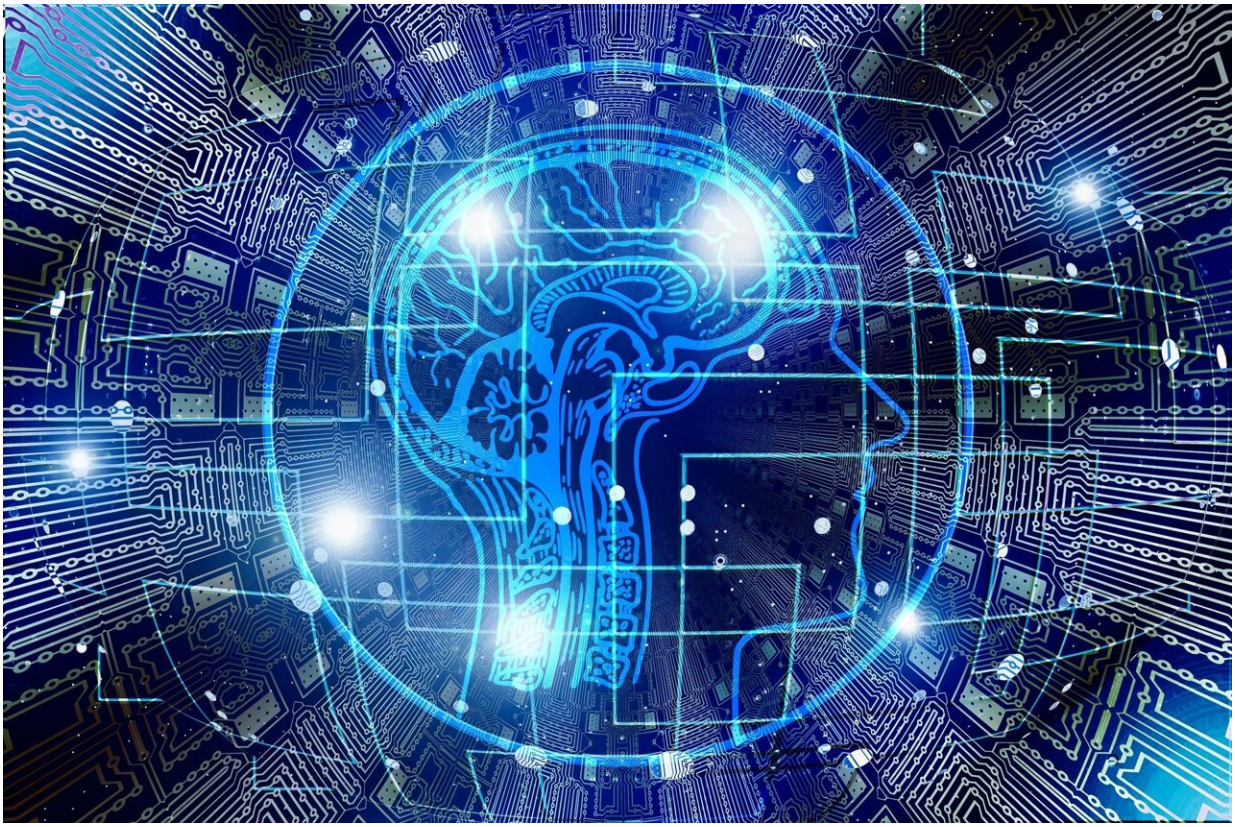
Tackling the ethical dilemma of responsibility in large language models

May 5 2023



Credit: Pixabay/CC0 Public Domain

Researchers at the University of Oxford, in collaboration with international experts, have published a new study in *Nature Machine Intelligence* addressing the complex ethical issues surrounding

responsibility for outputs generated by large language models (LLMs).

The study reveals that LLMs like ChatGPT pose crucial questions regarding the attribution of credit and rights for useful text generation, diverging from traditional AI responsibility debates that primarily focused on harmful consequences.

"LLMs such as ChatGPT bring about an urgent need for an update in our concept of responsibility," say Sebastian Porsdam Mann and Brian D. Earp, the study's joint first authors.

A key finding of the research, according to co-authors Sven Nyholm and John Danaher, "is that while human users of these technologies cannot fully take credit for positive results generated by an LLM, it still seems appropriate to hold them responsible for harmful uses, such as generating misinformation, or being careless in checking the accuracy" of generated text.

This can lead to a situation Nyholm and Danaher, building on previous work, have termed the "achievement gap": "Useful work is being done, but people can't get as much satisfaction or recognition for it as they used to."

Julian Savulescu, the senior author on the paper, adds, "We need guidelines on authorship, requirements for disclosure, educational use, and intellectual property, drawing on existing normative instruments and similar relevant debates, such as on human enhancement." Norms requiring transparency are especially important, continues Savulescu, "to track responsibility and correctly assign praise and blame."

The study, co-authored by an interdisciplinary team of experts in law, bioethics, machine learning, and related fields, delves into the potential impact of LLMs in critical areas such as education, academic publishing,

intellectual property, and the generation of mis- and disinformation.

Education and publishing are particularly in need of rapid action on guidelines for LLM use and responsibility. "We recommend article submissions include a statement on LLM usage, along with relevant supplementary information," state co-authors John McMillan and Daniel Rodger. "Disclosure for LLMs should be similar to human contributors, acknowledging significant contributions."

The paper points out that LLMs may be helpful in education, but warns that they are error-prone, and overuse might affect critical thinking skills. Institutions, the authors write, should consider adapting assessment styles, rethinking pedagogy, and updating academic misconduct guidance to handle LLM usage effectively.

Rights in generated text, such as intellectual property rights and human rights, make up another area in which the implications of LLM use need to be worked out quickly, notes co-author Monika Plozza. "IP rights and human rights pose challenges since they rely on notions of labor and creativity established with humans in mind. We need to develop or adapt frameworks like 'contributorship' to handle this fast-evolving technology, while still protecting rights of creators and users."

Not all foreseeable uses of LLMs are benign. "LLMs can be used to generate harmful content, including large-scale mis- and disinformation," warns co-author Julian Koplin. "That's why we need to hold people accountable for the accuracy of LLM-generated text they use, alongside efforts to educate users and improve content moderation policies to mitigate risks."

To address these and other risks related to LLMs, say co-authors Nikolaj Møller and Peter Treit, LLM developers could follow the example of self-regulation in fields like biomedicine. "Building, and deserving, trust

is crucial to the further development of LLMs. By promoting transparency and engaging in open discussions, LLM developers can demonstrate their commitment to responsible and ethical practices."

More information: Sebastian Porsdam Mann et al, Generative AI entails a credit–blame asymmetry, *Nature Machine Intelligence* (2023). DOI: 10.1038/s42256-023-00653-1

Provided by University of Oxford