual property rights (*The Guardian*)

NORDIS

UNIVERSITAS BERGENSIS

Inspired by the EU risks-based approach



Prohibited. Realistic images, sounds or videos whose realism is likely to mislead. No distribution without human approval.

Disclosure for all type of content, distinguishing machine outcomes, human oversight (ethics), human editorial control (recommenders)

No impact on information quality, no disclosure

CDJM, France

NORDIS

Experimenting with caution

The ethical perspective is part of a risk mitigation strategy that promotes responsible practices.

- Other strategies include testing and approval mechanisms to prevent AI "hallucinations" from being published (*Der Spiegel*, Germany).
- Due to potential errors, STT (Finland) avoids using AI for data exploration.
- Yle (Finland) emphasises ongoing risk assessment and vigilance to monitor and correct biases.
- General acknowledgement of use under human supervision and should be a human responsibility, and its use must be transparent to the audience.

Reflecting journalism ethics

- **Accuracy** (critically assessing and verifying sources and facts)
- **Trustworthiness** (sources may be unreliable, harmful or inaccurate)
- **Respect for facts** (not manipulating or distorting them)
- **Fairness** (avoid bias and echo chambers, encourage diversity)
- **Respect for (data) privacy**
- **Human responsibility** and accountability

Current issues and gaps

- Ethics is a matter of practice (ethical dilemmas)
- Internal vs common rules
- Transparency has many drawbacks that are not considered
- Only Yle (Finland) considers the environmental impact of AI.
- Less focus on the possibility of private or internal data leaks (only SVT in Sweden and Ringier in Switzerland)
- Professional organisations blurred messages about political and trade union positions (Belgium, Germany)
- New actors in journalism, such as data scientists and computer scientists, are not considered
- Risks mitigation also involve data and AI literacy + training

NORDIS

UNIVERSITAS BERGENSIS

Thank you for your attention!

Contact: @ohmyshambles @Gusse

See also on the risk mitigation strategies in fact-checking

Dierickx, L., van Dalen, A., Opdahl A.L. and Lindén C.G. *Striking the Balance in Using LLMs for Fact-Checking: A Narrative Literature Review*, in Proceedings from 6th Symposium on Multidisciplinary International Symposium on Disinformation in Open Online Media (MISDOOM), Lecture Notes in Computer Science (Springer).