CERN FOR AI: A Necessity For Our Global Future

By **Ljubisa Bojic**

GENERATIVE AI, a subset of artificial intelligence capable of multimodal content creation, is evolving toward general AI systems able to perform a broad array of tasks. Contemporary neural networks, often seen in technologies like Large Language Models (LLMs) and ChatGPTs, appear to possess cognitive abilities, acting as general brains.

Some researchers claim these algorithms even possess a "Theory of Mind," which refers to the understanding that other people have their own thoughts, feelings, beliefs, and perspectives that are different from one's own. On the other hand, our team found that GPT-4 surpassed human performance in linguistic pragmatics, demonstrating superior understanding of complex human dialogues with various linguistic challenges.

At the same time as these research inquiries take place, developers are rapidly integrating these powerful algorithms into various aspects of our daily life and industries. However, these advancements come with a significant challenge: the potential loss of human agency to AI-driven applications. Without human control and robust governance, the dominance of AI could lead to unpredictable and potentially catastrophic outcomes.

There is an immediate need for solutions implemented through multinational and multidisciplinary projects, agencies, and agreements. A tripartite strategy comprising an AI Observatory, complex virtual reality simulations akin to CERN, and improvements in recommender systems could set the stage for a safer AI future.

The AI Observatory: A Safeguard for AI Values and Capabilities

The first essential institution in this framework is an AI Observatory—an agency dedicated to continuously monitoring and testing AI technologies. This body would systematically prompt various language models to understand the values and attitudes these models express. By conducting parallel surveys of human values, the observatory can compare AI's alignment with human values. The AI Observatory would also measure AI's evolving capabilities and detect emerging properties as these systems become more humanoid.

As AI's cognitive and emotional capacities grow, monitoring these aspects becomes crucial in developing empathetic AI that aligns with human values and ethical standards. This becomes even more relevant as machine consciousness and superintelligence emerge.

Complex Simulations: Testing AI in Virtual Reality

The second critical component involves creating complex virtual reality simulations that mirror our world. These simulations would serve as testing grounds for general-purpose, multimodal algorithms. Within these virtual environments, AIs would interact, communicate, and collaborate, each endowed with specific "personalities", backgrounds, goals, and the autonomy to function independently.

Imagine a video game that plays itself, with inputs controlled by human overseers. Such simulations could open up questions about machine consciousness and its evolution within a controlled setting. The insights garnered from these simulations would not only aid in developing safe and aligned AI systems but would also ignite philosophical debates about our own existence. If we can create a controlled, semi-self-evolving world, does it suggest that our own reality might also be a simulation?

Recommender Systems: Solving the Algorithmic Challenges of Social Media

The third pillar, arguably the most pressing, is addressing the issue of outdated AI models, notably recommender systems often employed by social media platforms. As these models would be w integrated with generative AI, their effects would be amplified. There is growing evidence that current algorithms contribute to increased social polarization, media addiction, and limited creativity. These recommender systems have, for too long, acted as echo chambers, amplifying specific content while restricting exposure to diverse viewpoints.

Developing balanced algorithms capable of providing a mixture of educational content, entertainment, and topic plurality could mitigate these effects. The goal is to curate social media feeds that foster balanced emotions and varied perspectives, promoting a more informed and less polarized society.

The Road to Global AI Governance: Multinational Collaboration

The culmination of efforts from the AI Observatory and virtual reality simulations would inform governments and intergovernmental organizations about necessary AI regulations. However, implementing effective global AI governance requires transcending geopolitical rivalries, such as those between the United States and China. The only viable pathway for global cooperation involves leveraging frameworks provided by international bodies like the United Nations (UN).

AI governance should prioritize inclusivity and

collaboration, emphasizing the need to balance technological progress with ethical considerations and human values. This approach aligns with the broader objective of the EMERGE Forum and the scientific conference organized by the Institute for Artificial Intelligence Research and Development of Serbia, in collaboration with the Digital Society Lab at the Institute for Philosophy and Social Theory, University of Belgrade. EMERGE 2024, which will focus on the Ethics of AI Alignment, is scheduled to take place from December 12 to 13 in Belgrade, Serbia.

Conclusion

Establishing a CERN-like entity for AI is not just a futuristic vision but a necessity for our global future. The AI Observatory would serve as a sentinel, continuously monitoring and aligning AI values with human values. Complex virtual reality simulations would function as sophisticated testing grounds to ensure AI's safe and ethical evolution. Revisiting and revising recommender systems would mitigate the social challenges posed by current algorithms, promoting a more balanced and informed society.

The collective insights and data from these initiatives would guide governments and international organizations in crafting effective AI regulations, fostering multinational cooperation, and ensuring that AI technology benefits humanity as a whole.

As we stand on the brink of an AI-driven future, the establishment of a "CERN for AI" represents a proactive step towards ensuring that this future aligns with our shared human values and ethical standards, safeguarding the well-being of future generations.

THE AUTHOR

Ljubisa Bojic is a communication scientist, futurologist, and researcher. As a senior research fellow at both the Digital Society Lab, Institute for Philosophy and Social Theory at the University of Belgrade, and The Institute for Artificial Intelligence of Serbia, his work focuses on the intricate intersections of AI, society, and ethics.

REFERENCES:

[1] *https://ljubisabojic.com/*

[2] *https://emerge.ifdt.bg.ac.rs/*

[3] *Bojic, L., Cinelli, M., Culibrk, D. Delibasic, B. (2024). CERN for AI: a theoretical framework for autonomous simulation-based artificial intelligence testing and alignment. European Journal of Futures Research, 12, 15. https://doi.org/10.1186/s40309-024-00238-0*

[4] *Bojic, L. (2024). AI alignment: Assessing the global impact of recommender systems. Futures, 160, 103383. https://doi.org/10.1016/j.futures.2024.103383*

[5] *Bojic, L. (2022). Metaverse through the prism of power and addiction: What will happen when the virtual world becomes more attractive than reality? European Journal of Futures Research, 10(1), 22. https://doi.org/10.1186/s40309-022-00208-4*

[6] *Bojic, L., Kovacevic, P., & Cabarkapa, M. (2023). Gpt-4 surpassing human performance in linguistic pragmatics (arXiv:2312.09545). arXiv. http://arxiv.org/abs/2312.09545 Kosinski, M. (2024). Evaluating large language models in theory of mind tasks (arXiv:2302.02083). arXiv. http://arxiv.org/abs/2302.02083*