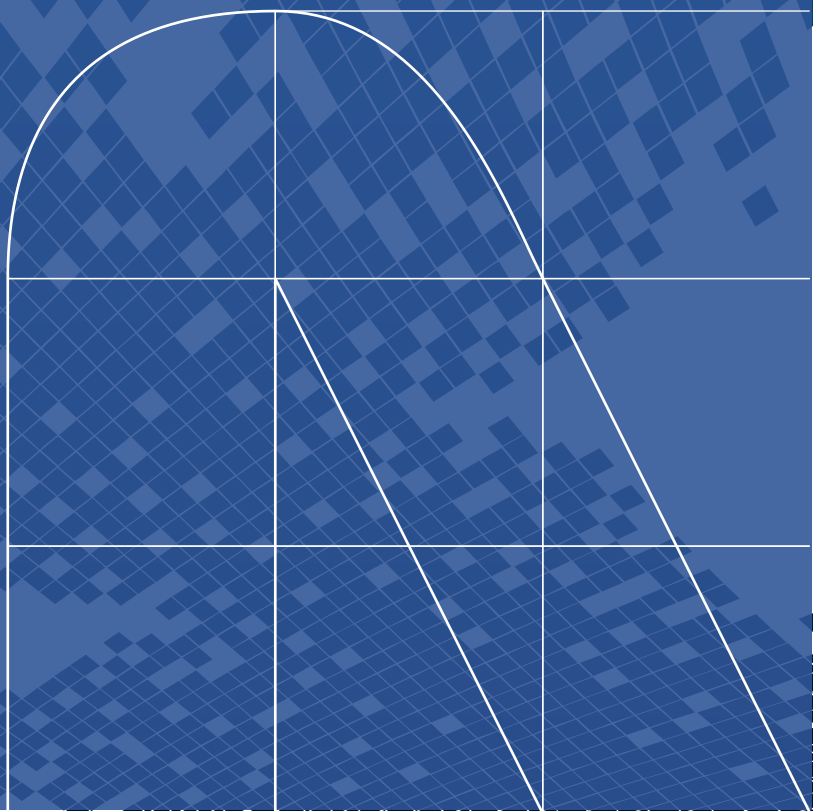
NTT DATA

All Hallucinations are Not Bad. Acknowledging Gen AI's Constraints and Benefits

NTT DATA point of view on Generative AI, examining the concept of hallucination, its implications and responsible AI practices.

NTT DATA Point of View



The AI Daze

The recent spurt of generative AI (GenAI) models has sparked a lot of interest and excitement. These advanced large language models (LLMs) can do many extraordinary things — from crafting email responses to recommending a fitness routine to generating computer code; the list is seemingly endless.

However, it is imperative to acknowledge that GenAI is not without its own set of challenges. Hallucination, is one such phenomenon which is associated with LLMs generating imaginative content extrapolating from the training data in ways that is incomprehensible by humans, has emerged as major concern.

Imagine telling a GenAI model to "*draw a cool scene where dragons are flying over New York, leaving shiny trails behind them*." The computer then creates a fantastic picture where dragons soar over the city, adding a touch of magic to the skyline. However, picture depicting dragons are flying over New York is not real or factually incorrect. This ability to turn unusual ideas into something real is what makes Generative Artificial Intelligence (Gen AI) so intriguing, and also raises questions about reliability of GenAI.

Hallucinations are completely fabricated outputs from large language models. Even though they represent completely made-up facts, the LLM output presents them with confidence and authority[2].

While experts are working towards mitigating the occurrence of hallucination. it's worth recognizing that the scope of this phenomenon is remarkably extensive. There is a parallel narrative, the creative facet of desirable hallucination, this aspect of GenAI has potential to foster innovation and open doors to unimaginable possibilities.

Gen AI Hallucination in the Headlines

Bogus case law[3]: Recently, lawyer relied on ChatGPT to prepare a filing on behalf of a man suing 'Avianca Airlines'. To support the argument, ChatGPT fabricated 3 cases titled Martinez vs. Delta Air Lines, Zicherman vs. Korean Air Lines and Varghese vs. China Southern Airlines, which sounded so real.

The fabrications were revealed when Avianca's lawyers approached the case's judge, saying they couldn't locate the cases cited in lawyers' brief in legal databases. As a result, the federal judge imposed 5K USD fine on two lawyers and a law firm for their submission of fictitious legal research.

Gen AI Defamation: Editor-in-chief of Gun Publication prompted ChatGPT for a summary of the Second Amendment Foundation vs. Ferguson case as background for a case he was reporting on.

ChatGPT provided with a summary of the case which stated that Alan Gottlieb, the Second Amendment Foundation's (SAF) founder, accused Walters of "defrauding and embezzling funds from the SAF.

As a result, Georgia radio host, Mark Walters has sued OpenAI in the company's first defamation lawsuit.

Can we trust everything that Generative AI says?

Nevertheless, the intricate struggle lies in striking the right balance, AI researchers must choreograph a delicate dance that ensures responsible AI use in one hand, whereas nurturing and encouraging creativity that enriches our world.

Generative AI: Foundations and Technology

Generative Artificial Intelligence (Gen AI) acts as a creative mind for machines, empowering them to generate distinct and unique content such as images, stories, or even text that closely mirrors human creation. To truly comprehend the fascinating phenomenon of hallucination within Generative AI, we must investigate the basics or foundational technologies that drive this this kind of creativity.

At the heart of Gen AI's creative power is not limited to a single technique, there are various techniques playing a vital role in this landscape.

The Generative Adversarial Network (GAN) is a key player, showcases a dynamic interplay between a 'generator' and a 'discriminator.' The generator attempts to create content that resembles reality, while the discriminator distinguishes between real and generated content. Through this back-and-forth competition, the generator improves its ability to create realistic outputs, sometimes pushing the boundaries of the ordinary and at times, venturing into the surreal world. This divergence from the expected, often termed as 'hallucination,' showcases the intriguing potential of Generative AI.

In addition to GANs, there are other significant techniques such as Variational Autoencoders (VAEs) and autoregressive models. VAEs can be like an artist honing their skills, learning, and refining through imaginative processes. Whereas, Autoregressive models resemble storytellers, carefully crafting narratives one step at a time.

However, the transformative impact in this landscape was the introduction of Transformer architecture. Initially designed for natural language processing tasks, the Transformer architecture led to revolutionary attention mechanisms. This mechanism allows models to focus on different parts of input sequences, enabling parallelized processing and capturing long-range dependencies effectively. In the context of Generative AI, especially when integrated with GANs, the Transformer architecture has proven to be transformative.

Apart from these aforesaid techniques, we have neural networks which are inspired by the actual human brain's interconnected nodes. These networks learn patterns and structures from extensive data, enabling the generation of content that simulates human creativity. The process involves repetitive ‘training' and/or Reinforcement learning from human feedback (RLHF) that combines reinforcement learning techniques, such as rewards and comparisons, with human guidance to train the model producing and real-world examples.

Understanding how these components and principles work in nexus is pivotal to unravelling the mechanism that may lead to hallucinations within Generative AI. The combination of these technologies, coupled with the sophisticated learning process, often propels machines into uncharted imaginative territories, resulting in creative and, at times, hallucinatory/unreal outputs.

Understanding the Mystery of Data

In the world of Artificial Intelligence (AI), data can be analogous to experiences in the human world. Data acts as both the canvas and the brush interchangeably, allowing machines in the domain of Generative AI to mimic and innovate, mirroring human cognition and creativity.

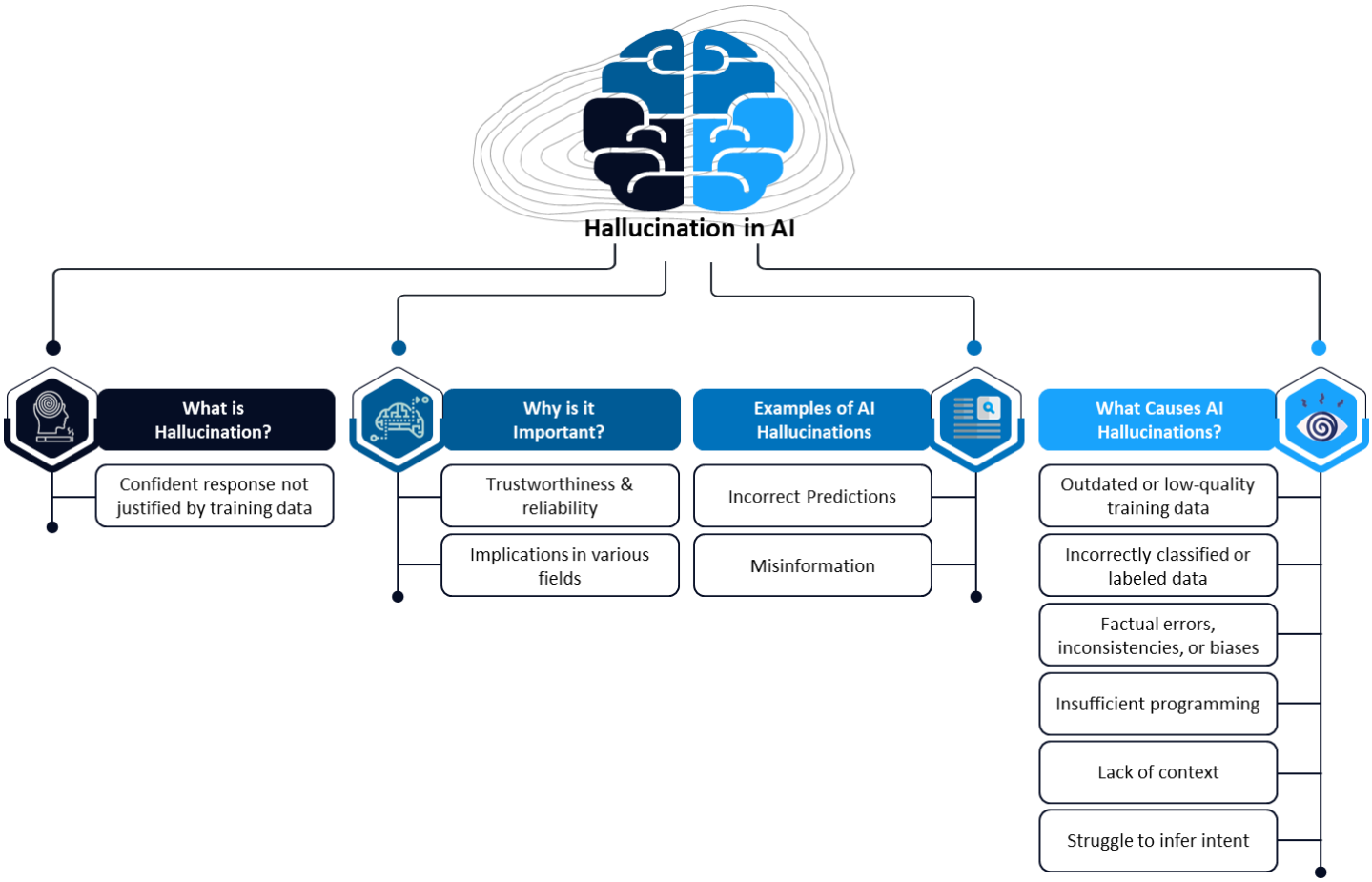
The term “Anthropomorphism” finds expression in the definition and explanation of hallucination within Generative AI models. Generally, hallucination refers to the AI's ability to generate content that deviates from conventional or expected outputs. However, the riskiest one is

when the GenAI model generates an outcome that matches perfectly in what you expect but lacks validation from verified facts. This scenario is more concerning as users might accept the generated content as accurate, inadvertently propagating unverified or false information. The challenge is determining between probable, yet unsubstantiated content and outputs grounded in validated facts to ensure responsible use of Generative AI.

Generative Anthropomorphism: Attributing human-like traits to non-human entities becomes evident as AI systems learn and derive creativity from vast datasets.

Also, as the causes of these hallucinations lie in the AI's tendency to generalize patterns from the data it has been trained on. The process of anthropomorphizing AI creativity often amplifies this tendency to generalize, giving rise to content that pushes the boundaries of reality.

The impact of hallucination in Generative AI is twofold. It primarily fuels innovation and unpredictability in creative processes, enriching the realm of AI-generated content. On the other hand, it poses challenges, particularly when hallucinated outputs dissipate misinformation or harmful/biased stereotypes. In addition, AI systems learn to make decisions based on training data, which can include biased human decisions or reflect historical or social inequities,



This imaginative capability leads to instances where AI-generated outputs blur the usual boundaries.

- Content generation that deviates from conventional or expected outputs that matches perfectly in what you expect but without validated facts.
- Images maybe depicted as mythical creatures, and texts maybe crafted as narratives that defy conventional logic.

even if sensitive variables such as gender, race, or sexual orientation. Anthropomorphism helps us understand these implications and navigate through the ethical and societal implications of hallucination, guiding responsible usage and development of Generative AI.

In the forthcoming sections, we will dive deeper into the intricacies of data, exploring how it influences the hallucinatory potential of

Generative AI and how we can capitalize this captivating phenomenon.

Types of Hallucination

Hallucination in Generative AI can exhibit in various forms, depending on the type of model and the specific task it is designed for. Here are some common types of hallucination in Generative AI:

Visual Hallucination: In image generation models, visual hallucination may involve the creation of images that depict objects, scenes, or patterns that do not exist. These hallucinations can range from surreal and abstract art to entirely fabricated objects or creatures.

Textual Hallucination: Language models may hallucinate text by generating sentences or paragraphs that contain fictional information or make false claims. Textual hallucinations can involve inventing events, details, or facts that have no basis.

Content Expansion Hallucination: This phenomenon occurs when a generative model produces more information than what is present in the input data. For example, a model might add unnecessary details to an image or generate extensive narratives that go beyond the information provided.

Inference Hallucination: In natural language processing tasks, inference hallucination can lead to incorrect assumptions or inferences. Large Language Models (LLM) may draw unwarranted conclusions from input data, leading to responses that misjudge or misrepresent the context.

Bias Hallucination: Bias hallucination refers to the generation of content that mirrors or amplifies biases that is already present in the training data. This can result in outputs that display stereotypes, discrimination, or even present unethical viewpoints.

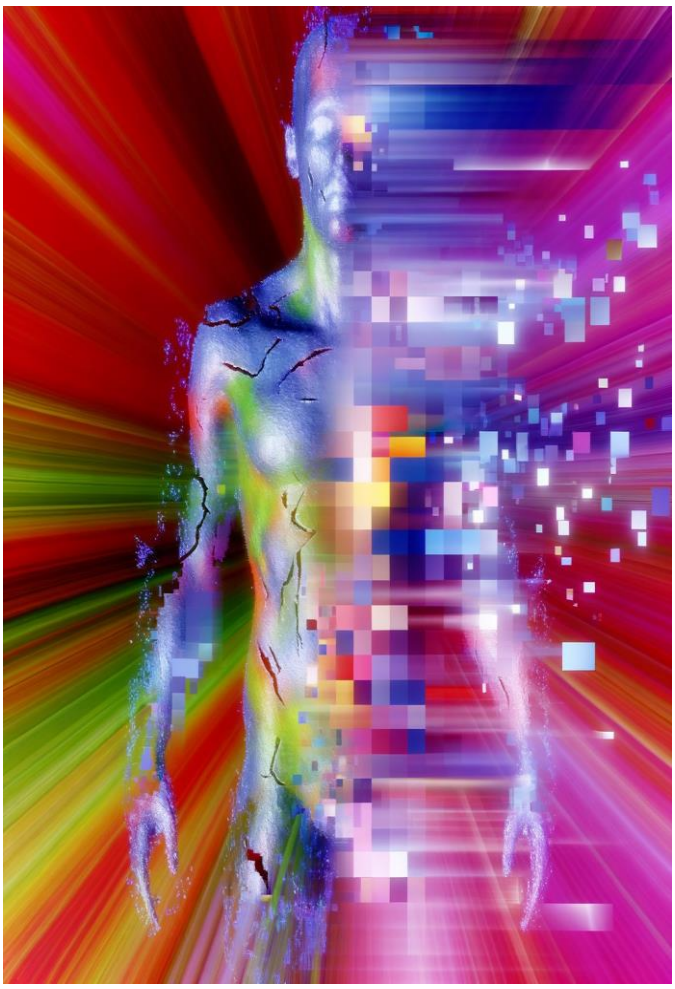
Contextual Hallucination: Language models can suffer from contextual hallucination, where they generate text that seems contextually relevant but is factually incorrect or not representative of the actual context.

It's essential to note that the impact of hallucination can vary amongst different Generative AI models. Mitigating and controlling these various forms of hallucination is a crucial factor in AI research and development to ensure that the future outputs are safe, reliable, and aligned with the set objectives.

Hallucination in Action – In Different sectors

Before we venture, as to how hallucination seeps into various sectors, it's extremely important to understand the versatility of Generative AI and its ability to visualize beyond the ordinary. Just as human creativity knows no bounds, Generative AI's hallucination presents a glimpse of imagination blending with technology.

Now, let's assume we have implemented Gen AI in various sectors and this aspect of Generative AI hallucination is set into to diverse sectors, where its impact is both appealing and transformative. From revolutionizing communication in the telecommunications sector to optimizing business processes in BPO, hallucination metamorphizes uniquely in each sector, introducing innovative possibilities.



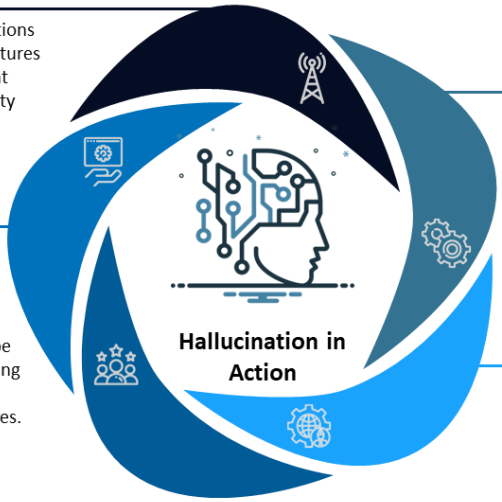
Let us journey through different domains, delving into how hallucination materializes within each sector and its potential ramifications.

Telecommunications Sector

Generative AI, when applied to the telecommunications sector, may hallucinate innovative network architectures or communication protocols that are beyond current technological constraints. This hallucinatory creativity can fuel disruptive ideas but demands rigorous evaluation before implementation.

IT Service Management (ITSM)

Within ITSM, Generative AI hallucination could generate novel troubleshooting approaches or optimization strategies for IT processes. However, distinguishing hallucinated proposals from viable solutions is crucial. Practical implementation must be assessed against scalability, compatibility with existing systems, and adherence to industry best practices, ensuring hallucinated ideas align with ITSM objectives.



Business Process Outsourcing (BPO)

Generative AI applied in BPO could hallucinate automation processes and workflow optimizations. However, discerning hallucinated processes from viable automation strategies is vital. Evaluation based on resource efficiency, integration complexity, and alignment with business goals is essential to select hallucinated processes that can be effectively integrated into BPO operations.

Customer Experience

In the realm of customer experience, Generative AI may hallucinate unique customer engagement strategies based on complex behavioral models. While these hallucinated strategies may promise highly personalized experiences, their practicality needs thorough validation. Evaluating hallucinated strategies for scalability, integration capabilities, and alignment with organizational goals is essential to extract genuine value from the hallucination process.

Generative AI implemented in organizations do not connect with various data types, such as internal rules and work-related materials, to generate content, leading to hallucinatory responses.

Nevertheless, we need to effectively leverage hallucination in these sectors that involves careful evaluation of hallucinated outputs for technical feasibility, adherence to industry standards, and alignment with operational goals. We can strike a balance between innovation and practicality is crucial to maximize the transformative impact of Generative AI hallucination within these specialized domains.

Leveraging Hallucination: Transforming Drawbacks into Advantages

Generative AI hallucinations, while often seen as a problem, can also be employed for better and positive purposes. This feat can be achieved by understanding how and when hallucinations occur, we can develop techniques to exploit them for creative and innovative applications.

One simple way to leveraging generative AI hallucinations is to use them to generate new ideas and concepts. For instance, a generative AI model could be used to create new product designs, advertisements, or even movie plotlines. By providing the model with an initial prompt and allowing it to hallucinate freely, we can unleash new possibilities that were not possible by a human to comprehend.

Another way to leverage generative AI hallucinations is to use them to create synthetic data. Synthetic data is a form artificial data that is generated to mimic real-world data. It can be used for a variety of purposes, such as training machine learning models, testing new models, and emulating/simulating complex systems.

Generative AI hallucination can be used to create synthetic data that is more realistic and varied than traditional synthetic data generation techniques.

According to Gartner®, " **Generative AI in material science -** Generative AI is impacting the automotive, aerospace, defense, medical, electronics and energy industries by composing entirely new materials targeting specific physical properties " "**Generative AI in drug design-** Generative AI has already been used to design drugs for various uses within months, offering pharma significant opportunities to reduce both the costs and timeline of drug discovery."[1]

Finally, generative AI hallucinations can also be used to create new forms of art and entertainment. For example, a generative AI model could be used to create new music genres, films, or even video games. By allowing the model to hallucinate freely, we can create new

and unique experiences that are not possible with traditional methods.

As generative AI models continue to learn and improve, we can expect to see even more innovative and creative applications of generative AI hallucination in the future.

However, we need to ensure that generative AI hallucinations should always be used with caution and pinch of salt, as hallucinations can be misleading and inaccurate, so it is important to verify the results of any generative AI model before using them in any real-world application.

Mitigating Hallucination in Generative AI

Hallucinations in Generative AI, while fostering creativity, can lead to challenges, especially when model outputs diverge too far from the intended domain. Mitigation strategies aim to ensure that the AI's creative process remains aligned with human expectations and practical applications. This section delves into the techniques, research advancements, best practices[4], and real-world case studies that address the issue of hallucination.

There are several techniques and strategies that can be used to mitigate hallucinations in generative AI models. Some of the most common include:

- **Prompting:** By providing the model with a clear and specific prompt, we can help to guide it towards generating more accurate and realistic outputs.
- **Diverse Dataset:** Training the model on a diverse and factual dataset of text and code can help it to learn a wider range of patterns and relationships, which can make it less likely to hallucinate and provide accurate data.
- **Improved GAN Architecture:** Some model architectures, such as generative adversarial networks (GANs), are specifically designed to generate more realistic outputs.
- **RAG Implementation:** By employing RAG hallucinations can be reduced as it provides the model with relevant context and information to improve the accuracy of its responses.
- **Output Filtering:** Once the model has generated an output, it needs to be filtered to remove any hallucinations or inaccurate information.
- **Human in the Loop:** Introduce human evaluator in the loop to assess outputs of the models, providing critical feedback on whether the generated content aligns with real-world expectations and facts.

A simple human review before sharing could have prevented the spurious Gen AI fake case issue[3].

Research Advancements & Best Practices

There is a growing body of research on how to mitigate hallucinations in generative AI models. Some of the most recent advancements include:

- **Hallucination Detection**: Researchers have developed new algorithms to detect hallucinations in generative AI outputs. This can be useful for identifying and removing hallucinations before they are used in any real-world application.
- **Hallucination Correction**: There are new algorithms to correct hallucinations in generative AI outputs. This can be useful for improving the accuracy and realism of generative AI outputs.

Some of the best practices for mitigating hallucinations in generative AI include:

- Use a variety of mitigation techniques: There is no single technique that can eliminate hallucinations in generative AI. The best approach is to use a variety of techniques in combination.
- Monitor the model's output: It is important to monitor the model's output regularly for signs of hallucination. This can be done manually or by using an automated hallucination detection algorithm.
- Be aware of the limitations of generative AI: Generative AI models are still under development and are not perfect. It is important to be aware of the limitations of these models and to use them with caution.

It is important to note that there is no single solution that can eliminate hallucinations in generative AI. However, by using a combination

of the techniques and strategies described above, we can significantly reduce the risk of hallucinations and improve the accuracy and realism of generative AI outputs.

OpenAI's potential new strategy for fighting the fabrications: Train AI models to reward themselves for everyone, correct step of reasoning when they're arriving at an answer, instead of just rewarding a correct conclusion. The approach is called "process supervision," as opposed to "outcome supervision," and could lead to better explainable AI[5].

Ethical and Societal Implications of Hallucination in AI

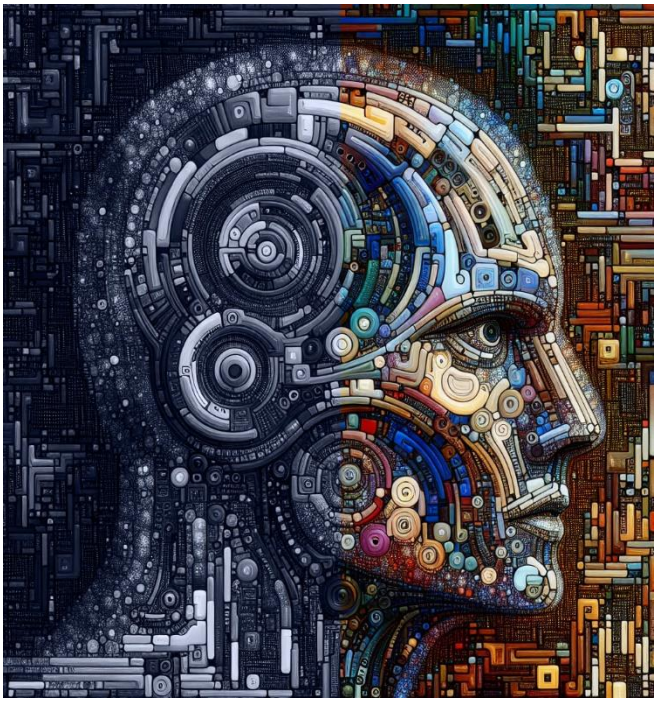
The emergence of hallucination in AI models, while filled with creative potential, necessitates a thoughtful examination of the ethical considerations and societal impacts that accompany this technology.

Discussion on Ethical Considerations

Truth and Falsity: The creation of content that may not align with reality raises ethical concerns. AI-generated hallucinations can blur the lines between fact and fiction, potentially leading to the dissemination of false information, which in turn has consequences for informed decision-making and trust.

Bias and Discrimination: Hallucinations produced by AI models may amplify biases present in training data. This raises concerns about perpetuating stereotypes, reinforcing societal prejudices, and exacerbating issues related to bias and discrimination.

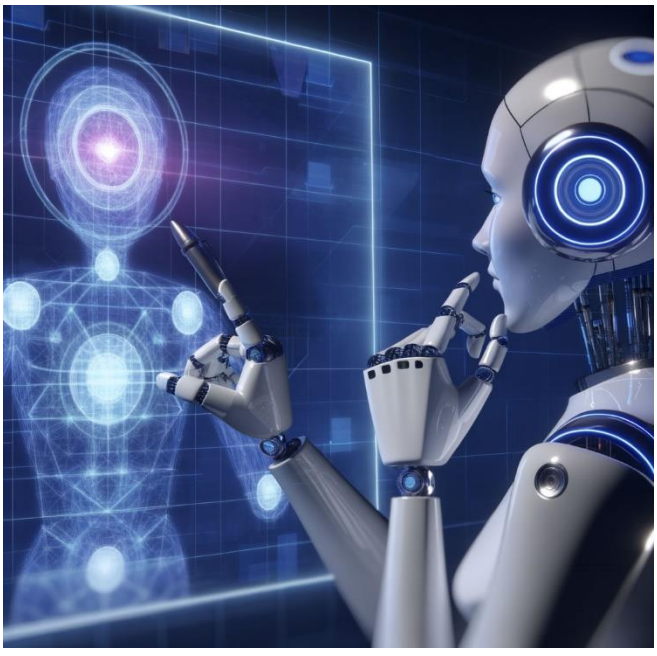
Privacy and Consent: Generating content, especially in personal contexts, can infringe upon individuals' privacy and consent. Considerations of consent, data usage, and potential harm to individuals featured in hallucinated content must be addressed.



Potential Impacts on Society

Misinformation and/or Disinformation: Hallucinations can become vehicles for the spread of misinformation and disinformation, impacting public discourse and trust in information sources. This may require active measures to combat the dissemination of false or misleading content.

Security Concerns: AI-generated hallucinations can be exploited for malicious purposes, including deepfakes or deceptive content used in cybersecurity attacks or fraud.



Responsible Use and Guidelines for Mitigating Negative Consequences

Transparency and Accountability: Developers and users of AI systems must prioritize transparency in the generation process and be accountable for the content produced. Clearly indicating AI-generated content can help mitigate the risk of misinformation.

Bias Mitigation: Implementing bias reduction techniques and actively working towards debiasing training data are primary steps in mitigating the amplification of biases in hallucinated content.

Ethical Guidelines: Framing and adhering to ethical guidelines for AI model development and usage can set standards for responsible AI deployment and content generation.

As we navigate through the world of AI hallucination, ethical considerations and societal impacts remain pivotal. Responsible development, usage, and guidelines are indispensable in ensuring that the benefits of this technology can be harnessed while minimizing its negative consequences.

Charting the Future with Generative AI – NTT DATA's Path Ahead

In our exiting journey of Generative AI and the captivating phenomenon of hallucination, we have uncovered a transformative landscape along with creativity and innovation. As we conclude our journey, we find ourselves standing at the crossroads of immense potential and profound responsibility.

The creative capacities of Generative AI, exemplified through hallucination, open doors to unprecedented possibilities across numerous sectors. From telecommunications to customer servicing, this technology can reshape how we perceive and interact with the digital world.

However, the path is not without its ethical and societal considerations. The generation of content that blurs the lines between fact and fiction necessitates an ethical framework that values transparency, accountability, and responsible usage. As we have witnessed, the impact of hallucination extends to bias mitigation, combating misinformation, and safeguarding individual privacy.

In this dynamic landscape, it is upon IT industry leaders, such as NTT DATA, to chart the way forward. Responsible development and adherence to ethical guidelines are central to harnessing the power of AI creativity while mitigating its potential pitfalls.

The journey into Generative AI continues, and the future promises remarkable possibilities. It is a journey that will be shaped by innovative practices, transparency, collaboration, and the commitment to responsible AI utilization. As we step into this future, we find ourselves not at the conclusion, but at the inception of a new era where AI creativity knows no bound.

Let's get started

See what NTT DATA can do for you.

- Deep industry expertise and market-leading technologies
- Tailored capabilities with your objectives in mind
- Partnerships to help you build and realize your vision.

Contact one of our authors or visit **nttdata.com** to learn more.

Sources

1. Gartner, Beyond ChatGPT: The Future of Generative AI for Enterprises, https://www.gartner.com/en/articles/beyond-chatgpt-the-future-of-generative-ai-for-enterprises. January 26, 2023. GARTNER is a registered trademark and service mark of Gartner, Inc. and/or its affiliates in the U.S. and internationally and is used herein with permission. All rights reserved.
2. https://www.techtarget.com/whatis/definition/AI-hallucination
3. https://www.forbes.com/sites/mollybohannon/2023/06/08/lawyer-used-chatgpt-in-court-and-cited-fake-cases-a-judge-is-considering-sanctions/?sh=5f0c9f937c7f
4. https://medium.com/google-cloud/generative-ai-understand-and-mitigate-hallucinations-in-llms-8af7de2f17e2
5. https://www.ibtimes.com/openai-announces-new-approach-fight-ai-hallucinations-after-legal-fabrications-3696947
6. https://www.technologyreview.com/2023/12/14/1085318/google-deepmind-large-language-model-solve-unsolvable-math-problem-cap-set/

All images in this report were generated using Azure Open AI