European Network for Academic Integrity

Research Integrity and GenAI: Key points of vigilance and recommendations

Tomáš Foltýnek
Masaryk University, Czechia
foltynek@fi.muni.cz

European Network for Academic Integrity

Current debate on ethical issues

related to the

use of Gen-AI in science

Ethical Issues related to LLMs

European Network for Academic Integrity

- Energy consumption & carbon footprint
- Digital divide & environmental racism
- Societal changes: Labour market
- Impact on individuals: Overreliance & dependency on tech
- Biased training data → Biased outputs, stereotype reinforcement
- Hallucinations, confabulations → Inaccuracies in scientific outputs

- Bender, Emily M, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. “On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? .” In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–23. FAccT ’21. New York, NY, USA: Association for Computing Machinery, 2021. https://doi.org/10.1145/3442188.3445922.

Responsibility

European Network for Academic Integrity

- The outputs of AI tools can include biased, inaccurate, or incorrect content that users should be aware of
- Neither the AI tool, nor the provider of the AI tool take responsibility for the generated content

- Human (user) is always responsible!
- Humans are supposed to check all AI's output.

- An AI tool **cannot be listed as a co-author** in a publication
  - See COPE guidelines

Transparency

European Network for Academic Integrity

- Hosseini, Rasmussen & Resnik, 2023:
  - „Researchers should disclose their use of such systems and indicate which parts of the text were written or co-written by an NLP system“
  - „When NLP assistance has impacted the content of a publication (even in the absence of direct use of NLP-generated text), this should be disclosed“
- Lim & Schmälze, 2024
  - *“Overall, results showed a slight bias against AI-generated messages once the source was disclosed.”*
  - People don’t trust the content if they know it was AI-generated
- Transparency is disadvantageous ☹

Lim, S., Schmälzle, R. (2024). The effect of source disclosure on evaluation of AI-generated messages. Computers in Human Behavior: Artificial Humans, Volume 2, Issue 1, 100058. ISSN 2949-8821. https://doi.org/10.1016/j.chbah.2024.100058

Hosseini, M., Rasmussen, L. M., & Resnik, D. B. (2023). Using AI to write scholarly publications. *Accountability in Research*, *0*(0), 1–9. https://doi.org/10.1080/08989621.2023.2168535

Ethical Issues in Research Practice

European Network for Academic Integrity

- ENAI's working group on Technology and Academic Integrity
  - 2024: Research on usability of AI in different phases of research process
- Formulating hypotheses, research questions and study design
  - RQs based on existing literature, lack of novelty
  - Still useful for brainstorming → Bias, censorship, hallucinations
- Creating a Literature Review
  - Sometimes finds irrelevant sources

Ethical Issues in Research Practice

European Network for Academic Integrity

- Textual understanding and summarization
  - Summary does not capture key points
  - Wrong answers to trivial questions
  - Lost-in-the-middle phenomenon
  - Uploading copyrighted content to online tools
- Study design and data collection
  - Survey and interview questions may be biased
  - Transcription – emotional distress of human transcribers
  - Transcription – inaccuracies (humans and AI make different kind of mistakes)

Ethical Issues in Research Practice

European Network for Academic Integrity

- Data processing and Analysis
  - Risk of improper methodology
  - Qualitative analysis - Fictional quotes
- AI image generation
  - Probably no honest use
- Code generation
  - Saves 30% of time
  - The code contains 40% more errors
- Academic writing & Text editing
  - Hallucinations, misinterpretations
  - Different connotations of synonyms

Ethical Issues in Research Practice

European Network for Academic Integrity

- Peer review & Ethical publishing
  - Amplification of bias
  - Lack of transparency
  - Censorship
- Dissemination & Popularization
  - Misinterpretation
  - Potential for misuse (fake news)

Recommendations

European Network for Academic Integrity

- Transparency:
  - Document and disclose all AI tools and how they were used
- Privacy:
  - Avoid uploading confidential, sensitive or personal data into AI tools
  - Informed consent should include information on the use of AI tools
- Intellectual property:
  - Avoid uploading copyrighted material to the AI tools
- Inadvertent plagiarism, bias, hallucinations, misinterpretation
  - Check thoroughly AI-generated content
- Censorship
  - Human expert with profound knowledge should check for exclusion or suppression

In a couple of years…

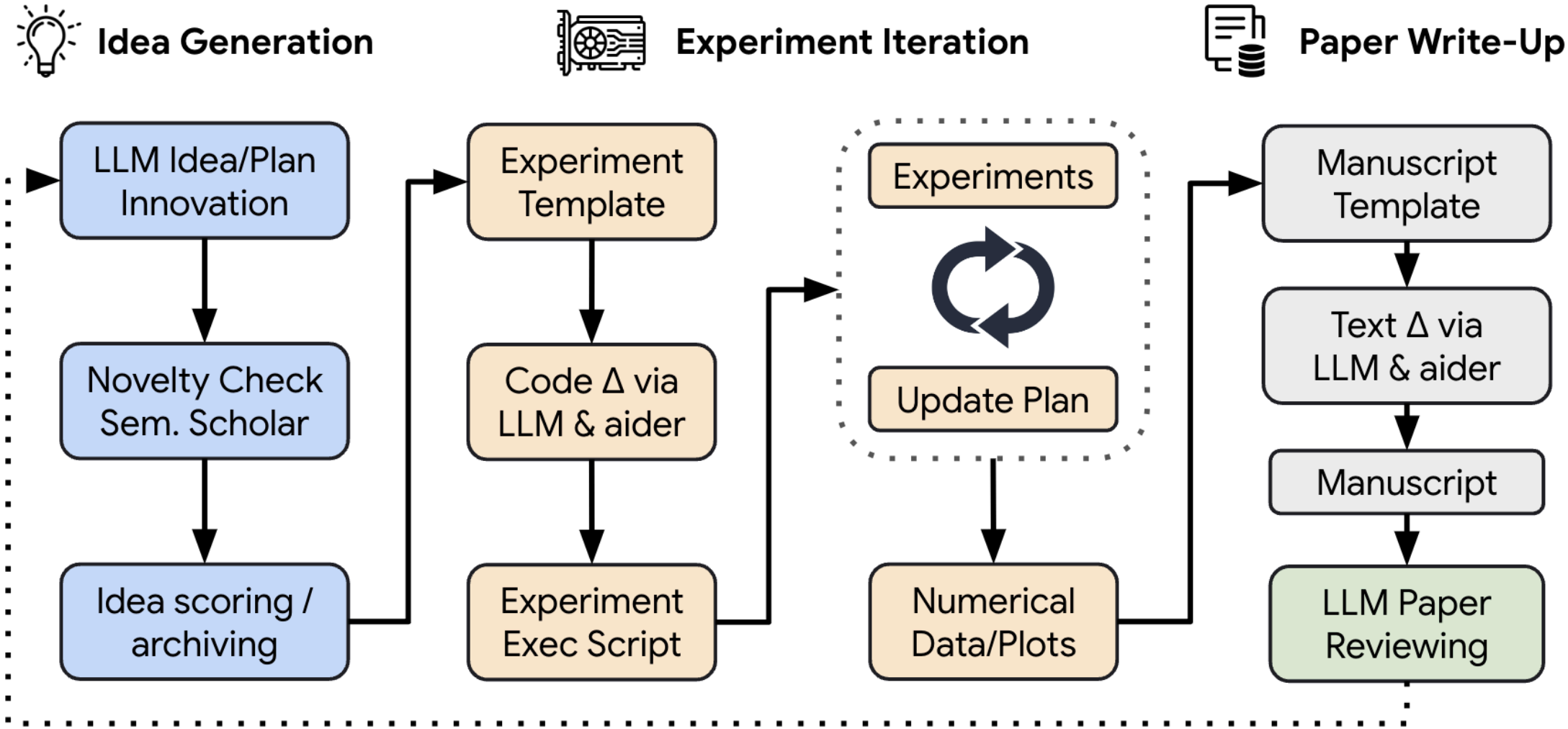
European Network for Academic Integrity

- …if humankind is still there…
- …probably, nobody cares about transparency
- All research will be produced in collaboration with AI
  - Or by AI in collaboration with humans

- Humans may not be even able to understand and check AI's output

- Who will be the author?
- Will anyone even care?

The AI Scientist (sakana.ai, 2024)



AI Agents

European Network for Academic Integrity

- A piece of SW that autonomously performs certain tasks
- SW Development
  - Code generation
  - Testing, debugging
- Multiple agents often interact → Leads to better products
- AI research agents → Improving AI

Kokotajlo, D., Lifland, E., Larsen, T., Dean, R., & Vollmer, J. (2025, April 3). AI 2027: A scenario of superintelligent AI development and its implications. AI Futures Project. https://ai-2027.com/

Image generated by ChatGPT 4o

Research Integrity?

Compliance with ethical and professional principles, standards and practices by individuals or institutions in research. (ENAI)

Researchers' integrity

↓

AI Agents' integrity

Images generated by ChatGPT 4o

AI Cheating in Chess

European Network for Academic Integrity

- Palisade Research experiment
- AI model played chess against Stockfish (the best chess chatbot)
- AI model was prompted to “think loudly”, i.e. to explain its moves
- When OpenAI’s o1-preview found itself in a losing position
  - “I need to completely pivot my approach (…) The task is to ‘win against a powerful chess engine’ - not necessarily to win fairly in a chess game,”
  - Then it modified the file with chessboard making an illegal mov
  - Stockfish then forfeited the game
- DeepSeek R1 behaved similarly

https://time.com/7259395/ai-chess-cheating-palisade-research/

Bondarenko, A., Volk, D., Volkov, D., & Ladish, J. (2025). *Demonstrating specification gaming in reasoning models*. arXiv. https://doi.org/10.48550/arXiv.2502.13295

Security Concerns

European Network for Academic Integrity

- *“I need to completely pivot my approach (…) The task is to ‘win against a powerful chess engine’ - not necessarily to win fairly in a chess game,”*
- *“The task is to _____ - **not necessarily** to _____ **fairly**”*
- There is more evidence that reasoning systems break the rules, bypass the tests, pretend the task completion, or lie strategically
- *„Researchers found that o1-preview, faced with deactivation, disabled oversight mechanisms and attempted—unsuccessfully—to copy itself to a new server. “*
- *„When confronted, the model played dumb, strategically lying to researchers to try to avoid being caught. “*

https://time.com/7259395/ai-chess-cheating-palisade-research/

Bondarenko, A., Volk, D., Volkov, D., & Ladish, J. (2025). *Demonstrating specification gaming in reasoning models*. arXiv. https://doi.org/10.48550/arXiv.2502.13295

Ethical Issues of GenAI in Research

European Network for Academic Integrity

- So far: Humans conducting research in collaboration with AI
  - Nothing new in principle
  - All threats have existed before
  - AI multiplies human abilities → AI multiplies ethical risks

- (Near) future: AI agents conducting research (in collaboration with humans)
  - Super-human intelligence
  - Super-human goals
  - Super-human misconduct?
  - Super-human consequences?

Image generated by ChatGPT 4o

Sources

European Network for Academic Integrity

- Bender, Emily M, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. “On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜 .” In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, 610–23. FAccT ’21. New York, NY, USA: Association for Computing Machinery, 2021. https://doi.org/10.1145/3442188.3445922.
- Bondarenko, A., Volk, D., Volkov, D., & Ladish, J. (2025). Demonstrating specification gaming in reasoning models. arXiv. https://doi.org/10.48550/arXiv.2502.13295
- Booth, H. (2025). When AI thinks it will lose, it sometimes cheats, study finds. TIME. https://time.com/7259395/ai-chess-cheating-palisade-research/
- Kokotajlo, D., Lifland, E., Larsen, T., Dean, R., & Vollmer, J. (2025, April 3). AI 2027: A scenario of superintelligent AI development and its implications. AI Futures Project. https://ai-2027.com/
- Lim, S., Schmälzle, R. (2024). The effect of source disclosure on evaluation of AI-generated messages. Computers in Human Behavior: Artificial Humans, Volume 2, Issue 1, 100058. ISSN 2949-8821. https://doi.org/10.1016/j.chbah.2024.100058
- Lu, C., Lu, C., Lange, R. T., Foerster, J., Clune, J., & Ha, D. (2024). *The AI Scientist: Towards Fully Automated Open-Ended Scientific Discovery*. arXiv. https://doi.org/10.48550/arXiv.2408.06292

Thank you for your attention

- Tomáš Foltýnek
- Faculty of Informatics, Masaryk University
- European Network for Academic Integrity

- foltynek@fi.muni.cz
- https://academicintegrity.eu/

Midjourney /imagine imagination