February 26, 2024

Today's Most Disruptive Technologies: Spotlight on Multimodal AI

Welcome back to Today's Most Disruptive Technologies! We turn from [quantum computing](#) to a spotlight on multimodal AI. Artificial intelligence (AI) continues to dominate the news and the markets and while some of us are still mulling over existential questions of what it means to be human, AI is about to take yet another fast turn—AI is going multimodal. The next generation of AI will be looking to connect through the very real—and very human—five senses of sight, sound, touch, smell, and taste, not to mention a whole new set of modalities from thermal radiation and haptics to brain waves and who knows what else.

What Is Multimodal AI?

The term "multimodal" or "having or involving [several modes, modalities or maxima](#)" is taking on increased significance in this age of AI, particularly in the context of generative (GenAI), with some calling it "[the next big thing in our path to achieving Artificial General Intelligence](#)."

- **Multimodal GenAI.** [Multimodal GenAI](#) is a type of AI that "combines multiple types or modes of data — text, images, video, audio, depth, etc." and "draws outputs from a combination of multiple data types" in order "to create more accurate determinations or make more precise predictions about real-world settings, scenarios, or problems." In other words, we are talking about a chatbot that can see, hear, and speak; a virtual assistant that inputs, outputs, and interacts via images, voice, and text; and what some are calling the [third wave of AI](#). While the first wave of AI was about classification and the current wave is about generative AI, the next wave might be about interactive AI.

- **How does multimodal AI work?** How does the model go from text to audio, video, and other modalities, and back again?

- ○ **Stacked models.** One way to accomplish this is by [stacking AI models](#) on top of one another. Here, (1) a user inputs an image into a chatbot; (2) that image is filtered by a separate AI; (3) that separate AI is built specifically to output detailed image captions in text; (4) those text descriptions are fed back to the chatbot; and (5) the chatbot responds to the now translated prompt.
- ○ **Grafted models.** Another way to do this is to have "much tighter coupling" where computer engineers "insert segments of one AI algorithm into another by combining the computer code infrastructure that underlies each model." This method is "[sort of like grafting](#) one part of a tree into another trunk" and then the grafted model continues to retrain on multimedia data sets until the AI is able to accurately link all that data and all those patterns across different modalities.
- ○ **Shared embeddings.** "[O]ne of the most powerful applications of deep learning models is to build an [embedding space](#), which is essentially a map of meanings for texts, images, audio, etc." A multimodal model (e.g., a vision language model) might have three [sub-models](#): (1) an image model to obtain image embeddings; (2) a text model to obtain text embeddings; and (3) a model to learn the relationships between them. These [embeddings](#) provide numerical representations that allow the model to easily understand semantic meanings and relationships among data.
- ○ **Vector databases.** A key evolution behind multimodal AI is how it "leverages multidimensional embeddings or indexing," stores all of that data in [vector databases](#), and "rel[ies] on vector databases for operation."
- ○ **Transformer model.** Multimodal AI might be the next generation of large language models (LLMs), but it is worth noting that most of today's LLMs rely on an earlier advance in AI called the [transformer model.](#) The transformer model is a "[neural network](#) architecture that can automatically transform one type of input into another type of output." This pivotal advance entered the AI scene way back in 2017, introduced to the tech world in the now sacrosanct research paper titled "[Attention Is All You Need](#)."

Why Does Multimodality Matter Now?

- **AI arms race.** Multimodal AI has been attracting renewed attention in the escalating AI arms race among big and small tech players alike. The ability to intake prompts and produce results in multiple formats like text, image, and voice "is emerging as a [competitive necessity](#) in the large language model (LLM) market." "[M]ultimodality has become a critical component for the [next generation of LLMs](#) and LLM-powered products."
- **More holistic AI.** "In order for AI to become a more useful tool, it has to learn how to accurately interpret content more holistically. This means working in [multiple modalities](#) (such as text, speech, and images) at once." A more [holistic AI](#) would be capable of learning not just from text, image/video, and audio, but also from other modalities such as record depth (3D); thermal energy (infrared radiation); and inertial measurement units (IMU), which calculate motion and position, as well as [haptics (touch)](#) and [emotion AI](#). In the ongoing search for general AI, multimodality could be key in enabling [future innovations](#) such as memory, planning, reasoning, and cognition.
- **More immersive experiences.** Multimodal understanding could be crucial in building "more [interactive, immersive, and smarter AI systems](#)." Multimodal AI has been described as "a concept devised, theorized, and now being implemented to deliver [multisensory immersive experiences](#)."
- **Computer vision**. "Machines don't have eyes," but "[machine eyes](#)" or "[computer vision](#)" might be better than a human doctor's eyes when it comes to detecting features in medical images that are not readily discernable by humans. Just imagine the ability for multimodal AI to more accurately interpret [X-rays, CT, and MRI scans, pathology slides](#), blood pressure, glucose levels, etc., and predict the likelihood of Parkinson's, [Alzheimer's](#), diabetes, kidney failure, [heart attacks, strokes](#), [cancer](#), and so on.

What Are the Potential Obstacles for Multimodal AI?

- **Technical challenges.** "Historically, analyzing such [different formats of data](#) together — text, images, speech waveforms, and video, each with a distinct architecture — has been extremely challenging for ?? machines."
- **Development costs.** "Initial model [development] is easily the [most costly aspects](#) because it includes perfecting the data science in parallel." "It's very [capital-intense](#). And it's probably even worse for multimodal, because consider how much data is in the images, and in the videos."
- **Proprietary training data.** "It's possible that the [cost of good training data](#) might become prohibitive (as original content creators/capturers seek to put monetization barriers around their data sets) for some time."
- **Bias.** "The combination of different content types brings [new risks of unintended harms](#) that may happen even when using "safe" system inputs." "[N]ew risks of unintended harms can arise on "[both sides](#)" of vision + language models." On the other hand, perhaps, multimodal AI, "if done correctly, … has less chance to produce bias to [more siloed models](#)."
- **Hallucinations.** Multimodal AI has just as much [potential to hallucinate](#) (and perhaps more) than unimodal AI. [Hallucinations](#) could become a greater risk in the case of vision-based models.
- **Malicious actors.** With multimodal GenAI capable of quickly replicating realistic synthetic voices from just a few seconds of real speech, malicious actors have a new and effective way to impersonate public figures or commit fraud. "[T]he versatility of multimodal LLMs 'presents a visual attacker with a wider array of achievable adversarial objectives,' essentially [widening the attack surface](#)." As multimodal AI continues to be integrated into our everyday lives, "[multimodal jailbreaks](#)" become a significant risk factor. Advances in multimodal AI models could mean "higher-quality, machine-generated content that'll be easier to personalize for '[misuse purposes](#).'"

What's Next for Multimodal AI?

- **Content moderation.** Multimodal AI could enhance the ability "to comprehensively understand the [content of social media posts](#) in order to recognize hate speech or other harmful content." "[Computer vision models](#) are used to detect and filter visual content that may contain toxicity or sensitive content." These models transcribe the raw video, analyze that transcription "by [leveraging block words](#) (i.e., removing any text containing words from a list of selected words) and [advanced NLP techniques](#) to filter content that may contain political, socio-economic, or demographic bias."
- **The metaverse.** Multimodal AI could help to "build AR glasses that have a more comprehensive understanding of the world around them, unlocking exciting [new applications in the metaverse](#)."
- **Hyperpersonalization.** "The future of generative AI is [hyper-personalization](#). This will happen for knowledge workers, creatives, and end users." On the other hand, perhaps "there will always be a [pendulum](#) between general [AI] tools and specialty tools."
- **Standardization.** While hyperpersonalization will continue at the user level, on the development side, "the next sweeping change in multimodal AI will be building more [standardized linkages](#) between different types of siloed models … and more open source offerings, making it easier and less expensive to train and run experiments."

- **Emerging use cases.** The growing ability for multimodal AI to input and output data in multiple modalities simultaneously is generating a multitude of new use cases all seemingly aimed at making chatbots more human.

    - **Sales and customer service.** Customer relationship management (CRM) platforms are "comparing foundation models across multiple criteria—including [qualitative aspects](#) such as friendliness, style,

and brand relevance." With phone-based automated support systems which "translate the sentiment apparent in our tone of voice into textual data [a] company can use for reporting and analysis," the art of customer service may never be the same again.

- ○ **Healthcare.** Multimodal AI promises to revolutionize the future of healthcare. "For people at risk for developing chronic medical conditions, a virtual health assistant could provide frequent feedback about their data to achieve prevention or better manage preexisting conditions." While "[t]here are already virtual AI chatbot health assistants for specific conditions such as diabetes, hypertension, obesity, and depression, … none have yet become holistic or preventive." Multimodal AI could also "make remote monitoring a reality, allowing a "hospital-at-home" with continuous vital-sign capture that is equivalent to an intensive care unit." Digital twins are already making their way into other industries such as manufacturing, design, architecture, retail, and entertainment, but in the healthcare context, a digital twin "would be informative for a person with a new diagnosis by providing a digital facsimile on which to find a successful treatment."
- ○ **Assistive technology.** Multimodal AI could lead to more accurate, more intuitive assistive technology tools, such as AI-powered description services for blind and low-sighted people.
- ○ **Shopping.** The ability for shoppers to search by image and for sellers to sell and market using images, while having the make and model, the relevant e-commerce categories, the required product descriptions, and other text and data automatically generated could transform the entire virtual shopping experience.
- ○ **Home security.** The ability to capture limitless security camera images, search for specific clips among countless hours of video, and track individuals through facial recognition tools stands to transform the home security business, not to mention electronic surveillance and law enforcement more generally.
- ○ **Autonomous driving.** The ability for multimodal AI to see, hear, and contextualize could lead to safer, more eco-friendly, and less traffic-heavy autonomous driving.
- ○ **Manufacturing.** "Multimodal generative AI can be leveraged to improve quality control in manufacturing, predictive maintenance of automobiles, and supply chain optimization in manufacturing."
- ○ **Knowledge economy.** "However, like previous technological inventions, multimodal and regular GenAI allow dozens of professions to evolve. Lawyers, writers, scientists, teachers, and more could optimize time-consuming tasks such as research, strategy development, document drafting and generation, and more, provided it falls under the purview of the underlying data the multimodal GenAI tool is trained on. In short, the knowledge economy could see a massive shift if the right data is available." Imagine being able to rely on your own personal AI assistant to handle all of those bothersome tasks that are holding you back!

Concluding thoughts. Multimodal AI opens up the possibility of a more holistic, a more interactive, and perhaps a more human AI in all of our futures. However, while the chatbots of tomorrow might talk, see, feel, and even emote their way into being our favorite helper, co-worker, or friend, businesses will need to adopt a "human-in-the-loop" approach to protect consumers and workers against fraud, bias, and other potential harms presented by this seemingly more human face to AI.

Follow us on social media @PerkinsCoieLLP, and if you have any questions or comments, contact us here. We invite you to learn more about our Digital Media & Entertainment, Gaming & Sports industry group and check out our podcast: *Innovation Unlocked: The Future of Entertainment*.

Authors

Explore more in

[Technology Transactions & Privacy Law](#)