VCDNP

Emerging Tech Brief No. 1

Artificial Intelligence Models and Systems

Dr. Sarah Case Lackner and Zaheed Kara

New technologies are emerging at an unprecedented pace and are influencing international relations in ways that are impossible to avoid. A basic understanding of their functioning, benefits, and shortcomings is essential for policymakers and diplomats working in fields as diverse as international security, nuclear energy, counter-terrorism and international development. This is particularly true for artificial intelligence (AI) models and systems, a technology that has moved from relative obscurity to a daily tool for millions in just a few years.

This brief, aimed at policymakers and diplomats internationally, provides a concise introduction to AI models and systems. It discusses the origin and development, beneficial use cases, shortcomings and vulnerabilities, and potential future developments in the field that may have implications for policymakers and diplomats.

What is Artificial Intelligence?

The term "artificial Intelligence" was first coined by Professor John McCarthy to describe machines that could perform tasks characteristic of human intelligence. Today, AI systems can discover patterns in large quantities of data, process images and speech, make decisions, and generate novel images and text, among other tasks.

While AI has been a field of research since the 1950s, developments in **machine learning**, a field that seeks to mimic human learning, and, in particular, the use of **neural networks**, have largely fueled the recent explosion of progress in development and use of AI models and systems. Machine learning allows software programmes, known as **AI models**, to detect patterns from large data sets, rather than being explicitly programmed to do so. Once the AI model has been trained, it is then able to use the connections identified in the training data to produce **output data** based on **input data** provided to it, such as a typed prompt. **AI systems** are applications that use trained AI models to perform a specific task.

Two macro-trends have fueled the exponential increase in AI model capabilities: **big data** and **computational power**. Machine learning requires a significant quantity of high-quality data on which a model can be trained to make the needed connections for its intended application, known as **training data**. The ubiquity of digitised data in the internet age has played a key role in the recent progress in AI systems and models, as vast databases of images and and large amounts of digitised text have provided needed input for these algorithms. Technological developments over the past few decades have also led to a massive increase in available computational power, enabling these machine learning algorithms to make use of this vast amount of data in shorter and shorter timeframes.

AI models can be broadly categorised as predictive or generative. **Predictive AI models** use statistical techniques and machine learning to analyse and solve computationally complex and data intensive problems. Traditionally, most AI applications in industry have focused on predictive AI systems used for predicting customer behaviour, identifying fraudulent transactions, or forecasting weather patterns. **Generative AI models** create novel or derived content based on a user's prompt by identifying patterns and relationships in the data they are trained on, such as images, computer code, and text. Generative AI models include the well-known subset of AI models referred to as *Large Language Models* (LLMs), which specialise in understanding and generating natural language.

A final important distinction is between narrow and general-purpose systems. **Narrow AI** is designed to perform a single, specific task and cannot operate outside of its predefined domain. Most AI systems in use today are examples of narrow AI, such as recommendation systems or email spam filters. While highly effective at their specialised functions, their capabilities are not transferable. In contrast, **general-purpose AI** refers to models with significant generality which are capable of competently performing a wide range of distinct tasks, and that can be integrated into a variety of downstream systems or applications. The remainder of this brief focuses on general-purpose, generative AI models.

What are some shortcomings of general-purpose AI models and systems?

As discussed above, machine learning algorithms are used to determine connections between the training data provided to AI models and create outputs without explicitly being given rules for making these connections by a human.

The lack of human involvement in this learning process leads to a conundrum: even after the AI model has been trained and is ready for use, it may not be clear which connections have been made or what rules the AI model is applying to make them. This phenomenon is often referred to as the **black box problem**. A classic example of this problem is illustrated by Professor Melanie Mitchell in her book *Artificial Intelligence: A Guide for Thinking Humans*, where a graduate student training an AI model to identify animals in photographs arrived at the conclusion that the machine learning model had instead learned to identify the defocused background in photographs including animals, concluding that this background constituted an "animal".[1] While the results were frequently reliable, the reasoning was far from human.

Complicating this picture, the training data used to teach the AI model to make connections may also be unknown, or at best, only partially known, to the end user. This creates the possibility that the model or system has hidden biases it learned from its training data, such as the misclassification of dark-skinned and female faces by AI models, highlighted in 2018 by Buolamwini and Gebru.[2] The heavy reliance on high-quality training data and integrity opens avenues for new types of cyber-attacks. For instance, in **data poisoning attacks**, malicious actors intentionally corrupt or otherwise manipulate the training data used for machine learning algorithms in order to introduce biases, create erroneous outputs, introduce vulnerabilities, or otherwise influence the decision-making of the model.

There is also the possibility that an AI system itself, including those installed in critical applications, could be infiltrated or attacked by a malicious actor, once in use. Instead of just tricking the model's inputs, a malicious actor could target the core components of the system. For example, they could launch a **model extraction attack**, repeatedly querying a system to reverse-engineer and steal the valuable, proprietary AI model. Alternatively, an attacker may attempt to compromise the servers hosting the AI, planting malware that subtly alters its decision-making logic or executing a denial-of-service attack to shut down an essential AI service. It may also be possible to embed a "trojan" or hidden backdoor into a model, causing it to perform normally until a specific, secret trigger causes it to fail or grant unauthorised access.

The current generation of AI models have a tendency to reach false conclusions and assert them with certainty, commonly referred to as **hallucinations**. For example, AI models have been shown to cite non-existent court cases to support a legal argument. These instances of confidently presented misinformation make rigorous

[1] Melanie Mitchell, "Artificial Intelligence: A Guide for Thinking Humans", Farrar Strauss & Giroux, 15 October 2019.
[2] Joy Buolamwini and Timnit Gebru, "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification", Proceedings of the 1st Conference on Fairness, Accountability and Transparency, Proceedings of Machine Learning Research, 81:77-91, 2018. Available at : https://proceedings.mlr.press/v81/buolamwini18a.html.

human verification essential, especially in high-stakes fields. That said, model developers are taking steps to decrease hallucinations and users can do so by implementing techniques such as chain-of-thought verification, iterative refinement, and external knowledge restriction.

How can AI models and systems be misused, and what protects against it?

While highly capable general-purpose AI models have incredible potential for the general public and those planning beneficial applications, they may also be misused by malicious actors. For example, the generation of highly convincing fake video and audio using AI models, known colloquially as **deepfakes**, is implicated in the increasing spread of misinformation and disinformation, particularly on social media. The potential for deepfakes to enable criminal activity is a significant concern in many industries, particularly those that rely on high-integrity remote identification. Models may also have the potential to uplift malicious actors in more severe, large-scale risks, such as in creating and deploying weapons of mass destruction or lowering barriers to carrying out cyber-attacks on critical infrastructure. For instance, criminals may use advanced AI models to enhance their abilities to undertake sophisticated cyber-attacks on critical infrastructure, or complex social engineering attacks.

Most major AI developers have acknowledged the potential risks arising from model misuse and are focused on developing **model guardrails** (otherwise known as **technical safeguards** in the AI community), which are technical measures designed to prevent LLMs from assisting malicious actors in engaging in harmful behaviour, for example by providing useful and detailed information on the development of chemical, biological, radiological, and nuclear weapons. That said, multiple researchers have demonstrated that many such guardrails can be bypassed by a determined and skilled attacker.[3,4]

In the future, AI systems used by security professionals may prove to be one of the most effective methods for defending against AI-enhanced attacks. For example, AI systems are showing promise in detecting deepfakes, and could be used by cyber security professionals to automate threat detection, identify patterns, and respond to threat incidents.

What is coming next for AI?

Given the pace of change of AI technology over the past few years, it is difficult to say with any certainty what the next major breakthrough for AI is likely to be. Recent trends in the AI community include:

---

[3] Andy Zou, et al. “Universal and Transferable Adversarial Attacks on Aligned Language Models”, arXiv preprint arXiv:2307.15043, 20 December 2023.
[4] Xiangyu Qi, et al. "Fine-tuning Aligned Language Models Compromises Safety, Even When Users Do Not Intend To!", arXiv preprint arXiv:2310.03693, 5 October 2023.

Agentic AI

This refers to AI systems that perform tasks and make decisions autonomously, not requiring human prompts or oversight to pursue complex goals. Agentic AI systems are still at a relatively early stage of development, and there are challenges associated with their deployment, notably ensuring that their goals align with human values (known as the alignment problem). Future capabilities of agentic systems may include breaking down complex problems into multiple, actionable steps, operating independently to complete multi-step tasks over extended periods, taking decisions and adapting strategies independently based on outcomes and changing conditions, coordinating multiple sub-tasks to achieve overarching objectives, and potentially managing other AI tools or systems to accomplish goals (known as multi-agent systems).

Tool Use and Code Generation

AI systems are increasingly capable of using external tools to write functional code and interface with other software systems. Tool-use for code generation is still at an early stage of development, and there are still challenges associated with their operation, notably ensuring that generated code is secure and doesn't introduce vulnerabilities. In the future, this may include generating scripts for data analysis or exploiting system vulnerabilities, automating repetitive tasks, interfacing with databases and information systems, or creating custom tools to overcome specific technical challenges.

Advanced Reasoning and Chain-of-Thought Capabilties

These are capabilities of AI systems that can maintain context over long interactions, break down complex problems into steps, and perform multi-step reasoning to solve these problems without requiring continuous human guidance, which are critical to addressing complex technical problems systematically. Advanced reasoning systems are still in an early stage of development, and there are challenges associated with their deployment, such as ensuring accurate and reliable outputs when handling complex technical domains. Future capabilities may include solving complex, domain-specific technical problems, maintaining coherent reasoning chains across many intermediate steps, and potentially identifying novel solutions to problems by connecting disparate fields of knowledge.

Multimodal AI Systems

AI models that can process and integrate multiple types of data inputs simultaneously - text, images, video, audio, and even scientific data formats. These systems exist today and can act as useful co-scientists. In the future they may be able to analyse satellite imagery while cross-referencing technical documents, process facility blueprints alongside security camera footage, or interpret scientific papers about nuclear materials while examining visual data about facility layouts. These are just a few examples of where AI could go next. More possibilities will likely emerge in the coming months and years, and the potential of these "next directions" will gradually become more concrete.

VCDNP

Vienna Center for Disarmament
and Non-Proliferation

The VCDNP is an intenational non-governmental organisation that promotes peace and security by conducting research, facilitating dialogue, and building capacity on nuclear non-proliferation and disarmament.

vcdnp.org

info@vcdnp.org

@VCDNP

VCDNP