INSIGHT 12

AI COULD BECOME MORE RELIABLE AND TRANSPARENT

In the future, AI could have improved reasoning skills, allowing it to produce better analyses, make fewer factual errors, and be more transparent. However, these improvements may not be enough to overcome problems related to bad data.

TODAY

AI can generate high-quality text and images but can make logical or factual errors. Neural networks[1] are good at recognising patterns without necessarily understanding content or context. For example, a large language model (LLM) uses probability to generate output word by word, based on how often words appear next to other words.[2] It does not understand what the words mean, so its output can contain **hallucinations**.[3] Some developers are seeking to improve factuality and accuracy by giving LLMs access to external knowledge bases through a process called **retrieval augmented generation** (RAG).[4] However, RAG relies on the quality of the source data. For instance, Google Search's AI Overview uses RAG to generate summaries of search queries, but its inability to distinguish authoritative sources from jokes on social media has led it to make recommendations such as putting glue on pizza and looking directly at the sun.[5]

Neural network

A type of AI modeled on the human brain, in which interconnected nodes (neurons) process information in layers to recognize patterns, learn from data, and make decisions. LLMs and image generators are a type of neural network.

Hallucination

When a generative AI system presents false or misleading information as true. This can include false claims, made-up sources, responses to content that was not in the prompt, or images depicting things that are impossible in reality.

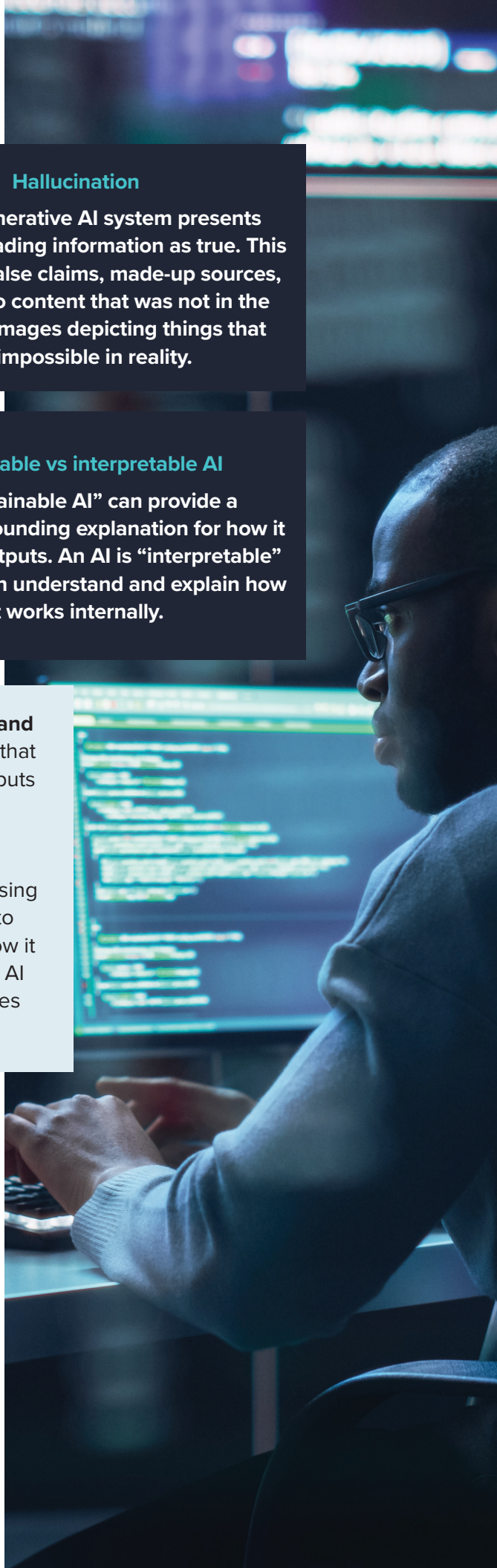
Retrieval augmented generation

An LLM referencing an authoritative knowledge base, outside of its training data, before generating a response.

Explainable vs interpretable AI

An "explainable AI" can provide a reasonable-sounding explanation for how it generates outputs. An AI is "interpretable" if a human can understand and explain how it works internally.

Neural networks lack transparency and can be difficult to understand and control. The calculations they routinely perform are so intricate that even human experts struggle to comprehend how they transform inputs into outputs.[6] In other words, neural networks lack **interpretability**. This makes it difficult to ensure that a generative AI system cannot create harmful content, or that a decision-making AI system is not considering prohibited factors such as race or gender (see insight Using AI to predict human behaviour may not work). Some developers try to overcome this by making their system provide an **explanation** for how it made a decision. However, it cannot currently be known whether an AI model's explanation accurately reflects its actual weighing of variables in a decision.[7]



FUTURES

Future AI systems could combine different approaches to become more functional and versatile. Hybrid AI systems use more than one type of AI: for example, neuro-symbolic AI combines the pattern recognition of neural networks with the human-interpretable rules of **symbolic AI**.[8] In the future, more AI systems could be composed of multiple systems with strengths that make up for each other's weaknesses. These hybrid systems could be more capable, accurate and high performing. They could lead to the development of entirely new types of AI.

AI systems could be more transparent and interpretable, making them less biased. Better interpretability could make it easier for developers to prevent a decision-making AI system from considering factors it should not, such as race. For example, a bank could show a customer why their AI system denied them credit, enabling the customer to seek recourse if – for example – they think the AI put too much weight on their postal code, which reflects that they live in a diverse part of the city. However, this is unlikely to fully eliminate bias (see Insights Bias in AI systems could remain forever and Using AI to predict human behaviour may not work), especially as biases may be inherent in the data.

AI could be less likely to hallucinate, although it is only as good as the data, logic, and training it has access to. Hybrid models, with the ability to reason in different ways, could hallucinate less and provide higher quality analysis. For example, they might use RAG to gather information online then use a symbolic AI to evaluate if the information is credible or a joke. Improved reasoning skills will not solve the underlying problem of a polluted information ecosystem, however (see Insight AI could break the Internet as we currently know it). Future AI systems may approach this problem by collecting more data directly via sensors. They might also be smart enough to recognise when they do not have adequate information to provide an accurate answer.

Symbolic AI

An approach that attempts to mimic human reasoning, in which knowledge is represented as symbols and manipulated in accordance with the rules of a formal logic system, such as deduction and induction.

IMPLICATIONS

- Improved transparency and a reduction in bias could **allow AI to be used in areas where it would be deemed inappropriate today**, such as law enforcement or legal proceedings.
- AI tools could make it **faster and easier to gather high-quality information and perform analysis**, improving decision-making processes and research.
  - **AI research aides could disrupt entry-level jobs** like research assistants, junior analysts, or junior lawyers.
  - **AI analysis would likely be best when limited to trusted, high-quality sources**, like an organization's internal documentation or a database of academic journals.
- **Higher quality analysis and reasoning could speed up adoption of AI** among currently risk-averse organizations. It could reduce the need for oversight and quality control, but also make it less likely that mistakes will be noticed.
  - **Hallucinations may become less obvious**, making them harder to detect.
- Even with improved reasoning, **false or biased inputs could undermine AI performance and reliability**.
  - **AI could still spread misinformation**, where the argumentation and reasoning are valid, but the premises are false.

Endnotes

1 'What Are AI Hallucinations? | IBM', 1 September 2023. https://www.ibm.com/topics/ai-hallucinations.

2 Naveed, Humza, Asad Ullah Khan, Shi Qiu, Muhammad Saqib, Saeed Anwar, Muhammad Usman, Naveed Akhtar, Nick Barnes, and Ajmal Mian. 'A Comprehensive Overview of Large Language Models'. arXiv, 27 December 2023. http://arxiv.org/abs/2307.06435, 4.

3 'What Are AI Hallucinations? | IBM', 1 September 2023. https://www.ibm.com/topics/ai-hallucinations.

4 Martineau, Kim. 'What Is Retrieval-Augmented Generation?' IBM Research Blog, 22 August 2023. https://research.ibm.com/blog/retrieval-augmented-generation-RAG.

5 Friedman, Alan. 'Google's AI Overview Tells Users to Eat Rocks and Glue, Says Google Search Is a Monopoly'. PhoneArena, 24 May 2024. https://www.phonearena.com/news/ai-overview-spews-ridiculous-statements_id158677.

6 Fan, Feng-Lei, Jinjun Xiong, Mengzhou Li, and Ge Wang. 'On Interpretability of Artificial Neural Networks: A Survey'. *IEEE Transactions on Radiation and Plasma Medical Sciences* 5, no. 6 (November 2021): 741–60. https://doi.org/10.1109/trpms.2021.3066428.

7 Fan, Feng-Lei, Jinjun Xiong, Mengzhou Li, and Ge Wang. 'On Interpretability of Artificial Neural Networks: A Survey'. *IEEE Transactions on Radiation and Plasma Medical Sciences* 5, no. 6 (November 2021): 741–60. https://doi.org/10.1109/trpms.2021.3066428.

8 Hitzler, Pascal, Aaron Eberhart, Monireh Ebrahimi, Md Kamruzzaman Sarker, and Lu Zhou. 'Neuro-Symbolic Approaches in Artificial Intelligence'. *National Science Review* 9, no. 6 (4 June 2022): nwac035. https://doi.org/10.1093/nsr/nwac035.

PDF: PH4-222/2025E-PDF
ISBN: 978-0-660-76905-9

Aussi disponible en français sous le titre : L'IA pourrait devenir plus fiable et transparente.