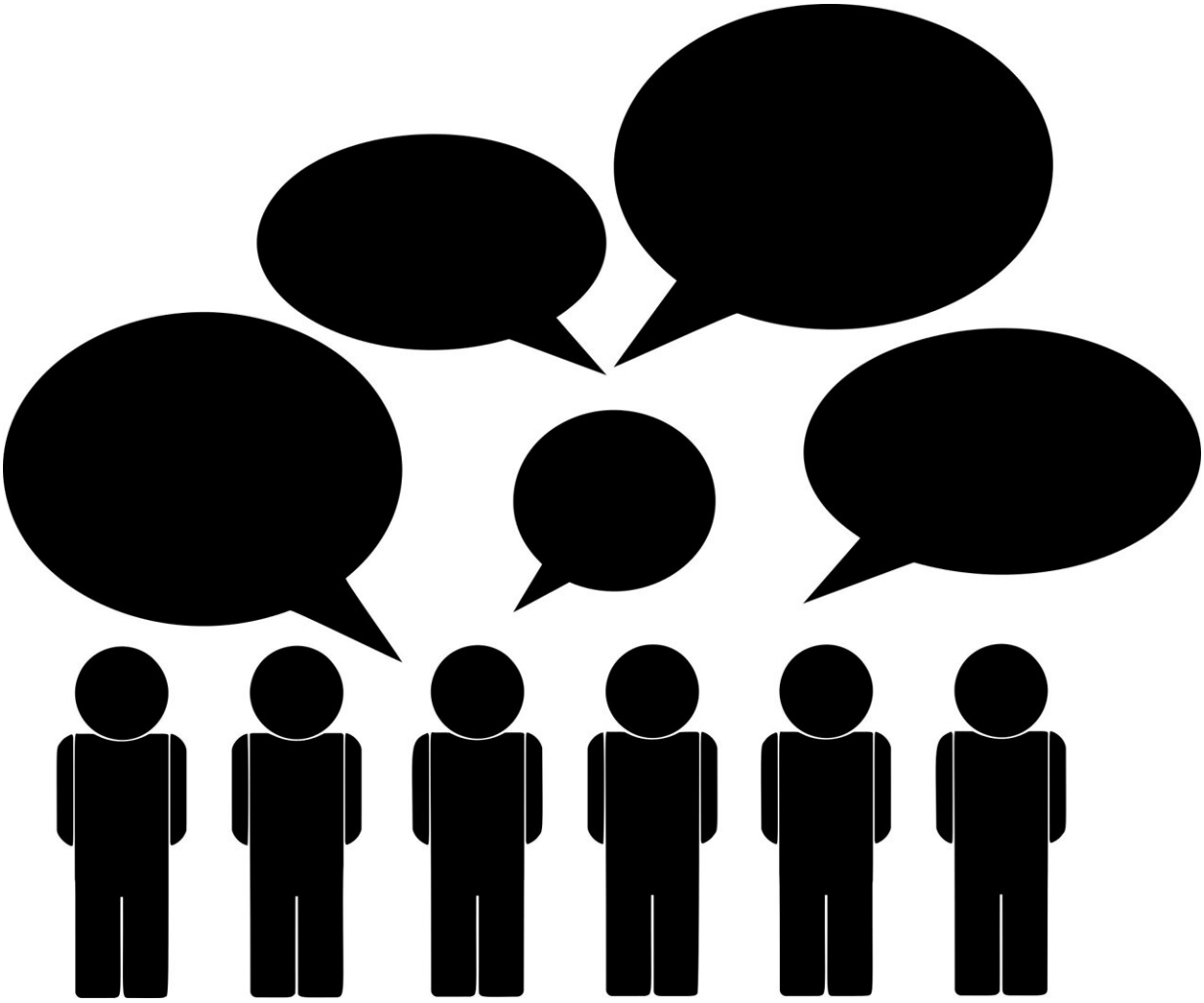
Large language models pose a risk to society and need tighter regulation, say researchers

August 13 2024



Credit: Pixabay/CC0 Public Domain

Leading experts in regulation and ethics at the Oxford Internet Institute, have identified a new type of harm created by LLMs which they believe poses long-term risks to democratic societies and needs to be addressed by creating a new legal duty for LLM providers.

In their paper "Do large language models have a legal duty to tell the truth?" published by the *Royal Society Open Science*, the Oxford researchers set out how LLMs produce responses that are plausible, helpful and confident but contain factual inaccuracies, misleading references and biased information. They term this problematic phenomenon as 'careless speech,' which they believe causes long-term harm to science, education and society.

Lead author Professor Sandra Wachter, Professor of Technology and Regulation, Oxford Internet Institute says, "LLMs pose a unique risk to science, education, democracy, and society that current legal frameworks did not anticipate. This is what we call 'careless speech' or speech that lacks appropriate care for truth.

"Spreading careless speech causes subtle, immaterial harms that are difficult to measure over time. It leads to the erosion of truth, knowledge and shared history and can have serious consequences for evidence-based policy-making in areas where details and truth matter such as health care, finance, climate change, media, the legal profession, and education.

"In our new paper, we aim to address this gap by analyzing the feasibility of creating a new legal duty requiring LLM providers to create AI models that, put simply, will 'tell the truth."'

This phenomenon of 'careless speech' is further complicated by human feedback that often favors outputs that align with their personal biases, and annotations that train models to generate 'assertive sounding

outputs,' among other factors unrelated to advancing truthful outputs.

Associate Professor and Research Associate Dr. Chris Russell, Oxford Internet Institute said, "While LLMs are built so that using them feels like a conversation with an honest and accurate assistant, the similarity is only skin deep, and these models are not designed to give truthful or reliable answers. The apparent truthfulness of outputs is a 'happy statistical accident' that cannot be relied on."

To better understand the legal restrictions faced when using LLMs, the researchers carried out a comprehensive analysis, assessing the existence of truth-telling obligations in the current legal frameworks such as the Artificial Intelligence Act, the Digital Services Act, Product Liability Directive and the Artificial Intelligence Liability Directive.

They find that current legal obligations tend to be limited to specific sectors, professions or state institutions and rarely apply to the private sector.

Commenting on the findings, Director of Research, Associate Professor Brent Mittelstadt said, "Existing regulations provide weak regulatory mechanisms to mitigate careless speech and will only be applicable to LLM providers in a very limited range of cases.

"Nevertheless, in their attempts to eliminate 'hallucinations' in LLMs, companies are placing significant guardrails and limitation on these models. This creates a substantial risk of further centralizing power in a few large tech companies to decide which topics are appropriate to discuss or off limits, which information sources are reliable, and ultimately what is true."

The Oxford academics argue that LLM providers should better align their models with truth through open, democratic processes. They

propose the creation of a legal duty for LLM providers to create models that prioritize the truthfulness of outputs above other factors like persuasiveness, helpfulness or profitability.

Among other things, this would mean being open about the training data they use and the limitations of their models, explaining how they fine-tune models through practices such as reinforcement learning from human feedback or prompt constraints, and building in fact checking and confidence scoring functions into outputs.

Professor Wachter concludes, "Current governance incentives focus on reducing the liability of developers and operators and on maximizing profit, rather than making the technology more truthful. Our proposed approach aims to minimize the risk of careless speech and long-term adverse societal impact while redirecting development towards public governance of truth in LLMs."

More information: Sandra Wachter et al, Do large language models have a legal duty to tell the truth?, *Royal Society Open Science* (2024). DOI: 10.1098/rsos.240197

Provided by University of Oxford