IMAGE: Adobe Stock

BY EUGENE YIGA

The Unreal World of Synthetic Data

As AI-generated synthetic data becomes increasingly realistic, it raises profound questions about the nature of truth and authenticity in the digital age.

In Accenture's Digital Health Technology Vision report, the consulting firm explores the new frontier of artificial intelligence (AI). The report discusses the concept of "synthetic realness", where AI-generated data, images, and chatbots convincingly mimic the physical world and blur the lines between what's real and what isn't.

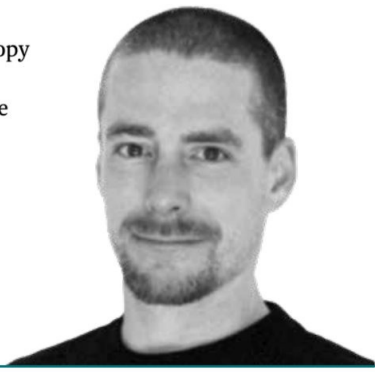
"As we enter the world with synthetic realness, where AI-generated data convincingly reflects the physical world, we are forced to face the questions of what's real, what's not, and perhaps more importantly, when we should care," the report states.

As head of the Critical Media Lab in Basel (Switzerland), Johannes Bruder sees this report as a window into the strategies of corporations and consulting firms that heavily influence the design of the systems that increasingly impact the world. Indeed, he believes that Accenture's dominance allows it to shape not only individual corporate strategies but also the very infrastructures and processes through which large companies interact.

The rise of synthetic data

At the heart of Accenture's vision of the "unreal" is the production and use of synthetic data, i.e., data artificially created by computer simulations or algorithms rather than being generated by actual events. Generative AI means computers can learn patterns from large amounts of real-world data and then generate new content that mimics those underlying patterns without duplicating the original data.

"It's data which is not a trace and not a copy and not a recording, but the product of a computational process such as generative AI," Bruder explains. This synthetic data can be shared and used, for instance, by health practitioners and researchers around the world to develop insights, treatments, and vaccines without privacy concerns.



Johannes Bruder

A rapid proliferation of start-ups focused on synthetic data...

American IT industry research and advisory firm Gartner predicts that the majority of the data used for AI development and analytics projects will be synthetically generated. This has led to a rapid proliferation of start-ups focused on synthetic data, all catering to a wide range of industry sectors. Synthetic data is also used to train machine learning systems, which allows for the inclusion of edge cases, outliers, and even "black swan" events that are possible but not necessarily likely or typical.

"Such events and data need to be ingested by machine learning algorithms so that they can be anticipated in the future to avoid, for instance, catastrophic crashes in financial markets and health crises," Bruder notes.

The economics of machine learning

However, the inability of machine learning systems to recognise specific patterns and predict specific events is not only due to a shortage of real-world data but also a result of the economics of machine learning itself. To remain manageable in terms of size and energy consumption, and to remain open to new discoveries, machine learning systems need to unlearn and forget.

"The so-called long tail of the data distribution will often be categorised as insignificant and therefore deprioritised or simply forgotten," Bruder says.

Synthetic data offers an opportunity to increase the sensitivity of machine learning systems for detecting or predicting precisely the data that is erased due to model bias and other unlikely but plausible events. It also allows for the curation of this process, as synthetic data generators emphasise the opportunity to decide what the synthetic reality suggested by the dataset should look like.

From realness to authenticity

To legitimise this operation, Accenture suggests substituting the term "real" with "authentic". The goal is to shift the conversations about AI from good (real) and bad (fake) to focus on authenticity. So instead of asking, "Is this real?" we'll begin to ask, "Is this authentic?"

"Privileging authenticity over realness is tantamount to moving the conversation towards what we should care about or what is necessary or right, regardless of whether or not it's real," Bruder says.

In machine learning, data used to train algorithms is often referred to as "ground truth" as it represents real-world events as a benchmark for statistical operations. However, synthetic data producers are not in the business of data collection but rather in the business of synthesising new worlds based on statistical properties extracted from existing real-world datasets.

"We are one level removed from the metaphorical ground, where you collect data via the concept of statistical properties," Bruder explains. These statistical properties refer only to the relations the data points have to each other, not to any actual events. The ground truth, in this case, is not real in a traditional sense but grounded in an opaque process of mimicking real data and in decisions about how close or similar any curated "unreal" needs to be to what we still conceive as authentic.

Architects of the unreal world

Accenture's report promotes a new willingness to sidestep many validation measures that characterise more traditional statistics in favour of increased possibilities for curating alternative realities. As the report states, "More and more enterprises are becoming architects of the unreal world, and as they push AI into more collaborative and creative roles, they are blurring the lines between what's real and what isn't."

"In this context, machine learning systems can turn into future-oriented tools for the creation of truths, as opposed to past-oriented tools for the uncovering of a pre-existing truth," Bruder says.

The Internet of ownership

Accenture's vision of synthetic realness is intimately linked to the concept of an "Internet of ownership", a digitally native infrastructure powered by blockchain, decentralised identities, confidential computing, and more.

"It's about faking it until it becomes real through instilling fervour, or conversely, through generating anxieties that have us embrace a vision of things to come," Bruder says.

Insisting on truth might not be an adequate response to computational data streams that are endlessly subject to manipulation.

"I think what we need to talk about are the very principles of discovery and prediction as a source of social and political orders," Bruder says. "As algorithmic societies increasingly blur the boundaries between the real and the unreal, the principles that underlie these transformations demand a closer look. It's only by confronting the very foundations that we can hope to navigate the complex landscape of synthetic realness and reclaim a sense of political agency in the face of an ever-shifting digital reality."

So instead of asking, 'Is this real?' we'll begin to ask, 'Is this authentic?'

THE PRESSURE FOR PURPOSE

In the age of big data and artificial intelligence, the very nature of information is undergoing a profound transformation. The traditional paradigm of purposeful data collection, where organisations gather and retain data for specific, well-defined objectives, is giving way to a new reality in which the potential value of data lies not just in its immediate utility but in the hidden insights and future applications that may be gleaned from it through the power of machine learning.

"Machine learning has created a pressure and often a desire in governments to gather data without a predefined purpose," explains Nanna Bonde Thylstrup, Associate Professor at the University of Copenhagen. "This data would previously have been deleted by archival institutions, but now they're kept, often through political pressure."

This shift puts significant pressure on archival institutions to rethink their role in society.

Beyond simply gathering and preserving data, they must now also consider issues of accountability and how to communicate the gaps that exist in their archives, which are becoming exacerbated by cloud technologies.

"Computer scientists will often look at public sector data as very rich, quality data," Thylstrup notes. "But the archivists are saying, well, actually, it's not anymore. There are many gaps."

As the pressure for purpose limitation continues to grow, businesses and organisations relying on public sector data for machine learning projects must be aware of the limitations and potential biases inherent in the data. There also needs to be more collaboration between data scientists, archivists, and policymakers to ensure that the data being used is fit for purpose and that any gaps or limitations are clearly communicated and understood.

Nanna Bonde Thylstrup

THE POLITICAL LOGIC OF GENERATIVE AI

As generative AI models become increasingly prevalent, understanding their underlying political logic is important for businesses and policymakers alike. Ludovico Rella and Alexander Campolo, postdoctoral researchers at Durham University in the UK, have identified four key aspects of this political logic: generativity, latent spaces, sequences and orders, and zero-shot learning. These can be mapped to simpler concepts as follows:

1. **Creating something new:** Generative AI models represent a revolutionary approach to data analysis and content creation, moving far beyond traditional computational methods. By deeply understanding underlying data structures, these models can predict, synthesise, and create new information that feels contextually authentic, almost as if the AI is developing its own creative intuition that can generate novel insights, text, images, or solutions previously unimagined.
2. **Hidden insights:** Latent spaces represent a breakthrough in how artificial intelligence conceptualises and organises information. These compressed representations allow AI to identify and explore intricate connections that might be invisible to human perception. These systems can extract core patterns, attributes, and relationships, essentially creating a nuanced, abstract understanding that transcends traditional data analysis.
3. **Flexible understanding:** Unlike rigid, step-by-step computational models, these systems can dynamically explore and interconnect ideas across vast information landscapes. This approach mirrors human-like cognitive flexibility, allowing the system to understand context, make intuitive leaps, and generate insights that feel genuinely intelligent and adaptive.
4. **Adaptive problem-solving:** Intelligent prompting has emerged as a groundbreaking technique for guiding AI behaviour. By carefully crafting precise, context-rich instructions, businesses can effectively "program" the AI to approach challenges with remarkable creativity and adaptability. This method allows organisations to leverage AI's vast knowledge base while maintaining precise control over its problem-solving approach.

"We see this ambivalent character of generative AI as an opening for thinking about different types of politics," Rella and Campolo explain. "This could involve not only mitigating bias or regulating the use of models from the outside, which is, of course, completely welcome, but also our ability to imagine or estimate other political distributions, other spaces, and other orders." GIBS